

DEPARTAMENT DE GENÈTICA EVOLUTIVA

DISEÑO, DESARROLLO E IMPLEMENTACIÓN DE UNA
PLATAFORMA BIOINFORMÁTICA ORIENTADA AL
ANÁLISIS Y GESTIÓN DE INFORMACIÓN EN
EPIDEMIOLOGÍA MOLECULAR.

ALICIA AMADOZ NAVARRO

UNIVERSITAT DE VALÈNCIA
Servei de Publicacions
2011

Aquesta Tesi Doctoral va ser presentada a València el dia 21 d'octubre de 2011 davant un tribunal format per:

- Dr. Julio Rozas Lirias
- Dra. Elisa Martró Català
- Dr. Juan José Abellán Andrés
- Dr. Francisco Xavier López Labrador
- Dra. María Esther De Ves Cuenca

Va ser dirigida per:

Dr. Fernando González Candelas

©Copyright: Servei de Publicacions
Alicia Amadoz Navarro

I.S.B.N.: 978-84-370-8529-6

Edita: Universitat de València
Servei de Publicacions
C/ Arts Gràfiques, 13 baix
46010 València
Spain
Telèfon:(0034)963864115

Diseño, desarrollo e implementación de una plataforma bioinformática orientada al análisis y gestión de información en epidemiología molecular.

Alicia
Amadoz
Navarro
2011



Facultat de Ciències Biològiques
Institut Cavanilles de Biodiversitat i Biologia Evolutiva
Departament de Genètica Evolutiva

**Diseño, desarrollo e implementación de una
plataforma bioinformática orientada al análisis y
gestión de información en epidemiología molecular.**

Memoria presentada por Alicia Amadoz Navarro para optar al grado de
doctor en Ciencias Biológicas por la Universitat de València.

Director

Dr. Fernando González Candelas
Catedrático de la U.V.E.G.

Tesis doctoral
Valencia, 2011

Portada diseñada por David Chanzá Téllez.

D. Fernando González Candelas, Doctor en Ciencias Biológicas y Catedrático del Departament de Genètica de la Universitat de València.

CERTIFICA: Que Alicia Amadoz Navarro, Licenciada en Bioquímica por la Universidad de Navarra, ha realizado bajo su dirección el trabajo que lleva por título: "Diseño, desarrollo e implementación de una plataforma bioinformática orientada al análisis y gestión de información en epidemiología molecular.", para optar al Grado de doctor en Ciencias Biológicas por la Universitat de València.

Y para que conste, en el cumplimiento de la legislación vigente, firmo el presente certificado en Valencia, a 19 de Abril de 2011.

Fdo.: Dr. Fernando González Candelas

A mis padres, Emilio Amadoz López y Maribel Navarro Mendive

A mi hermano, Gustavo

A David Chanzá Téllez

"Would you tell me, please, which way I ought to go from here?"

"That depends a good deal on where you want to get to," said the Cat.

"I don't much care where – " said Alice.

"Then it doesn't matter which way you go," said the Cat.

" – so long as I get somewhere," Alice added as an explanation.

"Oh, you're sure to do that," said the Cat, "if you only walk long enough."

(Alice's Adventures in Wonderland. Lewis Carroll, 1865)

Agradecimientos

Son muchas las personas a las que quisiera agradecer su ayuda y apoyo durante el tiempo que ha llevado la elaboración de esta tesis. En primer lugar quiero expresar mi más profundo agradecimiento a mi director, Fernando González-Candelas, por aceptar dirigir esta tesis, por guiarme a lo largo del doctorado, por estar siempre disponible para resolver todas mis dudas y preguntas y por permitirme ampliar mi formación mediante la asistencia a cursos y congresos. Doy las gracias también a Andrés Moyá por permitirme formar parte del Departamento de Genética Evolutiva del Instituto Cavanilles. Gracias a la Consellería de Sanidad Valenciana por la financiación durante los primeros 4 años de este trabajo. Muchas gracias a todos los compañeros y profesores del programa de doctorado “Biodiversidad y Biología Evolutiva” por las innumerables discusiones y reflexiones científicas durante las clases. Gracias a Iñaki Comas por compartir su despacho conmigo, por su paciencia con mis recurrentes cuestiones evolutivas y por ser una magnífica persona. Gracias a Yolima Carrillo-Cruz por su amistad, por su eterna sonrisa y por contagiar su alegría y carácter colombianos. Gracias a Pascual Asensi por su ayuda con los servidores, por sus consejos informáticos y por tantos cafés sin café. Gracias a José Manuel Cuevas, Victoria Furió y María José López por hacerme sentir como en casa en el “contenedor 1”. Gracias a Teresa Cortés por su apoyo y amistad. Gracias a Vicente Sentandreu por hacernos reír tanto. Gracias a todos los demás miembros del departamento: Ana González, Amparo Latorre, David Martínez, Paco Silva, Juli Peretó, Sari Gil, Xavier López, Alex Neef, Alma Bracho, María José Gosálbes, Silvia Ramos, Rafa Sanjuán, Nuria Jiménez, Giuseppe D'Auria, Miguel Pignatelli, Javier Tamames, Manoli Torres, Eugeni Belda, Araceli Lamelas, Benja Ortiz, Mireia Coscollá, Vicente Pérez, Laura Gómez, Francisco Peris, Ana Durbán y Rafa Patiño por hacer que mi paso por el Cavanilles sea siempre un agradable recuerdo.

La escritura de esta tesis comenzó a la par que mi trabajo en Adapting S.L., donde Luis Blasco me dio la oportunidad de ampliar mis conocimientos de programación junto a un equipo de grandes profesionales. Gracias a Victor Muñoz por su apoyo y profundo conocimiento en tecnologías web. Gracias a Laura Gil y Vanessa Jiménez por su alegría, optimismo y amistad. Gracias a Victor Sáez por su ejemplo de buena actitud frente a las dificultades. Gracias a todos los demás compañeros: Luis Llorens, Luis Sáez, Rubén Lafuente, Dolo Nebot, Belén Aguilera, Teresa Quílez, Diego Hernández, Carlos Millán, Juan Honrubia, Marc Climent, Rubén Parra, Javier Cerdá y Hugo Bárzena por todo lo que me enseñaron y los buenos momentos compartidos.

Quisiera extender este agradecimiento al Departamento de Bioinformática y Genómica del Centro de Investigación Príncipe Felipe donde trabajo actualmente. Gracias a Ximo Dopazo por darme la

oportunidad de volver al mundo de la investigación y formar parte de un excelente grupo de trabajo tan diverso y enriquecedor. Gracias a David Montaner, Paco García, Sonia Tarazona y Quique Vidal por los muchos consejos estadísticos que han sido aplicados en este trabajo. Gracias a Stefan Götz, Alejandro de María y Roberto Alonso por sus comentarios informáticos sobre esta tesis. Gracias a Hernán Dopazo, Adriana Cucchi, Eva Alloza, Patricia Sebastián, Luz García, Marta Bleda, David Montaner, Paco García, Stefan Götz, Quique Vidal, Martina Marbá, Davide Baù, Sonia Tarazona, Fernando García, Ana Conesa, Javier Santoyo, Marc Martí-Renom, Nacho Medina, Jose Carbonell, Joaquín Tárraga, Pablo Escobar, Luis Pulido, Roberto Alonso, Alejandro de María, François Serra, Rubén Sánchez, Maria Jaime, Aaron Weimann, Jorge Jiménez, Arturo Sylveira, Patricia Díaz, Rodrigo Lomas, David Dufour, Pablo Arce, Jorge García, Dani Navarro, Marina Naval, Chiara Russo, Fede García y Jordi Durbán por estar siempre dispuestos a ayudar y por crear, con su energía y entusiasmo, un fantástico ambiente de trabajo que me ha motivado en estos últimos meses.

Doy las gracias a Marta Campos, Azahara Abedine, Andrea Urrecho, Itxaso Ilundáin, Uxue Muguerza, Aintzane Zabaleta, Nora Ibargoyen, Olatz Alberdi, Estíbaliz Ruiz, María Gutiérrez, Beatriz Lorente y Nelo Puchades, Beatriz González y Daniel García, Rubén Arnal y Amparo Fayós por los buenos y malos momentos, por su amistad y por su apoyo durante tantos años. Gracias a María Téllez-Plaza por su amistad, por tantas conversaciones sobre la ciencia y la vida y por sus comentarios epidemiológicos sobre la introducción de esta tesis.

Quiero agradecer especialmente a mi padres, Emilio Amadoz y Maribel Navarro, la educación que me proporcionaron, todo su amor y apoyo incondicional. A mi hermano Gustavo, por enseñarme a ser más generosa, y a toda la familia, por estar siempre ahí.

A David, por pensar y creer en mí, por estar a mi lado y apoyarme siempre, especialmente durante los más de dos años que llevo escribiendo esta tesis. También quiero agradecer a Ramón, Reme, Ana Sofía, Rafa y al resto de familia valenciana por acogerme como una más.

ÍNDICE GENERAL

1 INTRODUCCIÓN GENERAL	5
1.1 CONCEPTOS SOBRE EPIDEMIOLOGÍA MOLECULAR.....	5
1.2 GESTIÓN DE LA INFORMACIÓN.....	7
<i>1.2.1 Conceptos sobre bases de datos</i>	7
<i>1.2.2 Bases de datos de biología molecular</i>	9
<i>1.2.3 Integración de la información</i>	10
1.3 MECANISMOS VIRALES DE ADQUISICIÓN DE RESISTENCIAS	11
1.4 LA HEPATITIS C	15
<i>1.4.1 Epidemiología</i>	16
<i>1.4.2 Estructura del genoma viral y proteínas virales.....</i>	18
<i>1.4.3 Tratamiento y variabilidad del genoma viral</i>	20
1.5 SISTEMAS DE INFORMACIÓN EN EPIDEMIOLOGÍA MOLECULAR	22
2 OBJETIVOS	27
3 DISEÑO E IMPLEMENTACIÓN.....	31
3.1 ESPECIFICACIÓN DE REQUISITOS	31
3.2 ARQUITECTURA DE LA PLATAFORMA BIOINFORMÁTICA	35
3.3 BASES DE DATOS.....	38
<i>3.3.1 Análisis.....</i>	39
<i>3.3.2 Diseño</i>	42
<i>3.3.3 Implementación.....</i>	43
3.4 INTERFAZ WEB	45
<i>3.4.1 Análisis.....</i>	45
<i>3.4.2 Diseño</i>	47
<i>3.4.3 Implementación.....</i>	49
3.5 HERRAMIENTAS DE GESTIÓN DE DATOS	50
<i>3.5.1 Gestión de la base de datos local</i>	51
<i>3.5.2 Búsquedas en bases de datos externas.....</i>	58
<i>3.5.3 Gestión de ficheros propios del usuario</i>	61
<i>3.5.4 Gestión de ficheros de resultados</i>	62
3.6 HERRAMIENTAS DE ANÁLISIS DE DATOS	63
<i>3.6.1 Búsqueda de secuencias por similitud</i>	66
<i>3.6.2 EMBOSS</i>	70
<i>3.6.3 Alineamiento múltiple de secuencias</i>	72
<i>3.6.4 epiPATH-tools.....</i>	77
<i>3.6.5 Libsequence.....</i>	93
<i>3.6.6 Análisis estadísticos</i>	94
<i>3.6.7 Utils.....</i>	96
3.7 SEGURIDAD DE LA INFORMACIÓN	96

4 APLICACIONES.....	101
4.1 PREDICCIÓN DE LA RESPUESTA AL TRATAMIENTO EN PACIENTES CON HEPATITIS C	101
<i>4.1.1 Introducción</i>	<i>101</i>
<i>4.1.2 Material y métodos.....</i>	<i>103</i>
<i>4.1.3 Resultados</i>	<i>115</i>
<i>4.1.4 Discusión</i>	<i>123</i>
4.2 ANÁLISIS EVOLUTIVO Y POBLACIONAL DE UN BROTE CAUSADO POR EL VIRUS DE LA HEPATITIS C.....	133
<i>4.2.1 Introducción</i>	<i>133</i>
<i>4.2.2 Material y métodos.....</i>	<i>135</i>
<i>4.2.3 Resultados</i>	<i>137</i>
<i>4.2.4 Discusión</i>	<i>144</i>
5 DISCUSIÓN GENERAL	149
6 CONCLUSIONES	157
7 BIBLIOGRAFÍA	161
8 ANEXOS.....	201
8.1 ESPECIFICACIÓN DE REQUISITOS DE LA BASE DE DATOS PRINCIPAL	201
8.2 ESPECIFICACIÓN DE REQUISITOS DE LA BASE DE DATOS SECUNDARIA	219
8.3 ESQUEMA DE LA BASE DE DATOS PRINCIPAL	221
8.4 ESQUEMA DE LA BASE DE DATOS SECUNDARIA.....	223
8.5 DIAGRAMA ER DE LA BASE DE DATOS PRINCIPAL	224
8.6 DIAGRAMA ER DE LA BASE DE DATOS SECUNDARIA.....	225
9 LISTADO DE ABREVIATURAS.....	229

1. INTRODUCCIÓN GENERAL

1 Introducción general

1.1 Conceptos sobre epidemiología molecular

La Epidemiología es la disciplina científica que se dedica al estudio de la frecuencia y la distribución de los estados o eventos relacionados con la salud en poblaciones específicas (Porta 2008). Una de sus aproximaciones es la Epidemiología molecular, que estudia los mecanismos moleculares, la patofisiología y la etiología de una enfermedad mediante el empleo de biomarcadores. Un biomarcador es una sustancia detectable que permite identificar algún proceso biológico y que, en el contexto de la Epidemiología molecular, sirve para estratificar grupos y subgrupos en los que se observan mejor las asociaciones epidemiológicas (Ambrosone *et al.* 1997; Barreto *et al.* 2006). Los biomarcadores son diferentes según la enfermedad que se estudie y pueden obtenerse mediante distintas técnicas derivadas de la inmunología, bioquímica, biología molecular y genética con el propósito de identificar, genotipar o determinar los orígenes y la propagación de los microorganismos, particularmente los patógenos (Foxman *et al.* 2001; Morgan *et al.* 2001; Campoccia *et al.* 2009).

Las enfermedades de origen infeccioso, causadas principalmente por virus o bacterias, son uno de los tipos de enfermedades más estudiados en el área de la Epidemiología molecular (Preston 2003). Esto es debido a que el conocimiento de su biología, origen y propagación es fundamental para la realización de intervenciones en salud pública y medicina preventiva (Holmes 1998; Lappalainen *et al.* 2001).

Los virus son microorganismos de estructura muy sencilla, capaces de reproducirse únicamente en el interior de células vivas utilizando su metabolismo. Las bacterias son microorganismos constituidos por una única célula sin núcleo diferenciado. Ambos tipos de microorganismos tienen material genético formado por moléculas de ácido desoxirribonucleico (DNA) o ácido ribonucleico (RNA) en algunos virus.

El dogma central de la biología molecular (Crick 1970) postula que la información genética de un organismo se traduce mediante una serie de procesos moleculares en el que se pasa del DNA al RNA y posteriormente a proteínas. Por tanto, la información contenida en la secuencia de nucleótidos de DNA que conforman los genes codifica la síntesis de una secuencia de aminoácidos. Este dogma es una visión simplificada y actualmente se conocen muchos más detalles del flujo de información a nivel molecular (Shapiro 2009; Gao *et al.* 2010). Dentro de la estructura molecular de la información genética, se define como gen la unión de secuencias genómicas que codifican un grupo coherente de productos funcionales potencialmente solapantes (Gerstein *et al.* 2007). En el caso de una fracción importante de los virus patógenos para el ser humano la información genética se encuentra en el RNA. La interacción paciente-patógeno puede considerarse como una relación hospedador-parásito en la que los mecanismos de defensa del primero hacen que el segundo trate de evitarlos para poder continuar sus funciones vitales. De esta forma, el tratamiento administrado al paciente, que ayuda a sus defensas, hace que el patógeno cambie para adaptarse al nuevo medio y, en último caso, volverse resistente. El cambio en el patógeno queda reflejado en su información genética y se emplea para

evaluar su evolución, así como sus características genéticas y epidemiológicas.

1.2 Gestión de la información

La gestión de la información trata de recoger y distribuir datos e información entre diferentes fuentes y destinatarios para su posterior análisis. Actualmente constituye una parte crítica en toda organización porque permite una mayor eficiencia y calidad del trabajo desarrollado (López-Hernández 1990). Un sistema de información (SI) es un sistema que gestiona y controla los flujos por los que se distribuye la información que proviene de fuentes tanto internas como externas (Arias-Coello *et al.* 2006). Los elementos de un SI son, fundamentalmente, tres: el elemento humano, imprescindible para la ejecución eficaz de las tareas; los soportes informáticos para almacenar, distribuir los datos y el *software* que automatiza los flujos de trabajo o *workflows* (secuencia de actividades utilizadas para la ejecución de un proceso); y los datos que van a ser procesados. Las funciones de un SI son la recogida, el procesamiento, el almacenaje, la elaboración y la presentación de los datos.

1.2.1 Conceptos sobre bases de datos

Una base de datos se define como una colección de datos estructurada de forma que refleje las interrelaciones y restricciones que existen en el mundo real (Silberschatz *et al.* 2002). Los datos han de poder ser compartidos por varios usuarios y aplicaciones y deben mantenerse independientes de estos. El sistema gestor de bases de datos (SGBD) es el conjunto de programas y procedimientos que suministran los medios

necesarios para describir, recuperar y manipular los datos almacenados en la base de datos, manteniendo su integridad, confidencialidad y seguridad. La interacción con la base de datos se realiza con el lenguaje de definición (DDL) y el de manipulación (DML) de datos. El lenguaje estandarizado para manipular información en las bases de datos (SQL) es una combinación de los anteriores y es el empleado en las bases de datos relacionales.

Existen diversos modelos de bases de datos. Un modelo es un conjunto de reglas para la representación de los datos. El modelo más empleado en la actualidad es el modelo relacional (Codd 1970) compuesto por relaciones, que son básicamente tablas donde una tupla corresponde con una fila y un atributo corresponde a una columna. El diagrama entidad-relación (ER) ofrece una visión unificada de los datos. En él se representan las estructuras que constituyen el contenido de la base de datos junto con las restricciones que limitan las ocurrencias o instancias válidas. Para ello hace uso, fundamentalmente, de tres conceptos:

- Entidad: objeto acerca del cual queremos almacenar información.
- Atributo: propiedad de una entidad o de una relación.
- Relación: asociación o correspondencia entre dos o más entidades. Se define por su nombre, grado (número de entidades que relaciona) y por el tipo de correspondencia o cardinalidad, que indica el número de ocurrencias mínimo y máximo de una entidad que participa en la relación. Gráficamente esto se representa mediante 1:1 (uno a uno), 1:N (uno a muchos) y N:N (muchos a muchos).

1.2.2 Bases de datos de biología molecular

Las bases de datos públicas de biología molecular almacenan grandes cantidades de información derivada de distintas fuentes y de los grandes proyectos de secuenciación realizados hasta ahora, y que continúan realizándose.

Existen más de 1330 bases de datos sobre biología molecular y celular públicamente disponibles (Galperin *et al.* 2011). La principal base de datos de secuencias de DNA está formada por tres grandes grupos de bases de datos que colaboran entre sí, siendo conocidos actualmente como *International Nucleotide Sequence Database Collaboration* (INSDC, www.insdc.org), quien proporciona los estándares comunes para el envío de secuencias. Los grupos que forman parte del INSDC son *GenBank* (Benson *et al.* 2011), construida y distribuida por el Centro Nacional de Información Biotecnológica (NCBI) localizado en Estados Unidos; *EMBL Nucleotide Sequence Database* (EMBL-Bank) (Leinonen *et al.* 2011), que pertenece al Instituto Europeo de Bioinformática (EBI) y se encuentra localizada en el Reino Unido; y *DNA Data Bank of Japan* (DDBJ) (Kaminuma *et al.* 2011), producida y mantenida por el Instituto Nacional de Genética en Japón. Cada uno de los grupos recoge una porción del total de secuencias que se envían en todo el mundo para luego intercambiar esta información entre sí y producir un único almacén sincronizado (Cochrane *et al.* 2011).

En la Tabla 1 se encuentran las estadísticas correspondientes a la última distribución disponible a fecha de 10 de Enero de 2011.

Tabla 1. Número de registros en las bases de datos de secuencias.

Base de datos	Distribución	Número de secuencias	Número de nucleótidos
EMBL-Bank	Enero 2011	199.575.971	301.588.430.603
GenBank	Diciembre 2010	129.902.276	122.082.812.719
DDBJ	Diciembre 2010	128.607.782	120.919.931.265

Los números proporcionadas por EMBL-Bank son considerablemente mayores a consecuencia de que existen varios tipos de registros según la técnica de obtención de las secuencias (Tabla 2), de los cuales el tipo de registro *Standard* es el equivalente a los registros sincronizados con las otras grandes bases de datos.

Tabla 2. Relación de tipos de registros en EMBL-Bank tomado de <http://www.ebi.ac.uk/embl/Services/DBStats> a fecha de 10 Enero 2011.

Tipo de registro	Registros	Nucleótidos
<i>Standard</i>	128.262.666	120.603.334.814
<i>Constructed</i> (CON)	6.381.010	225.047.233.405
<i>Third Party Annotation</i> (TPA)	6.894	385.832.010
<i>Whole Genome Shotgun</i> (WGS)	64.925.118	180.599.264.067

1.2.3 Integración de la información

Una de las áreas de la Bioinformática es el manejo y análisis de datos biológicos de forma informatizada, de manera que agrupa diversas materias como la biología, la informática, la estadística y las tecnologías de la información para facilitar nuevas vías de comprensión de la información disponible y crear perspectivas globales en biología.

Los proyectos de investigación realizados en distintas áreas biológicas exigen actualmente el empleo de grandes cantidades de datos para contrastar las hipótesis propuestas, revelar asociaciones, o plantear nuevas hipótesis. La información disponible se encuentra distribuida en distintas

bases de datos con distintos formatos y formas de acceso, la mayoría de ellas no diseñadas para el acceso automatizado mediante máquinas (Stein 2003). Este hecho es uno de los problemas en la integración de la información biológica y la bioinformática se ha convertido en una parte esencial de las soluciones eficientes del manejo, integración, creación de nueva información y transformación de los datos crudos en conocimiento biológico (Searls 2000; Haas *et al.* 2001; Lacroix 2002; Jackson *et al.* 2003; Karasavvas *et al.* 2004; Köhler 2004; Maojo *et al.* 2004; Louie *et al.* 2007).

Un SI que integre la información biológica está compuesto, básicamente, por una base de datos que almacena y organiza los datos masivos además de herramientas específicas desarrolladas para su visualización y análisis (Fellenberg 2003; Brazhnik *et al.* 2007; Romano 2008). Este sistema debe facilitar el acceso a la información almacenada y métodos para la extracción de aquella información necesaria para responder una cuestión biológica concreta. En este sentido, la interfaz *web* es una de las tecnologías empleadas universalmente en el acceso a la información y que tiene un papel importante en la investigación genética, permitiendo un intercambio libre y universal de datos biológicos (Guttmacher 2001).

1.3 Mecanismos virales de adquisición de resistencias

Los mecanismos moleculares que provocan cambios en el material genético se encuentran modulados por las fuerzas evolutivas. Las resistencias observadas en los virus se originan principalmente por su capacidad de generación de variabilidad mediante los mecanismos de mutación, recombinación y

reordenamientos genómicos que, junto a altas tasas de replicación y elevados tamaños poblacionales, les confieren una gran capacidad de adaptación (Moya *et al.* 2004).

Se considera una mutación la alteración de la información contenida en el material genético. Los virus de RNA presentan altas tasas de mutación por ausencia o baja eficacia de la actividad correctora de las polimerasas implicadas en su replicación o transcripción (Steinhauer *et al.* 1992; Friedberg *et al.* 1995; Duffy *et al.* 2008). La continua aparición de mutaciones favorece la enorme adaptabilidad que presentan los virus de RNA a los cambios ambientales.

La recombinación genética implica la rotura de una parte de la cadena de material genético que a continuación se inserta en otra región diferente. En un contexto viral, ocurre generalmente cuando la polimerasa RNA dependiente de RNA (RdRp) para de copiar una cadena de RNA y la transfiere a otra durante la síntesis de una nueva cadena en el proceso conocido como 'intercambio de plantillas' (Lai 1992; Hill *et al.* 1997). Este mecanismo se observa de forma desigual entre los distintos virus de RNA (Makino *et al.* 1986; Evans 1999; Enjuanes *et al.* 2005; Barr *et al.* 2010) y parece ser una de las principales fuerzas evolutivas responsable de la pandemia del virus de la inmunodeficiencia humana (VIH) (Nájera *et al.* 2002; Rhodes *et al.* 2003; Prljic *et al.* 2004).

Los reordenamientos genómicos son cambios estructurales que eliminan, añaden o reordenan grandes regiones de material genético. Este mecanismo es importante en los virus fragmentados, aquellos que presentan el genoma segmentado, como el virus de la gripe, y en virus multipartitos,

aquellos en los que cada fragmento genómico se localiza en partículas virales separadas, como los bromovirus. Este mecanismo se reconoce como el principal causante de los cambios antigénicos en el virus de la gripe (Ferguson *et al.* 2003).

Dada la enorme variabilidad existente en las poblaciones de virus tipo RNA, estas son consideradas como quasi-especies moleculares (Eigen *et al.* 1988), conjunto con una gran heterogeneidad genética de variantes estrechamente relacionadas sobre el que actúan los distintos procesos evolutivos (Cuevas *et al.* 2005). La deriva genética (muestreo aleatorio de los genotipos de una generación a la siguiente) y la selección natural (que aumenta la proporción de variantes con una alta adaptabilidad) son los principales mecanismos evolutivos mediante los cuales disminuye la variabilidad en una población. Además, existen procesos importantes en la dinámica evolutiva de las poblaciones de los virus de RNA como son el trinquete de Muller, el principio de exclusión competitiva y la hipótesis de la Reina Roja, entre otros (Moya *et al.* 2000).

El trinquete de Muller (Muller 1964) hace referencia a la pérdida de eficacia biológica como consecuencia de la acumulación de mutaciones deletéreas en poblaciones asexuales que experimentan reducciones en el tamaño poblacional. Los mecanismos de propagación de las poblaciones virales se ajustan a condiciones bajo las que puede actuar el trinquete de Muller (Duarte *et al.* 1992; Elena *et al.* 1996; Yuste *et al.* 1999; Novella *et al.* 2004).

Según el principio de exclusión competitiva (Gause 1934), cuando dos poblaciones entran en competición y sus nichos son completamente solapantes, una de ellas elimina o excluye a su oponente en su favor. Durante el proceso de competición entre ambas poblaciones se observa un fenómeno conocido como la hipótesis de la Reina Roja (Van Valen 1973). Este proceso consiste en la ganancia absoluta de eficacia en ambas poblaciones aunque sin variación relativa. Ambos procesos han sido comprobados en poblaciones de virus de RNA (Clarke *et al.* 1994; Quer *et al.* 1996).

La actuación de todos estos mecanismos en la dinámica evolutiva y poblacional de los virus conducen a la aparición de resistencias a tratamientos. El estudio de su implicación en la interacción paciente-patógeno es importante para entender el mecanismo de la enfermedad y disponer de estrategias terapéuticas alternativas (Bonhoeffer *et al.* 1997; Domingo *et al.* 1997; Moya *et al.* 2004; Nuñez *et al.* 2005).

En el estudio de los aspectos evolutivos, poblacionales y epidemiológicos de diversas enfermedades se emplean como biomarcadores las secuencias nucleotídicas de los patógenos. Para la realización de estos estudios se requiere una gran cantidad de muestras, dada la enorme variabilidad que existe entre estos microorganismos (variación intrapaciente), el gran número de pacientes afectados y la variabilidad de los virus entre los diferentes pacientes (variación interpaciente). Además, las técnicas experimentales actuales como los métodos basados en la reacción en cadena de la polimerasa (PCR) junto con las técnicas de secuenciación (Tarasevich *et al.* 2003), genotipado (van Belkum *et al.* 2001), los *microarrays* (Jansen *et al.* 2006) y las últimas tecnologías de *next-generation*

sequencing (NGS) (Ansorge 2009; Nakamura *et al.* 2009; Metzker 2010; Sorek *et al.* 2010), entre otras (Zaidi *et al.* 2003) permiten obtener una gran cantidad de secuencias de material genético, por lo que una buena organización de toda esta información es esencial para llevar a cabo estos estudios.

1.4 La hepatitis C

La hepatitis C es una enfermedad causada por el virus de la hepatitis C (HCV) que afecta al hígado en humanos. Puede manifestarse de forma aguda y resolverse espontáneamente (hepatitis aguda) o evolucionar a crónica, aunque la hepatitis crónica no siempre es detectable en fase aguda. La hepatitis aguda se caracteriza por un aumento en los niveles de alanina aminotransferasa (ALT), una molécula presente en el suero sanguíneo indicadora del daño hepático. Aproximadamente un tercio de los adultos con hepatitis aguda desarrollan también síntomas clínicos e ictericia. Esta forma de la enfermedad puede ser grave, aunque raramente fulminante (Farci *et al.* 1996), y prolongarse durante 6 meses. En algunos casos la infección remite por sí sola, aunque un 55-85% de los pacientes infectados suelen desarrollar hepatitis crónica. La hepatitis crónica es a menudo asintomática aunque suele asociarse a altos niveles de ALT, y evoluciona hacia una fibrosis progresiva (sustitución del tejido hepático por tejido fibroso en el que hay un marcado aumento de matriz extracelular), cirrosis (atrofia del hígado) y hepatocarcinoma (cáncer de hígado) (Hoofnagle 2002). Además, el HCV está estrechamente relacionado con el desarrollo de determinadas enfermedades extrahepáticas como la crioglobulinemia, caracterizada por la activación de

linfocitos B y la producción de autoanticuerpos (Sansonno *et al.* 2005; Charles *et al.* 2009).

1.4.1 Epidemiología

La Organización Mundial de la Salud (OMS, www.who.int) estima que aproximadamente 170 millones de personas en todo el mundo, un 3% de la población mundial, se encuentran infectadas con el HCV. De ellos, 130 millones son portadores crónicos con riesgo de desarrollar cirrosis y cáncer de hígado. Se estima que entre 3-4 millones de personas se infectan nuevamente cada año, el 70% de las cuales desarrollará hepatitis crónica. El HCV es responsable de entre un 50-76% de los cánceres de hígado y de dos tercios de los trasplantes que se practican en el mundo. En la Figura 1 se puede observar la prevalencia (número total de individuos afectados en un momento determinado) de esta enfermedad en todo el mundo, siendo superior en los países menos desarrollados. Los datos más alarmantes se encuentran en varios países africanos, siendo Egipto donde alcanza su máximo con un 14,5% de población afectada. Respecto a la Comunidad Valenciana, la incidencia (número de casos nuevos en un momento determinado) de esta enfermedad en el año 2009 fue de un 0,57 por 100.000 habitantes y hubo una prevalencia de un 3,95 por 100.000 habitantes (Dirección General de Salud Pública 2009).

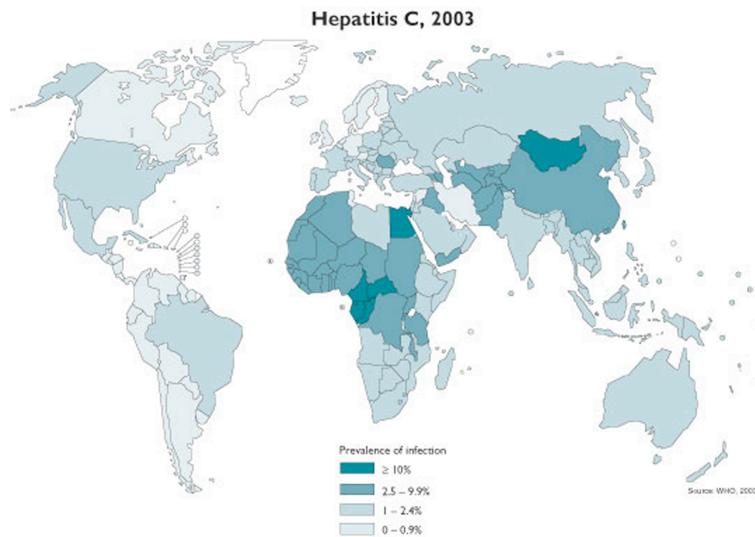


Figura 1. Prevalencia de la infección de la hepatitis C en 2003 (OMS).

El virus se contagia de persona a persona por vía parenteral, principalmente por la exposición directa o indirecta a sangre contaminada, bien por transfusiones de sangre, consumo de drogas por vía intravenosa, empleo de material quirúrgico mal esterilizado, etc. (Shepard *et al.* 2005; Sy *et al.* 2006).

El HCV pertenece al género *Hepacivirus* dentro a la familia *Flaviviridae* (Robertson *et al.* 1998). Se han identificado hasta seis genotipos y más de 50 subtipos, cuya distribución varía geográficamente y entre los grupos de riesgo (Simmonds *et al.* 2005). El genotipo 1a es el genotipo prototípico y es común en Estados Unidos y en el norte de Europa. El genotipo 1b se distribuye por todo el mundo, siendo el más común. Los genotipos 2a y 2b también se distribuyen por todo el mundo y causan entre el 10% y el 30% de los casos de hepatitis C, siendo particularmente comunes en Japón y el norte de Italia. El

genotipo 3 es el más frecuente en la India y su introducción en Estados Unidos y Europa parece ser reciente, como resultado del uso de drogas por vía intravenosa entre los años 60 y 70. El genotipo 4 es el más común en África y es extremadamente variable. Los genotipos 5 y 6 se encuentran en áreas geográficamente aisladas, Hong Kong y el sudeste de Asia, respectivamente (Hoofnagle 2002).

1.4.2 Estructura del genoma viral y proteínas virales

El genoma del HCV es una cadena sencilla de RNA y polaridad positiva que codifica una poliproteína de unos 3000 aminoácidos que, tras su procesamiento mediante proteasas virales y celulares, da lugar a 10 proteínas distintas (Penin *et al.* 2004; Dubuisson 2007). Las proteínas no estructurales (p7, NS2, NS3, NS4A, NS4B, NS5A y NS5B) se obtienen tras el corte de la poliproteína mediado por las proteasas virales NS2-3 y NS3-4A, mientras que las proteínas estructurales (proteína *core*, E1 y E2) son liberadas por proteasas de señal en el retículo endoplasmático (RE) del huésped (Figura 2).

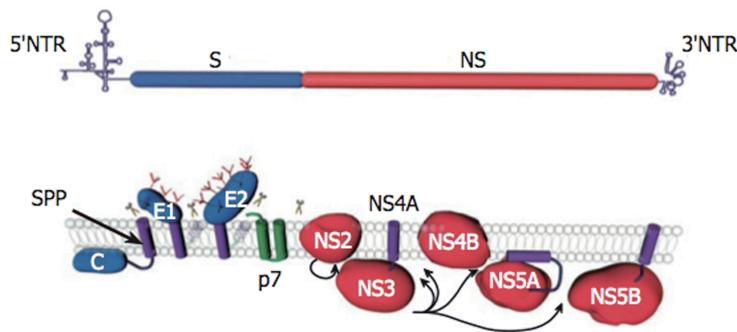


Figura 2. Arriba: Organización del genoma del HCV. S corresponde a la región que codifica proteínas de función estructural y NS corresponde a la región que codifica proteínas de función no estructural. Abajo: Procesamiento de la poliproteína y localización de las proteínas relativa a la membrana del RE. Imagen tomada de Dubuisson (2007).

La proteína C o *core* es una proteína de unión a RNA y se considera el único componente de la nucleocápside viral. Se cree que está implicada en la señalización celular, apoptosis, carcinogénesis y en el metabolismo lipídico. Las proteínas E1 y E2 son proteínas transmembrana de tipo I que forman un heterodímero y constituyen las proteínas de envuelta del virus. Son esenciales para la entrada del virus en la célula y participan en el ensamblaje de nuevas partículas virales. Además, se han identificado tres regiones hipervariables (HVR) en la proteína E2 que parecen ser las dianas de neutralización de los anticuerpos generados como respuesta a la infección viral (Weiner *et al.* 1991; Kato *et al.* 1992; Troesch *et al.* 2006; Torres-Puente *et al.* 2008c). Su extremada variabilidad podría dificultar la respuesta inmune del paciente (Taylor *et al.* 1999).

La proteína p7 se encuentra localizada entre la región de proteínas estructurales y la región de las no estructurales. Se ha visto que está implicada en canales iónicos de membranas

lipídicas artificiales y que es esencial para la infectividad del virus en chimpancés (Steinmann *et al.* 2007).

En cuanto a las proteínas no estructurales, la proteína NS2 participa en la proteólisis de la poliproteína en la unión NS2/NS3. La proteína NS3 parece presentar diversas actividades enzimáticas tales como actividad serina proteasa, NTPasa y RNA helicasa, que se encuentran potenciadas por el cofactor NS4A. Antes de ser liberada de la poliproteína, la proteína NS3 cataliza la proteólisis en el sitio de unión NS2/NS3, junto con la proteasa NS2. La proteína NS4B parece estar implicada en la inducción de alteraciones de membranas intracelulares, lo que sugiere su participación en la replicación del RNA. La proteína NS5A es esencial en la replicación del genoma viral y presenta una región denominada región determinante de la sensibilidad a interferón (ISDR) que parece modular la respuesta al interferón. Además, se ha visto que interacciona con numerosos componentes de rutas de señalización, como la inhibición de la proteína quinasa R (PKR), un supresor tumoral y regulador de la expresión génica celular en apoptosis, contribuyendo a la oncogénesis del HCV (Gale *et al.* 1999). Por último, la proteína NS5B es una polimerasa dependiente de RNA que sintetiza nuevas cadenas de genoma viral. Se ha demostrado la interacción de NS5B con NS3 y NS4A, así como con otros componentes del complejo de replicación, interacciones que intervendrían en la regulación de la actividad polimerasa.

1.4.3 Tratamiento y variabilidad del genoma viral

En el tratamiento de la hepatitis C crónica se emplean actualmente dos drogas antivirales: interferón (IFN) y

ribavirina (RBV). El tratamiento con IFN es efectivo en el 39% de los pacientes (Zeuzem *et al.* 2000; Lindsay *et al.* 2001); sin embargo, combinado con RBV supera el 60% (Ghany *et al.* 2009). Dentro de la respuesta con tratamiento combinado existen diferencias según se trate de un tipo u otro de virus. En el caso del genotipo 1, alrededor del 50% de los pacientes responden al tratamiento y en el caso de los genotipos 2 y 3, un 80% de los pacientes muestran respuesta (Alter 2006; Ghany *et al.* 2009; Pang *et al.* 2009). Además, el coste del tratamiento para la hepatitis crónica es elevado (Wong *et al.* 2003; Shepherd *et al.* 2005), tiene numerosos efectos secundarios y no es apropiado para algunos pacientes (Shepard *et al.* 2005; Manns *et al.* 2006; Patel *et al.* 2006).

La principal preocupación en la terapia anti-HCV es la aparición de virus resistentes a fármacos. Estos aparecen rápidamente bajo la presión selectiva que ejercen las drogas antivirales. La alta tasa de replicación y la baja fidelidad en la maquinaria de replicación permiten al virus explorar su espacio genómico y adquirir mutaciones que aumenten su resistencia a las drogas antivirales. Se estima que cada nuevo virus generado muestra, como media, un cambio de nucleótido por ciclo de replicación. La identificación de las mutaciones específicas y los patrones genéticos virales responsables de los distintos fenotipos clínicos podría mejorar el diagnóstico y tratamiento de los pacientes (González-Candelas *et al.* 2010).

Diversos trabajos han puesto de manifiesto que la variabilidad genética del HCV hace posible su escape de la respuesta inmune del paciente, teniendo como consecuencia la cronificación de la enfermedad, y han tratado de establecer su relación con la respuesta al tratamiento (Farci *et al.* 2002; Guo

et al. 2004; López-Labrador *et al.* 2004; Schinkel *et al.* 2004; Layden-Almer *et al.* 2005; Pawlotsky 2006; Sklan *et al.* 2009; Bittar *et al.* 2010).

La variabilidad observada en el HCV se distribuye de forma diferente a lo largo del genoma viral y afecta de forma distinta al tratamiento en cada región genómica (Maekawa *et al.* 2009). Se ha establecido que cuanto mayor sea la presión inmune en una región, mayor será su variabilidad genética (Reed *et al.* 2000) por lo que la mayoría de estudios relativos a la relación entre la variabilidad y el tratamiento se centran en estas regiones, aunque se ha sugerido que el nivel global de variabilidad en el genoma es el que influye en la respuesta al tratamiento (Cuevas *et al.* 2008a).

1.5 Sistemas de información en epidemiología molecular

La información empleada en la realización de estudios sobre Epidemiología molecular es, por una parte, la referente a los pacientes, como sus datos demográficos e historial clínico, y, por otro lado, la información sobre los biomarcadores elegidos, como por ejemplo las secuencias genéticas del patógeno de interés. En los laboratorios de tamaño medio esta información se encuentra almacenada habitualmente de forma separada e independiente en distintos programas informáticos, lo que dificulta su manejo y análisis. Los programas más populares en estos estudios son las hojas de cálculo o programas como EpiInfo (Dean *et al.* 1991; Ma *et al.* 2008) y EpiData (Lauritsen 2000) para la información relativa a los pacientes y ficheros de texto de uso interno para las secuencias genéticas obtenidas en el laboratorio. Este flujo de trabajo pone

de manifiesto la necesidad de una infraestructura para el almacenamiento y análisis de información sistematizado y común para todos los investigadores, de forma que la gestión de los datos disponibles sea eficiente y facilite el trabajo de investigación.

Las soluciones comerciales más extendidas son los sistemas de gestión de información de laboratorio (LIMS), que son aplicaciones diseñadas para seguir el procesado de las muestras, almacenar los datos generados en el laboratorio y producir informes con esta información. Estos programas suelen ser demasiado costosos y complejos para las tareas que realizan, ya que abarcan tareas generales a muchos laboratorios y no están diseñados a medida (Morris *et al.* 2008). Dado que los laboratorios pequeños o medianos no suelen disponer de un departamento de bioinformática ni de presupuesto, tiempo y esfuerzo para implantar un LIMS comercial, el desarrollo de herramientas de libre distribución técnicamente accesibles en cuanto a su instalación y uso cobra vital importancia en el área de Epidemiología molecular. Por otro lado, el éxito de las soluciones bioinformáticas en entornos de investigación depende de su fiabilidad, flexibilidad, facilidad de uso y transparencia hacia el usuario final (Kumar *et al.* 2007a).

2. OBJETIVOS

2 Objetivos

La realización de estudios en Epidemiología molecular requiere el manejo y análisis de grandes cantidades de datos recopilados de diversas fuentes. La integración de toda esta información es de vital importancia para poder llevar a cabo estos trabajos de forma eficiente. La principal motivación de esta tesis es cubrir la necesidad de un sistema de gestión de forma sencilla, flexible y fiable. Para ello se han planteado los siguientes objetivos generales:

- Desarrollo de una base de datos para el almacenamiento de datos clínicos y demográficos relativos a los pacientes y de datos genéticos y epidemiológicos relativos a los patógenos.
- Desarrollo de una plataforma de integración de la información almacenada junto con herramientas de análisis específicos para los datos.
- Desarrollo de nuevas herramientas de análisis evolutivo y poblacional para volúmenes masivos de datos.

La plataforma desarrollada tiene como objeto cualquier tipo de patógeno aunque para las aplicaciones realizadas en esta tesis se ha empleado únicamente el HCV. Los objetivos específicos en estas aplicaciones son los siguientes:

- Desarrollo de un modelo predictivo de la respuesta al tratamiento con IFN y RBV en pacientes afectados de hepatitis C con genotipo 1, subtipos 1a y 1b.
- Estudio evolutivo y poblacional de un brote de HCV en la Comunidad Valencia.

3. DISEÑO E IMPLEMENTACIÓN

3 Diseño e implementación

3.1 Especificación de requisitos

El propósito de esta especificación de requisitos es definir las necesidades de la plataforma bioinformática que se va a desarrollar en esta tesis con el objetivo de gestionar y analizar información en epidemiología molecular.

La plataforma debe tener las siguientes características:

- Ser de código abierto y libre distribución. Las ventajas de este tipo de *software* son enormes para la comunidad científica ya que permite el libre acceso y la adaptabilidad de un *software* especializado a cada entorno concreto.
- Almacenar información clínica y molecular de forma privada y segura. La información que se va a recoger en esta plataforma forma parte del trabajo interno de un laboratorio previo a la publicación de resultados. Por esta razón es necesario que la información sea únicamente accesible por parte de los investigadores implicados en estos estudios.
- Permitir el análisis de la información sin salir de la plataforma. La integración de herramientas de análisis ahorra el paso de modificación del formato para obtener ficheros apropiados en cada aplicación. De esta forma, el procesado y análisis de los datos es transparente de cara al usuario.
- Centralizar los datos y aplicaciones en un servidor. Con este tipo de arquitectura se facilita el acceso desde cualquier sitio con el único requisito de una conexión a internet. Por

tanto, no se requiere instalar ningún *software* en el ordenador del usuario para poder usar la plataforma. Además, el usuario se olvida de las actualizaciones ya que éstas las realiza el administrador de sistemas una única vez en el servidor. Otra ventaja de la arquitectura cliente-servidor es que permite múltiples usuarios simultáneamente. De esta forma los datos disponibles almacenados siempre estarán actualizados.

La plataforma se compone de 4 partes funcionales diferentes:

- Base de datos
- Interfaz web
- Herramientas de gestión
- Herramientas de análisis

A continuación se detallan los requisitos de cada componente.

Base de datos

- Almacenar información clínica de pacientes y molecular de patógenos. La base de datos debe recoger toda la información necesaria para realizar estudios de epidemiología molecular.
- Mantener la integridad de los datos almacenados. El sistema que permite un mayor control de referencias entre registros y una mejor recuperación de los datos en caso de fallos son las tablas INNODB.

- Adaptar la base de datos a cada usuario. Las vistas permiten un ajuste del modelo lógico a cada usuario concreto incrementando la personalización de la base de datos y la seguridad de los datos almacenados.

Interfaz web

- Integrar todas las herramientas necesarias para que el usuario gestione sus datos en la base de datos local y los pueda analizar. El hecho de disponer de todas las herramientas en un mismo entorno facilita el trabajo del investigador.
- Cumplir las recomendaciones del *World Wide Web Consortium* (W3C) respecto a los estándares de lenguajes de programación. De esta forma se garantiza que la implementación de la web sea correcta.
- Estructurar los contenidos y aplicaciones de forma sencilla e intuitiva para el usuario. La usabilidad de la plataforma mejora con una buena estructuración de los contenidos en el menú y las subsecciones.
- Permitir realizar análisis en modo *background*. En este modo, la interfaz sigue siendo operativa una vez que se envían los análisis en vez de esperar a obtener los resultados.

Herramientas de gestión

- Permitir la inserción de datos. Los formularios de inserción permiten insertar registros en cada tabla de la base de datos o en tablas del mismo módulo de forma sencilla.

- Permitir la inserción de gran cantidad de datos. En los casos en que se tengan muchos datos que insertar o que haya que insertarlos en muchas tablas distintas.
- Permitir la búsqueda de información a través de formularios sencillos.
- Permitir la realización de búsquedas a medida de cada usuario.
- Facilitar las búsquedas más comunes.
- Poder consultar todos los registros de una sola tabla.
- Recuperar las secuencias de la base de datos local en distintos formatos.
- Permitir la actualización de los registros en la base de datos.
- Permitir la eliminación de registros en la base de datos.
- Comprobar la consistencia de información que proviene de dos fuentes distintas.
- Facilitar la búsqueda de secuencias en bases de datos públicas generales.
- Facilitar la búsqueda de secuencias en bases de datos públicas específicas del HCV.
- Permitir al usuario cargar sus ficheros en el servidor.
- Mostrar al usuario los ficheros de resultados.

Herramientas de análisis

- Buscar secuencias por similitud frente a bases de datos externas, a la base de datos local y a un fichero cargado por el usuario.
- Traducir secuencias de nucleótidos en secuencias de aminoácidos.

- Realizar alineamientos múltiples de secuencias.
- Traducir alineamientos de aminoácidos en alineamientos de nucleótidos.
- Calcular parámetros poblacionales y evolutivos.
- Realizar pruebas estadísticas basadas en coalescencia.

3.2 Arquitectura de la plataforma bioinformática

La plataforma bioinformática fue diseñada siguiendo un modelo de arquitectura en tres capas o niveles (Fowler 2002). En la Figura 3 se muestra una breve descripción del SI que se desarrollará con más detalle en los siguientes apartados. En este esquema se representa la organización de los distintos componentes del *software*. En el nivel de presentación se sitúan la interfaz *web* y la línea de comandos (intérprete de órdenes de un ordenador) que posibilitan la interacción a través de una representación visual de la aplicación, proporcionando una forma de acceder y manipular los datos. En el nivel de aplicación se encuentran las herramientas de gestión y análisis de datos que responden a las peticiones del usuario para ejecutar las tareas de manejo y estudio de la información. Por último, en el nivel de persistencia se incluyen aquellos componentes que permiten a los objetos del nivel de aplicación interactuar con el almacén permanente de información.

Además, en este esquema se representa en línea continua la vía principal de uso de la plataforma y en línea discontinua la vía de uso puntual a través de la línea de comandos, orientadas hacia un usuario medio y un usuario avanzado respectivamente. Ambas vías pueden ser utilizadas

indistintamente sin que ello afecte al funcionamiento de la plataforma. En lo sucesivo, esta tesis se centra en la vía principal de uso de la plataforma.

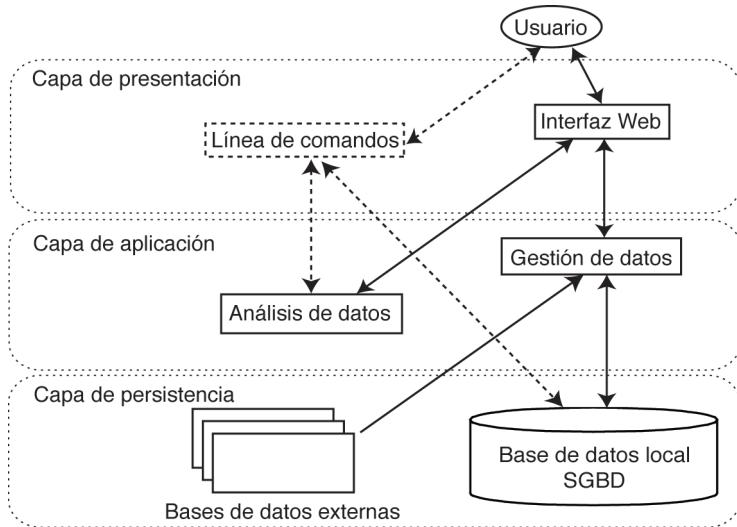


Figura 3. Esquema de la arquitectura de la plataforma bioinformática.

El acceso a la plataforma se realiza a través de la interfaz *web*, donde se encuentran centralizadas todas las herramientas implementadas. Para el diseño se tuvieron en cuenta los posibles usuarios, de modo que su acceso queda regulado en distintas partes de la plataforma (Figura 4). Los métodos de restricción empleados afectan también a la vía de uso puntual.

La primera restricción que se encuentra el usuario es la identificación con un nombre y una contraseña en la interfaz *web*. Estos datos se encuentran almacenados en el SGBD de tal forma que si un usuario no se encuentra registrado en el propio sistema de acceso al SGBD tampoco puede entrar en la interfaz de la plataforma. La interfaz se encuentra programada para

avisar al usuario en caso de que no haya introducido uno de los dos campos del formulario o los datos sean erróneos.

La segunda restricción de uso afecta a la gestión de la información de la base de datos local. El usuario debe tener permisos de acceso a la base de datos principal y únicamente el administrador de la plataforma puede proporcionarlos.

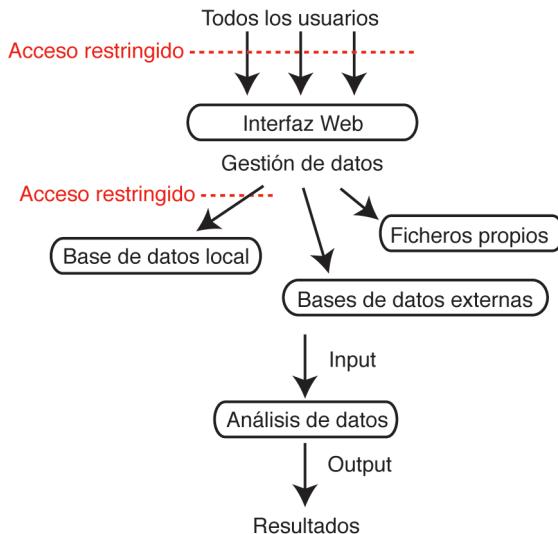


Figura 4. Esquema de acceso a la plataforma.

De este modo, sólo aquellos usuarios que se encuentren registrados en la plataforma tienen acceso a ella (primer acceso restringido). Dentro de este grupo general se diferenciaron aquellos usuarios con acceso únicamente a los datos de bases de datos externas y a los ficheros propios del mismo, de aquellos usuarios que, además, tienen acceso a la base de datos local (segundo acceso restringido). Todos ellos pueden utilizar las herramientas de análisis implementadas para analizar sus propios datos.

Se eligió una distribución física de los componentes de tipo cliente-servidor, donde el cliente (máquina remota donde trabaja un usuario) se comunica con el servidor local (máquina en la que se ejecuta el SI) por red (Figura 5). De esta forma, en el servidor local se encuentran la plataforma con la base de datos local, las herramientas de análisis, los ficheros cargados por los usuarios y los ficheros de resultados. Desde el cliente se accede a la plataforma a través de la red con un navegador y puede exportar la información almacenada. El acceso a los datos almacenados en bases de datos públicas localizadas en servidores externos se facilita desde la plataforma. Este tipo de arquitectura permite el acceso de múltiples usuarios simultáneamente.

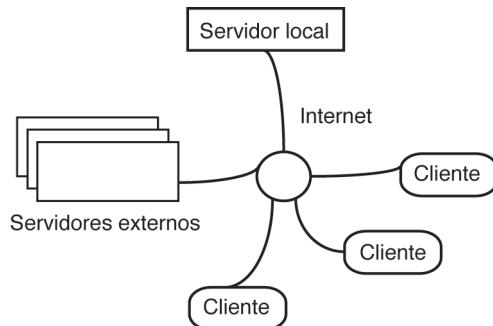


Figura 5. Esquema de la distribución física de la plataforma.

3.3 Bases de datos

El empleo de una base de datos como almacén fijo de información tiene el objetivo de proporcionar una forma práctica, eficiente y fiable para almacenar y recuperar información. Los sistemas de bases de datos son la mejor opción cuando se dispone de grandes cantidades de

información como es el caso de los estudios de epidemiología y genética molecular de poblaciones.

3.3.1 Análisis

En esta fase se identificó el ámbito de información que debía incluir la base de datos para alcanzar los objetivos propuestos. Para ello se siguió la metodología de entrevistas con las partes interesadas en la plataforma, científicos y médicos que trabajan actualmente en los campos de la Epidemiología molecular, la Genética de poblaciones y la Genética evolutiva. Además, se analizó la documentación existente sobre estándares y manuales de bases de datos similares ya existentes (Silberschatz *et al.* 2002; Nelson *et al.* 2003; Morris 2005; Araújo *et al.* 2006).

Desde el momento en que un paciente llega a un centro de salud hasta el análisis molecular del patógeno en el laboratorio, se recoge mucha información relativa tanto al estado clínico del paciente como a la epidemiología del patógeno. Toda esta información se dividió en 50 tablas organizadas en 12 módulos conceptuales sobre los que posteriormente se desarrolló el esquema de base de datos.

El flujo de información de la parte del mundo real de interés (Figura 6) comienza en el módulo de FUENTES de muestras que hace referencia a aquellos lugares donde se puede encontrar un patógeno, bien sea en pacientes o bien en entornos medioambientales. Después, se puede obtener INFORMACIÓN CLÍNICA del paciente cuando acude al médico, quien pide realizarle PRUEBAS CLÍNICAS de las que se obtienen RESULTADOS. Además, se puede obtener información sobre el

tipo de TRANSMISIÓN con respecto a cómo y por quién fue infectado el paciente y a quién pudo infectar este a su vez. Entonces, el médico ofrece al paciente un TRATAMIENTO y se recogen MUESTRAS del patógeno implicado. Las muestras siguen un PROCESADO DE LABORATORIO del que se obtienen SECUENCIAS de material genético que corresponden a un PATÓGENO específico asociado a un tipo de transmisión y, en algunos casos, a un BROTE epidemiológico. Además, las secuencias pueden analizarse con MÉTODOS FILOGENÉTICOS y ser publicadas en distintos trabajos, generando así una BIBLIOGRAFÍA.

De este estudio se obtuvo la especificación de requisitos de la base de datos principal MOLECULAR_EPIDEMIOLOGY que se encuentra como Anexo (7.1) en esta tesis.

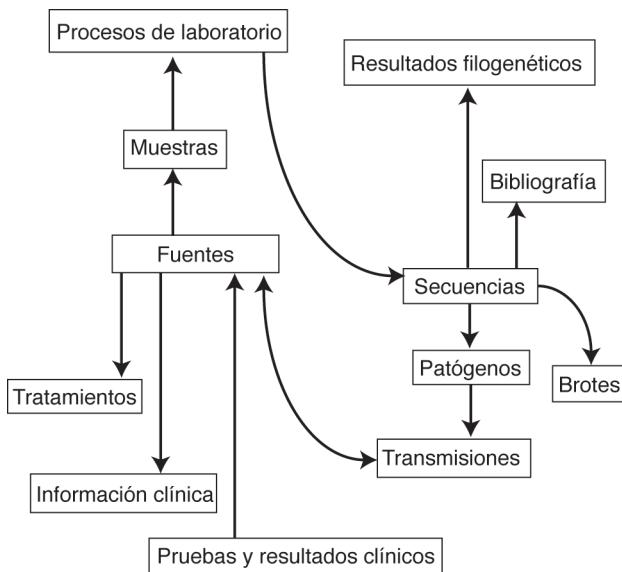


Figura 6. Diagrama de flujo de la información en la base de datos principal, *molecular_epidemiology*.

Con objeto de llevar un mínimo control de las modificaciones realizadas en la base de datos principal, se diseñó una segunda base de datos que sirve como registro de las inserciones, actualizaciones y eliminaciones de información realizadas por cada usuario en las distintas tablas de la base de datos principal desde la interfaz *web*. El acceso a esta base de datos secundaria se encuentra habilitado únicamente para el administrador de la plataforma. En la Figura 7 se observa el diagrama del flujo de información que se implementó en el esquema de base de datos secundaria, donde un USUARIO puede realizar MODIFICACIONES DE INFORMACIÓN en la base de datos principal en una FECHA concreta. Este segundo esquema de base de datos se estructuró en 7 tablas distribuidas en 3 módulos conceptuales. Del estudio y análisis de las necesidades de administración de la base de datos principal se obtuvo la especificación de requisitos de la base de datos secundaria, DBCONTROL_MOLEPI, que se encuentra como Anexo (7.2) en esta tesis.

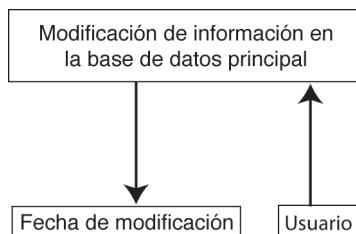


Figura 7. Diagrama de flujo de información en la base de datos secundaria, *dbcontrol_molepi*.

3.3.2 Diseño

Una vez establecidos los requisitos de los esquemas de base de datos, se desarrolló la estructura de cada esquema con el programa DBDESIGNER 4. Este programa fue elegido por ser de libre distribución y específico para el diseño de bases de datos que funcionan bajo el SGBD MYSQL, utilizado posteriormente en la implementación de la base de datos. Además, es capaz de traducir el esquema dibujado al lenguaje DDL. Los esquemas de bases de datos se encuentran como Anexo en esta tesis (7.3 y 7.4). Debido a que el esquema obtenido no especifica el tipo de cardinalidad entre las entidades, se desarrolló un diagrama ER donde se pueden observar las relaciones 1:1, 1:N y N:N (Anexos 7.5 y 7.6). A continuación, se codificaron los esquemas en lenguaje DDL y se depuró el código ‘a mano’ comprobando que se cumpliesen los requisitos especificados en la fase de análisis.

En esta tesis se presentan los esquemas definitivos desarrollados, pero hasta conseguir éstos se realizó un extenso trabajo previo de consulta y rediseño sucesivo de los esquemas.

En esta fase también se estudiaron a fondo todas las entidades y sus atributos para eliminar redundancias y conseguir la máxima simplicidad siguiendo el proceso de normalización (Codd 1970; Fagin 1977; Fagin 1981; Maier 1983). Este proceso implica realizar una reflexión sobre cada entidad, sus atributos y sus relaciones con otras entidades para establecer el esquema más optimizado posible.

3.3.3 Implementación

Se instalaron ambos esquemas de base de datos en el SGBD de MYSQL, que fue elegido por ser de libre distribución y tener un amplio uso en el ámbito académico, además de por su fiabilidad y flexibilidad para funcionar bajo distintos sistemas operativos. En este caso, se implementó en un sistema Linux Fedora. Las tablas utilizadas en la implementación fueron de tipo INNODB porque permite transacciones, bloqueo de registros e integridad referencial (condición que garantiza que una entidad siempre se relacione con otras entidades válidas, es decir, existentes en la base de datos). Además, este tipo de tablas ofrece una fiabilidad y consistencia muy superior a MYISAM, otra tecnología de almacenamiento de datos disponible en MYSQL.

Una vez instalado el esquema de la base de datos principal, se implementaron las vistas individuales de los usuarios. Una vista es el resultado de una búsqueda almacenada como tabla virtual que no forma parte del esquema físico, es decir, una abstracción de los datos (Figura 8). Las vistas permiten a los distintos usuarios acceder y manejar diferentes grupos de datos manteniendo la independencia de los mismos. De esta forma, se restringe el acceso e incrementa la seguridad de los datos almacenados, además de adaptar el modelo lógico a cada usuario concreto.

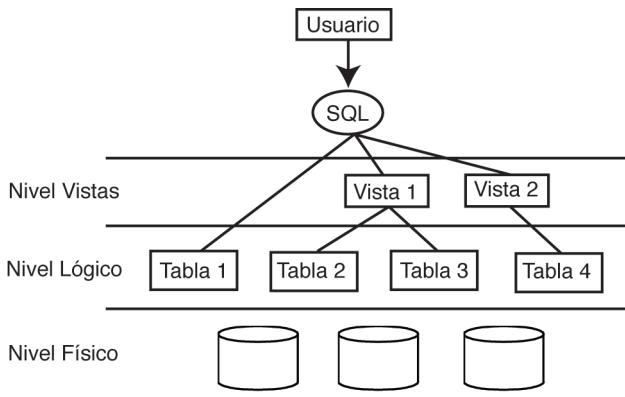


Figura 8. Esquema de los niveles de abstracción de datos en 3 capas.

El sistema de gestión de vistas utilizado en esta plataforma viene incluido en el SGBD MYSQL versión 5 o superior y se diseñaron tres niveles de vistas. El primer nivel de vistas (*general tables views*) incluye una vista por tabla para aquellas tablas que almacenan datos generales, donde cada usuario tiene únicamente permiso de búsqueda. El segundo nivel (*user project views*) incluye una vista por tabla incluida en la base de datos, donde cada usuario tiene los permisos para buscar, insertar, actualizar y borrar sus propios datos, definidos por el atributo ‘proyecto’ de cada tabla. Y, por último, un tercer nivel de vistas (*info patients views*), en las que hay una vista por tabla que incluye información clínica de los pacientes compartidos por diferentes usuarios, donde cada usuario tiene únicamente permiso de búsqueda. Además de estos tres niveles de acceso a la información, se implementó un cuarto nivel de vistas (*select views*) compuesto por una vista de nivel 1 más una vista de nivel 2 o una vista de nivel 1 más una vista de nivel 3 dependiendo de la tabla, que se emplea en la herramienta de búsquedas de la interfaz web.

3.4 Interfaz web

La comunicación entre el usuario y el SGBD se realiza por defecto a través de un terminal o línea de comandos. Dado que éste no es un sistema sencillo para un usuario normal, se diseñó e implementó una interfaz vía navegador *web* más amigable y sencilla de utilizar (Chanzá 2008). En la interfaz *web* se integraron las herramientas de gestión y análisis de datos. Todas estas herramientas se desarrollan con más detalle en los próximos apartados de esta tesis. En el actual ciclo de vida de la plataforma se incluyen dos versiones estables liberadas en las que la interfaz *web* ha ido evolucionando. En esta tesis se describe la interfaz de la última de ellas (versión 2).

3.4.1 Análisis

Se definieron las necesidades del usuario final y las características que debía tener la interfaz. El principal objetivo fue integrar distintas aplicaciones de forma que se encuentren centralizadas en el servidor local y el flujo de información entre ellas y hacia el usuario sea sencillo, transparente y rápido. Para ello, se utilizó un diseño basado en formularios que permitiese al usuario manejar la información y las aplicaciones de una forma fácil e intuitiva.

Con el sistema de formularios, el usuario rellena la información y pulsa un botón de envío para remitir un mensaje al servidor, que ejecuta un programa de aplicación. Este programa puede ser del grupo de las herramientas de análisis, y por tanto devolver al usuario los resultados, o bien puede ser del grupo de herramientas de gestión de la base de datos local, y

por tanto ejecutar una transacción en el servidor MYSQL para, posteriormente, devolver los resultados al usuario.

En la Figura 9 se observa el flujo de información en la interfaz de la plataforma. Cada rectángulo representa una sección del menú en la aplicación *web*. Normalmente, el usuario inicia su sesión a través de la PÁGINA INICIO con un nombre y contraseña. Una vez identificado accede a la PLATAFORMA donde selecciona el tipo de datos a utilizar. Puede manejar sus datos almacenados en la *EPIPATH DB* local, buscar información en bases de datos externas (*EXTERNAL DB*) y cargar ficheros de datos locales en el servidor (*LOCAL FILES*). Toda esta información la puede guardar como *input* para las herramientas de análisis en la sección *USER'S FILES*. A la hora de lanzar una aplicación en el servidor (*ANALYSIS TOOLS*), los datos de entrada los puede tomar de las secciones *USER'S FILES*, *LOCAL FILES* o directamente copiando y pegando (*COPY&PASTE*). Finalmente, los resultados del análisis o de la consulta a la base de datos local se muestran en la sección *RESULTS*.

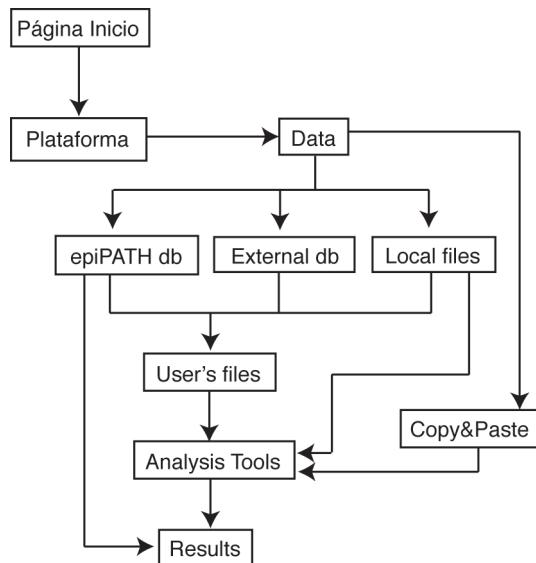


Figura 9. Diagrama de flujo de la versión 2 de la interfaz *web*.

3.4.2 Diseño

Esta fase comenzó con la maquetación de la *web* a partir del análisis realizado previamente. En este proceso se aplicó un formato a todos los elementos que componen la interfaz, como secciones, campos de formulario, imágenes, textos, etc. con el programa ADOBE FIREWORKS CS3. Durante esta etapa se obtuvieron los bocetos que daban una idea del aspecto visual que tendría la interfaz.

Una vez se obtuvo la estructura de la interfaz, se programó la *web* en los lenguajes XHTML 1.0, CSS y PHP con *scripts* (conjunto de instrucciones) en JAVASCRIPT (Figura 10). La programación se realizó utilizando el programa BLUEFISH. A pesar de que el proceso de adaptación de la versión 1 a la 2 fue muy laborioso, se logró mejorar y estandarizar el código de la interfaz siguiendo las recomendaciones del W3C.

epiPATH Analysis Platform for Molecular Epidemiology of Infectious Diseases

DATA >

Welcome, alicia :: If your username is not alicia, please logout

LOGOUT

DATA

This is the first step of the platform, data must be selected to be analyzed. There are three different ways to select data: (1) if you have access to epiPATH database server you can edit, search and retrieve your own data, (2) if you want to retrieve data from external databases you can select the external database and perform your own search and (3) if you have your data in a plain text file such as .fasta or .txt you can upload it to this platform.

1. epiPATH database [top]

epiPATH platform is distributed with a database schema that could be implemented locally. This database schema collects clinical, epidemiological and molecular information of different pathogens and users could administer their own data with different tools. The tools that this platform implements are those to add, retrieve, delete and update data. Contact with your System Administrator if you want to have access to epiPATH database.

alicia has access to epiPATH database.

Click on the appropriate tool to manage your data.

Insert Data Tool
Search Data Tool
Update Data Tool
Delete Data Tool

2. External databases [top]

To retrieve sequences from public databases, please select the database you want to search. Depending on the external database, (1) the external search form will be opened in a new window or (2) a local form will appear at this section.

This tool would search data remotely at the selected database server. Once you have retrieved data, please save your result file in your local computer. To continue analyzing these sequences, please upload your text file to the platform at 'Other data' section.

(1) - External databases that are included in this type of search system are EMBL-SRS (option 'EMBL - Standard Query Form'), LANL-HCV (option 'LANL - HCV') and euHCVdb (option 'euHCVdb') databases.
(2) - External databases that are included in this type of search system are EMBL (option 'EMBL - IDs') and NCBI (options 'NCBI - Query Form' and 'NCBI - IDs') databases.

Select an external database to search (Choose one) ▾

3. Other data [top]

To upload your data, please submit a text file. Your file will be stored in a temporary folder at our local server.

Select a data file

[] Examinar...
Load data Reset

Last modified :: 24 Jul 2008

Optimal visualization with Firefox

UNIVERSITAT DE VALÈNCIA
Institut Cavanilles de Biodiversitat i Biologia Evolutiva

MADE WITH

Figura 10. Versión 2 de la interfaz web.

En el menú general se incluyen tres secciones separadas claramente: DATA, ANALYSIS TOOLS y RESULTS, en las que se agrupan las distintas aplicaciones para facilitar el acceso y el flujo de información.

3.4.3 Implementación

En esta fase se cargaron los archivos de la interfaz *web* en el servidor local y se comprobó que todo el código programado tuviese el aspecto deseado y funcionase correctamente. Para ello, se instaló y configuró previamente el servidor HTTP APACHE junto con las librerías de los lenguajes utilizados en la programación y se realizaron las pruebas necesarias y la depuración del código. El servidor HTTP APACHE fue elegido por su libre disponibilidad, fiabilidad y flexibilidad para funcionar en distintos sistemas operativos. En este caso, se implementó en un sistema Linux Fedora. Los detalles sobre la instalación y configuración, tanto del servidor HTTP APACHE como de las librerías, se encuentran en el manual distribuido con la versión de la plataforma correspondiente.

A continuación se describen aquellas características generales que se implementaron en la interfaz *web*.

Una vez se accede a la plataforma *web*, la sesión del usuario se mantiene abierta bien hasta que se pulse *logout* o bien hasta que transcurra el tiempo programado de caducidad para una sesión inactiva. La sesión se renueva cada vez que el usuario carga una página de la interfaz. De este modo, la interfaz vuelve a la página de inicio tras el tiempo programado de inactividad. La opción *logout* destruye la sesión del usuario actual y hace que la interfaz vuelva a su página de inicio.

Por otro lado, en la versión 2 se añadió la opción de enviar los procesos al servidor en modo *background* en todas las aplicaciones salvo en las de gestión de la base de datos local. De esta forma, el usuario puede elegir cómo lanzar las aplicaciones

en el servidor. En el modo *background off*, la opción por defecto, la interfaz queda esperando a que termine el proceso iniciado en el servidor remoto. En cambio, en el modo *background on* la interfaz no espera a que el proceso enviado al servidor termine y el usuario puede continuar utilizándola mientras tanto. Esta segunda opción está indicada para aquellos procesos que van a ser prolongados en el tiempo, donde el análisis continúa en el servidor de forma independiente al funcionamiento de la interfaz. El código empleado para lanzar los procesos *background* en el servidor remoto puede verse en el Cuadro 1.

Cuadro 1. Código del proceso en *background* en la versión 2 de la plataforma.

```
/* Background process */
function bg_process($path,$file,$args) {
    if (substr(PHP_OS, 0, 3) == 'WIN') {
        $proc = popen('start /b "'.$path.'\"'.$file.'" "'.$args, 'r');
    } else {
        $proc = popen($path.'/'.$file.'.'.$args, '2>
$_SESSION["TEMPDIR"].'/delete &', 'r');
    }
}
```

3.5 Herramientas de gestión de datos

La gestión de la información es una parte fundamental de la plataforma, puesto que su objetivo es facilitar y estandarizar el manejo de los datos antes y después de ser analizados. Las herramientas que se incluyen en este apartado son aquellas implementadas para la gestión de la base de datos local, la búsqueda de información en bases de datos externas y la gestión de ficheros propios del usuario y los ficheros de resultados. Además, las tareas que realizan las herramientas de gestión de la base de datos local pueden realizarse también a

través de la línea de comandos directamente en el SGBD de MYSQL, aunque para ello son necesarias algunas nociones del lenguaje DML de SQL.

3.5.1 Gestión de la base de datos local

El objetivo de las herramientas de gestión de la base de datos local es permitir al usuario una interacción fluida con el SGBD para gestionar completamente sus propios datos con total independencia y de forma sencilla.

La integración de la base de datos en los archivos XHTML se realizó mediante la implementación de código PHP, que permite ejecutar consultas SQL dentro del código de la *web*.

Se describen a continuación las herramientas que se programaron para facilitar la gestión de la información en la base de datos local: inserción, búsqueda, actualización y eliminación de datos.

3.5.1.1 Inserción de información

Se implementaron dos formas de introducir datos a través de la interfaz, bien cargando un fichero que contenga la información o bien rellenando un formulario *web* específico para esos datos. La primera opción se diseñó para cuando se tiene que insertar una gran cantidad de datos en una o varias tablas y la segunda opción para el caso de insertar pocos datos en cada tabla.

En el formulario de la primera opción, se elige el fichero situado en el ordenador del usuario con el formato adecuado junto con la tabla donde se van a insertar los datos. En esta sección se especificaron las instrucciones de uso,

aunque las indicaciones sobre el formato y las consideraciones en cuanto al orden de las tablas a la hora de añadir la información se encuentran en el manual de la plataforma distribuido con la versión correspondiente.

Todos los formularios incluidos en la segunda opción de inserción se agruparon por módulos conceptuales y, a diferencia del diseño previo, en la misma página *web*. De esta forma se logró centralizar todas las opciones y simplificar la navegación del usuario a través del sitio *web*. En este sentido, al inicio de la sección de esta herramienta, se implementó un menú de módulos a través del cual es posible acceder rápidamente al formulario de interés sin necesidad de desplazarse por toda la página. Además, en cada formulario se añadió información y ayuda relativa a su utilización.

Cuando se trabaja en el área de la Epidemiología molecular, la información necesaria para realizar algunos estudios viene dada por distintas fuentes como hospitales y/o laboratorios independientes entre sí. Como consecuencia, pueden aparecer algunas inconsistencias en la información guardada en la base de datos debido al origen de esta información, por ejemplo, cuando en el informe que acompaña a una muestra proveniente de un hospital aparece que dicha muestra pertenece a un genotipo concreto y posteriormente en el laboratorio el resultado del análisis es un genotipo distinto. En este caso, el esquema normalizado de la base de datos no es suficiente para asegurar la consistencia de la información, ya que el usuario es quien debe decidir sobre cuál es el dato correcto entre ambas fuentes. Para evitar en gran medida la proliferación de información ambigua, se

implementaron unos *scripts* que alertan en la interfaz *web* de la existencia de inconsistencias de este tipo en la información almacenada en la base de datos local.

El primero de ellos hace referencia al caso del ejemplo anterior, en el que se comprueba que, para una secuencia concreta, la información del tipado en la tabla SECUENCIAS es coherente con el dato en la tabla RESULTADOS DE PRUEBAS. Este código se programó en los formularios de inserción de información de los módulos SECUENCIAS, PRUEBAS Y RESULTADOS y al cargar datos desde un archivo.

El segundo *script* de este tipo comprueba que la región a la que pertenece una secuencia concreta que se encuentra almacenada en la tabla PROCESOS DE LABORATORIO sea igual a la región almacenada en la tabla SECUENCIAS. Existen dos campos ‘región’ debido a que en el proceso de obtención de la secuencia en el laboratorio se dispone de una información *a priori* y de unos resultados una vez se termina el protocolo. Por tanto, ambas deben ser consistentes y si esto no fuese así, podría deberse a algún tipo de error en el proceso. Con el código que se programó se mantiene un mínimo control del proceso de laboratorio, lo que convierte a la plataforma en un LIMS. Este *script* se implementó en los formularios de inserción de información del módulo SECUENCIAS y al cargar datos desde un archivo.

3.5.1.2 Búsqueda de información

Se implementaron cuatro formas de búsqueda de información en la base de datos local: la búsqueda sencilla a través de formularios, las consultas complejas, las consultas

complejas más comunes y los informes de datos. La primera opción se diseñó para el caso en el que se quiere buscar información relativa a un módulo concreto y las consultas complejas para cuando se quiere buscar a través de distintos módulos. En la opción de consultas complejas más comunes se programaron aquellas consultas más utilizadas y con los informes de datos el usuario obtiene una idea general de toda la información contenida en una única tabla. Esta última opción aporta una visión general de toda la información incluida en una tabla, lo que resulta muy útil para la gestión de los datos.

Al igual que en las herramientas de inserción de datos, todos los formularios de búsqueda de información fueron agrupados por módulos conceptuales en la misma página *web*, con lo que se logró centralizar las búsquedas y simplificar la navegación del usuario. En esta sección también se incluyeron las instrucciones de uso y un menú dentro de la opción de búsqueda sencilla, que permitiese acceder directamente a los formularios concretos de un módulo.

En el formulario de la opción de consultas complejas (Figura 11), además de las instrucciones de uso, se incluyó un campo para insertar la consulta excepto la palabra *select*, que se implementó directamente en el código. De esta forma se consiguió aumentar la seguridad del texto enviado al SGBD. Esta opción se diseñó para dar una total libertad al usuario en la realización de consultas a medida a través de la interfaz *web*.

En cuanto a las consultas más comunes, se programaron las búsquedas de todas las secuencias de un paciente, especificando o no la región, la búsqueda de todas las secuencias de una muestra, especificando o no la región y esta

misma consulta pero obteniendo como resultado un fichero FASTA con todas las secuencias encontradas de la muestra.

3. Complex query

Please, enter your query avoiding "SELECT" argument (it is already implemented in this tool).

Complex query form

SELECT

Figura 11. Formulario de búsqueda compleja, versión 2.

Los resultados obtenidos con la búsqueda de información a través de cualquiera de las opciones anteriores se muestran en pantalla en formato tabular y se añadieron diversas opciones que incrementan la interactividad con el usuario. En la Figura 12 se observa un ejemplo de la página resultado en la que, al comienzo, aparece la consulta realizada y, a continuación, las instrucciones para exportar los datos bien en formato EXCEL, texto o FASTA, en el caso de obtener secuencias. Cuando se elige el formato FASTA, aparece la opción de exportar las secuencias ya en pauta, es decir, con el marco de lectura correcto. Además, los registros seleccionados pueden guardarse como *input* en la plataforma para continuar utilizándolos en el servidor. A continuación, se muestra la tabla con todos los registros del resultado de la consulta realizada, que llevan una casilla a la izquierda para seleccionar cada fila y exportarla. Al final de la página aparece el número total de resultados obtenidos.

Searching data - Complex query

Your complex query is: SELECT * from select_authors_alicia limit 1,3;

Export instructions

To save results, please select each row you want to export and a format to save data. Note that if you save data in fasta format, only 'clone_id' and 'sequence' fields would be exported and if you choose the first line of the table, field names also will be saved in your file.

Text format has fields delimited by tabs and Excel format has fields inside a table. In Fasta format, clone_id is the identification value for each sequence. If clone_id is Null or blank, it would be sequence_id as identification value for each sequence.

To obtain sequences in the correct reading frame (with the value stored in codon_start field in epiPATH database), please select "Yes" at in 'Save with correct reading frame' option.

If you are going to continue analysing these results in this platform, select 'Yes' option in 'Also save as input'.

Export options

Save results as

Also save as input? No Yes

<input type="checkbox"/>	author_id	author_firstname	author_lastname1	author_lastname2	comments	project
<input type="checkbox"/>	sam02	Ref02	Ref002	Ref0002	NULL	REVIEW
<input type="checkbox"/>	sam03	Ref03	Ref003	Ref0003	NULL	REVIEW
<input type="checkbox"/>	sam04	Ref04	Ref004	Ref0004	NULL	REVIEW

Number of query results obtained: 3

Figura 12. Página resultado de una búsqueda a través de la web, versión 2.

3.5.1.3 Actualización de información

Se diseñó e implementó una herramienta de actualización de información desde la interfaz de la plataforma que permite la modificación de datos ya almacenados en la base de datos local.

En la Figura 13 se presenta el formulario *web* de esta herramienta. Al inicio se añadieron las instrucciones para actualizar la información, con un ejemplo de cómo llenar el formulario situado a continuación. Para aumentar la seguridad de la información y evitar posibles errores a la hora de la gestión masiva de datos, se implementó esta aplicación de forma que sólo se permite actualizar un campo (columna) de un registro (fila) cada vez. Además, una vez se ha enviado el formulario y actualizado la información se realizan las comprobaciones de inconsistencias, al igual que en la herramienta de inserción, y se informa al usuario de las que puedan aparecer. En esta

aplicación se implementaron los dos tipos de comprobaciones, tanto para la información del genotipo como de la región.

UPDATE DATA

This tool provides a web interface to update data in the local epiPATH database. To update data you must have permission to update. If you have not this permission, please, contact with your System Administrator. Due to database security reasons, you can update a single value of the database per update.

INSTRUCTIONS

You should select the name of the database table, the field name at the table to which your value belongs, the (identification value in the table) of your record and the new value you want to update.

How to fill in the IDs

The update data tool gives at the different IDs list boxes the fields needed to identify your record at the selected table. You must fill in the suitable fields the values that identify your record. In example:

UPDATE:

centres_p	Select a field to update (Choose one)
ID1: centre_number	1
ID2: sip	22200001
ID3:	
ID4:	
New Value:	
<input type="button" value="Update"/>	<input type="button" value="Clear"/>

Red rectangles indicate the IDs required by the table 'centres_p' and in blue rectangles are shown the fields that the user must fill in to update a field of this table. In this case, ID1 and ID2 are given but selecting other tables, ID3 or ID4 can be given, even only ID1.

Update data

Select a table to update (Choose one)	Select a field to update (Choose one)
ID1:	
ID2:	
ID3:	
ID4:	
New Value:	
<input type="button" value="Update"/>	<input type="button" value="Clear"/>

Figura 13. Herramienta de actualización de datos, versión 2.

3.5.1.4 Eliminación de información

En la plataforma también se implementó una herramienta para eliminar datos ya almacenados en la base de datos local. El formulario *web* es similar al de la herramienta de actualización de datos, con las instrucciones de uso junto con un ejemplo al comienzo y el formulario a continuación. Para evitar posibles errores a la hora de la gestión masiva de datos y

mejorar la seguridad, esta aplicación se implementó de forma que sólo se permite borrar un registro (fila) cada vez.

3.5.2 Búsquedas en bases de datos externas

Dada la disponibilidad pública de una enorme cantidad de secuencias e información relacionada con ellas, el uso de estas bases de datos se encuentra muy extendido en el ámbito de la biología molecular. Por este motivo, se incluyó un módulo de herramientas que faciliten la búsqueda de secuencias en bases de datos públicas desde la plataforma.

De los cientos de bases de datos biológicas que se encuentran actualmente operativas y mantenidas por diversas organizaciones de todo el mundo, se seleccionaron cuatro para realizar búsquedas remotas desde la interfaz *web*.

Respecto a los tres grupos de bases de datos principales, se eligieron *EMBL-Bank* y *GenBank* porque disponen de acceso a través de consultas programadas en BIOPERL (Stajich *et al.* 2002). Actualmente, DDBJ permite el acceso a través de servicios *web* pero no tiene, por el momento, acceso a través de BIOPERL. Como las tres grandes bases de datos se encuentran sincronizadas, el no implementar la base de datos japonesa no representa una pérdida de acceso a información relevante.

Por otro lado, dado que la plataforma bioinformática se aplica a un caso específico en esta tesis, se añadieron algunas de las bases de datos que almacenan información y secuencias del HCV, ya que en el grupo de investigación se venía trabajando con este virus. Existen tres grandes bases de datos específicas del HCV (Kuiken *et al.* 2006): *Los Alamos Hepatitis C Sequence Database* (LANL-HCV) (Kuiken *et al.* 2008), orientada hacia el

análisis de secuencias de DNA, epidemiología molecular e inmunología; la base de datos europea sobre hepatitis C (euHCVdb) (Combet *et al.* 2007), orientada hacia las secuencias de proteínas, estructura y análisis funcional para comprender la resistencia a nivel molecular; y la base de datos japonesa sobre hepatitis C (HVDB) (Shin-I *et al.* 2008), orientada hacia el análisis filogenético. Se eligieron LANL-HCV y euHCVdb para su integración en la plataforma porque estas bases de datos disponen de un formulario propio vía *web* para realizar consultas. HVDB se encuentra únicamente accesible a través de gráficos implementados en JAVA en su interfaz *web*, lo que dificulta su integración y la recuperación de información de forma masiva. Además, requiere un usuario remoto para poder cargar datos propios y analizarlos, lo que no se consideró viable desde el punto de vista de integración de herramientas dentro de la plataforma bioinformática.

Las aplicaciones de búsquedas externas se implementaron en la interfaz *web* en tres formatos distintos: formulario estándar, formulario por identificadores (IDs) y formulario definido por el usuario. Todas estas búsquedas se realizan directamente sobre la base de datos correspondiente sin mediar ningún tipo de programa de búsqueda por similitud, es decir, se recupera información con un resultado equivalente a la herramienta de búsquedas en la gestión de la base de datos local. Con estas herramientas se consiguió simplificar y centralizar los procesos de recuperación, análisis y comparación de secuencias almacenadas tanto en la base de datos local como en las bases de datos externas.

A continuación se explica con algo más de detalle las diferentes búsquedas implementadas en esta sección.

3.5.2.1 Formulario de búsqueda estándar

El usuario puede recuperar secuencias de las bases de datos *EMBL-Bank*, LANL-HCV y euHCVdb a través de sus propios formularios de búsqueda. Para acceder a ellos desde la interfaz, se implementó un menú donde se selecciona la base de datos externa en la que se quiere realizar la búsqueda y, en el caso de ser una de las mencionadas anteriormente, se abre el formulario correspondiente en una nueva ventana del navegador. A continuación, el usuario realiza la búsqueda que desea y obtiene los resultados en su ordenador, que deberán ser cargados en la plataforma mediante la aplicación OTHER DATA, que se verá más adelante, para continuar su análisis.

3.5.2.2 Formulario de búsqueda por IDs

El usuario puede buscar mediante los IDs de las secuencias en las bases de datos de *EMBL-Bank* y *GenBank* eligiendo las opciones EMBL-IDs o NCBI-IDs respectivamente en el menú inicial de esta sección. Una vez seleccionado uno de ellos, aparece un formulario donde se introducen los IDs bien escribiéndolos en el espacio reservado para ello o bien cargando un fichero con varios IDs. Se puede elegir el formato en el que se obtiene el resultado entre las opciones FASTA, EMBL o *GenBank*. Los formularios se implementaron en PHP y PERL junto con los módulos BIO::DB::DBFETCH y BIO::DB::GENBANK de BIOPERL respectivamente.

3.5.2.3 Formulario de búsqueda definido por el usuario

El usuario puede recuperar información de *GenBank* empleando también un formulario definido por sí mismo. Se diseñó la opción NCBI-QUERY FORM con la que, una vez seleccionada en el menú de búsquedas externas, aparecen sus instrucciones de uso y campos para añadir y eliminar opciones de búsqueda en el formulario a medida. Una vez se encuentra definida la consulta, se envía al NCBI y se recuperan los resultados en el formato elegido (FASTA, EMBL o *GenBank*). El formulario se implementó en PHP, JAVASCRIPT, PERL y el módulo BIO::DB::GENBANK de BIOPERL.

3.5.3 Gestión de ficheros propios del usuario

Además de recuperar información de la base de datos local y las bases de datos externas, el usuario puede gestionar en la plataforma sus propios ficheros de datos. Esto permite ampliar los análisis a aquellos datos que no se encuentren en ninguna de las bases de datos integradas anteriormente.

Se implementó una herramienta específica llamada OTHER DATA que permite cargar ficheros desde la interfaz. Los ficheros se almacenan en una carpeta temporal en el servidor remoto. Una vez subidos, estos ficheros pueden analizarse con las diversas aplicaciones implementadas. El formato en que debe estar la información dependerá de la aplicación con la que se quiera analizar. Por ello, el usuario tiene que editar los datos en su propio ordenador y a continuación cargarlos en el servidor para analizarlos.

Los ficheros de usuario almacenados en la plataforma se muestran en la sección USER'S FILES del menú general. Los

archivos mostrados pertenecen al usuario que inició la sesión y aparecen ordenados por carpetas con el nombre del usuario más la fecha en que fueron cargados en el servidor. Esta aplicación se diseñó con las características de añadir ficheros, ver los que se encuentran almacenados y su contenido, pero no se permite borrarlos porque el propio servidor se encarga de esta tarea al almacenarlos en una carpeta temporal. De esta forma, se consiguió garantizar la disponibilidad de espacio en el servidor.

En esta sección del menú también se implementó el almacenamiento y la gestión de los ficheros resultado obtenidos en los análisis cuando se selecciona la opción *Save as input* en la aplicación correspondiente. Esta opción se diseñó para los casos en los que se quiere continuar analizando los resultados obtenidos previamente en la plataforma.

3.5.4 Gestión de ficheros de resultados

La gestión de ficheros de resultados se diseñó e implementó de forma prácticamente idéntica a la gestión de ficheros de usuario. El acceso a esta aplicación, llamada RESULTS, se situó en el menú principal de la izquierda. Los resultados se almacenan en una carpeta temporal en el servidor remoto y son accesibles a través de la interfaz *web* o directamente en el servidor si el usuario dispone de acceso. Este último tipo de acceso depende del tipo de servidor y se encuentra gestionado por el administrador del sistema. En la interfaz *web*, se muestran los ficheros de resultados del usuario que inició la sesión ordenados por carpetas. Los ficheros son mostrados conforme se obtienen, por lo que si una aplicación

tarda bastante la página debe ser refrescada hasta que aparezcan completados. Al igual que en la herramienta anterior, el usuario puede ver y descargar los resultados que tiene almacenados pero no puede borrarlos.

3.6 Herramientas de análisis de datos

El análisis de la información es el paso inmediato tras la obtención de los datos y para ello se integró en la plataforma un segundo grupo de herramientas orientadas a solventar aquellas cuestiones que puedan plantearse desde la Genética de poblaciones y la Epidemiología molecular. Se estudiaron el flujo de análisis y las aplicaciones necesarias para llevarlo a cabo. Las herramientas de análisis implementadas fueron elegidas siguiendo el flujo de procedimientos estandarizados en el análisis de datos genéticos (Figura 14) y se describen con más detalle en los siguientes apartados de este punto.

El flujo de análisis implementado comienza con la obtención de secuencias en el laboratorio y su almacenamiento en la base de datos local. El usuario puede recuperarlas de la base de datos, bien en pauta o no, para realizar búsquedas por similitud; tanto con secuencias nucleotídicas como con secuencias proteicas frente a la base de datos NCBI (REMOTE NCBI BLAST) y con secuencias nucleotídicas frente a la base de datos local (LOCAL EPIPATH DATABASE BLAST) o frente a un archivo propio del usuario que contenga un grupo de secuencias (FILE BLAST). Posteriormente, las secuencias nucleotídicas obtenidas pueden traducirse en secuencias proteicas con el programa TRANSEQ.

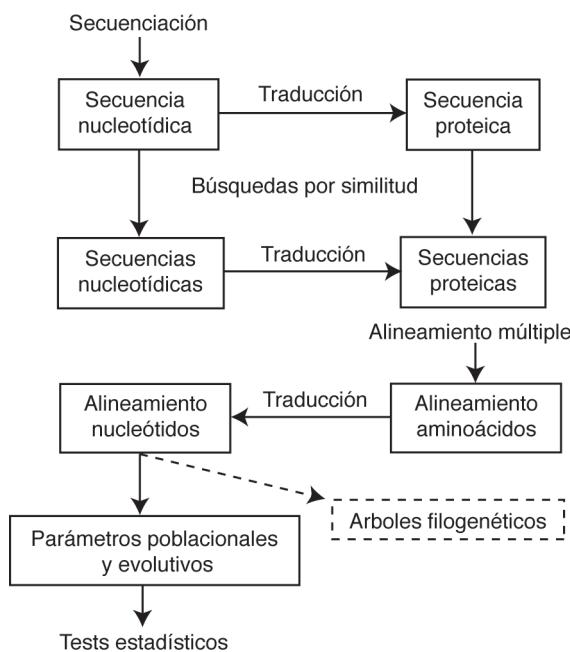


Figura 14. Diagrama de flujo de análisis de datos.

Tras obtener todas las secuencias relevantes en aminoácidos, se alinean con los programas de alineamiento múltiple de secuencias (MSA). En la plataforma se integraron MUSCLE, T-COFFEE, MAFFT y PROBCONS. Una vez obtenido el alineamiento proteico, se vuelve a traducir a nucleótidos con el programa REVTRANS. También se incluyó el programa TRANALIGN para realizar la alineación de secuencias nucleotídicas dado un alineamiento proteico.

Con el alineamiento de nucleótidos se calculan diversos parámetros poblacionales y evolutivos con DIVERGENCE, NEUTRALITY, POLYMORPHISM y SYNONYMOUS (EPIPATH-TOOLS), además de COMPUTE (LIBSEQUENCE). Sobre algunos de estos parámetros pueden realizarse pruebas

estadísticas basadas en coalescencia con la herramienta COALESCENCE. La obtención de árboles filogenéticos (representación gráfica de las relaciones de descendencia evolutiva entre diferentes variantes) no se encuentra disponible en la última versión liberada, quedando como posible mejora de la plataforma en un futuro.

Todas estas herramientas de análisis se integraron en la interfaz *web* de la plataforma, donde se organizaron en el menú lateral de la siguiente forma: en el apartado BLAST se incluyen REMOTE NCBI, LOCAL EPIPATH, FILE BLAST; en la sección EMBOSS se incluyen TRANALIGN y TRANSEQ; en la sección EPIPATH-TOOLS se incluyen DIVERGENCE, NEUTRALITY, POLYMORPHISM y SYNONYMOUS; en el apartado LIBSEQUENCE se incluye COMPUTE; en la sección MSA se incluyen MUSCLE, T-COFFEE, MAFFT y PROBCONS; en la sección STATISTICS se incluye COALESCENCE; y en el apartado UTILS se encuentra REVTRANS. Además, todas ellas pueden lanzarse a través de la línea de comandos directamente en el servidor.

Los formularios de las herramientas de análisis tienen una estructura común (Figura 15) que incluye, en aquellas aplicaciones que lo requieran, una zona para insertar los datos de entrada o *input*. Esta sección consta de tres opciones: un área de texto donde ‘pegar’ las secuencias, un menú desplegable donde seleccionar un fichero ya guardado en el servidor de la plataforma y la opción de cargar directamente un fichero desde el propio ordenador del usuario. A continuación, hay una sección con las opciones para lanzar la herramienta y el modo en que se guardarán los resultados obtenidos. Por último, se encuentra la zona donde se configuran los parámetros

específicos de cada herramienta, a la que se accede desplegando la opción *Algorithm parameters* o *Program options* según la herramienta.

The screenshot shows the '1. Remote NCBI Blast' interface. At the top, it says 'This tool searches the NCBI nucleotide database.' Below this is a large input area labeled 'Enter query sequences here in fasta format'. Underneath it, there's a section for selecting a sequence file from a temporary input folder or uploading one. A dropdown menu allows selecting the NCBI nucleotide database. There are options for 'Background mode' (Off or On), saving input and output as FASTA files, and choosing algorithm parameters like blastn, Megablast, alignments, and descriptions. The 'Expect threshold' is set to 10. There are also checkboxes for filtering low complexity regions, using a mask lookup table, and masking lowercase letters. At the bottom are 'Submit Blast' and 'Clear' buttons.

Figura 15. BLAST remoto frente al NCBI, versión 2.

3.6.1 Búsqueda de secuencias por similitud

La búsqueda de secuencias por similitud es muy importante en la investigación biológica ya que permite, entre otras cosas, determinar las relaciones evolutivas entre ellas (Attwood *et al.* 1999; Stormo 2009). Por ello, cuando disponemos de una secuencia problema es interesante encontrar otras secuencias que se le parezcan. La comparación entre cadenas biológicas mide la similitud como porcentaje de identidad, el número de bases en la secuencia de interés que

corresponden exactamente a la misma región en la secuencia de la base de datos; o puede medirse también por el grado de conservación, donde se buscan correspondencias que no alteren la función de una proteína. Estas medidas se emplean como indicio de si existe o no homología, que es la relación evolutiva que existe entre dos secuencias cuando descienden de un ancestro común.

A la hora de buscar similitud entre pares de secuencias (*Pairwise alignment*) hay dos formas de considerar el alineamiento: global, en el que se tiene en cuenta toda la extensión de las secuencias, o local, que únicamente se centra en las regiones de similitud de parte de las secuencias. Este último es el modelo que siguen algunos de los programas disponibles públicamente como BLAST (Altschul *et al.* 1990; Altschul *et al.* 1997) o FASTA (Lipman *et al.* 1985). Una búsqueda de similitud local tiene unos resultados con mayor significado biológico ya que los sitios funcionales suelen localizarse en pequeñas regiones.

Las búsquedas de secuencias en bases de datos pueden realizarse mediante consultas de texto o por búsquedas de similitud, siendo estas últimas las más habituales y empleadas. Realizar una búsqueda por similitud en una base de datos que contenga miles o millones de secuencias puede llevar un tiempo prohibitivo según el algoritmo empleado. Por ello se suelen emplear comparaciones entre pares de secuencias que utilizan alineamientos locales.

Se eligió BLAST para ser integrado en la plataforma bioinformática por ser uno de los programas disponibles públicamente más populares y que implementa de manera muy

eficiente un algoritmo de alineamiento local entre pares de secuencias. Sus características le permiten realizar búsquedas rápidas en servidores públicos de bases de datos.

A continuación se describe el algoritmo BLAST. Primero busca en la base de datos todos los segmentos, normalmente 3 residuos en proteína y 11 en DNA, que alinean con la secuencia problema por encima de un determinado umbral. A continuación, extiende los segmentos en ambas direcciones hasta que alcanza un parámetro fijado y el resultado son los llamados pares de alta puntuación (HSP), que se muestran en el *output* junto a la probabilidad de que las coincidencias no se den por azar. A mayor puntuación y menor probabilidad, la secuencia resultado es la más parecida encontrada en la base de datos.

En la plataforma que se desarrolla en esta tesis se pueden llevar a cabo tres tipos de búsquedas de secuencias por similitud con BLAST. Se implementaron las búsquedas en el NCBI mediante un acceso remoto, las búsquedas frente a la base de datos local y las búsquedas frente a un fichero cargado por el usuario que contenga un grupo de secuencias que actúe como base de datos y, frente al cual, se lanza la búsqueda mediante BLAST. Todas estas opciones se desarrollan con más detalle en las siguientes secciones.

3.6.1.1 Búsqueda remota en el NCBI

El usuario con acceso a la plataforma puede realizar búsquedas por similitud en la base de datos del NCBI desde la interfaz *web*. Para ello se integró un formulario con la configuración de los parámetros de búsqueda y los datos de

entrada para lanzar BLAST de forma remota. En la Figura 15 se muestra esta aplicación que tiene, como opciones específicas, una pestaña para seleccionar la base de datos del NCBI frente a la que se va a enviar la consulta. Una vez enviado el análisis, se obtiene un fichero de resultados en formato BLAST por cada secuencia enviada. Además, si se selecciona la opción para obtener los resultados en formato FASTA, se obtiene otro con el resultado de la búsqueda en este formato.

Esta herramienta se implementó con los lenguajes PHP, PERL y dos módulos de BIOPERL BIO::DB::GENBANK y BIO::TOOLS::RUN::REMOTEBLAST.

3.6.1.2 Búsqueda local frente a epiPATH

Se implementó la opción para realizar búsquedas por similitud en la base de datos local EPIPATH desde la interfaz web. Para ello se añadió un formulario con la configuración de los parámetros de búsqueda y los datos de entrada para lanzar BLAST de forma local. Se configuró por defecto la opción *blastn*, que busca secuencias nucleotídicas en bases de datos nucleotídicas. Una opción específica de esta herramienta es la selección de la base de datos sobre la que se realizará la consulta. En este caso se añadió *epiPATH all sequences*, que incluye todas las secuencias almacenadas en la base de datos local, pero es posible configurar tantas bases de datos como interesen. Los ficheros de resultados tienen, como en la aplicación anterior, formatos BLAST y FASTA.

Esta herramienta se implementó con los lenguajes PHP, PERL y el módulo BIO::TOOLS::RUN::STANDALONEBLAST de BIOPERL. Además, se emplearon los programas XDFORMAT para

obtener la base de datos y XGET para formatear los resultados. Ambos fueron obtenidos de la distribución de BLAST de la Universidad de Washington (WU-Blast).

3.6.1.3 Búsqueda local frente a un fichero

Por último, se implementó una tercera forma de realizar búsquedas por similitud mediante BLAST. En este caso las búsquedas se realizan de forma local frente a un fichero cargado en el servidor. En el formulario también se configuró la opción *blastn* por defecto. En este análisis, los ficheros de resultados tienen los mismos formatos descritos en las aplicaciones previas de BLAST. La implementación se realizó de igual forma que en la herramienta anterior.

3.6.2 EMBOSS

The European Molecular Biology Open Software Suite (EMBOSS) (Rice *et al.* 2000; Mullan *et al.* 2002; Olson 2002) es un paquete de programas de código abierto muy popular y empleado en biología molecular para el análisis de datos. El paquete incluye más de 100 aplicaciones susceptibles de integrarse en la plataforma bioinformática, aunque se eligieron únicamente dos programas (TRANALIGN y TRANSEQ) debido a su interés en el flujo de procedimientos analizado (Figura 14).

3.6.2.1 Tranalign

Esta aplicación alinea un grupo de secuencias nucleotídicas tomando su correspondiente grupo de secuencias proteicas alineadas. Ambos grupos deben tener sus secuencias en el mismo orden. El procedimiento que sigue se basa en la traducción de cada secuencia nucleotídica en las tres pautas de

lectura, empleando un código genético específico, y su comparación con su correspondiente secuencia proteica. El resultado es la secuencia nucleotídica que codifica para la proteína y todas ellas se devuelven en un fichero como alineamiento.

En las opciones específicas del formulario de TRANALIGN se elige el código genético que se utilizará en el análisis. Las secuencias pueden especificarse en distintos formatos, aunque únicamente si están en formato FASTA se podrán traducir varias secuencias dentro del mismo fichero. En el resto de formatos, traducirá la primera de ellas. El fichero de resultados devuelve las secuencias nucleotídicas alienadas en el formato elegido. Se utilizó el lenguaje PHP en la implementación de esta herramienta.

3.6.2.2 Transeq

Esta aplicación traduce secuencias nucleotídicas en su correspondiente secuencia proteica en los tres marcos de lectura posibles, tanto hacia adelante como en los reversos.

Se implementó en la plataforma mediante un formulario similar al anterior con lenguaje PHP. Como opciones específicas se incluyeron la selección del marco de lectura y el código genético que utilizará el programa entre otras. El resultado del análisis es un fichero en el formato elegido con la secuencia proteica en código estándar de una letra de la Asociación Internacional de Química Pura y Aplicada (IUPAC).

3.6.3 Alineamiento múltiple de secuencias

Los MSA tratan de encontrar aquellas características conservadas dentro de un grupo de secuencias de forma similar al alineamiento entre pares de secuencias (Kumar *et al.* 2007b), que emplean un tiempo de procesamiento y un espacio de memoria proporcional a la longitud de las secuencias que están siendo comparadas. Si extendemos este cálculo a tres secuencias, nos encontramos con una complejidad y tiempo de computación que aumentan considerablemente conforme se emplea un mayor número de secuencias. Por estas razones se suelen utilizar aproximaciones heurísticas, siendo el alineamiento progresivo o jerárquico, en el que se alinean primero las secuencias más similares basándose en un árbol, la técnica más utilizada (Batzoglou 2005; Wallace *et al.* 2005).

En la plataforma se implementaron aquellos programas más relevantes hasta la fecha en la obtención de MSA: MUSCLE, T-COFFEE, MAFFT y PROBCONS, todos ellos basados en el método de alineamiento progresivo.

3.6.3.1 MUSCLE

MUSCLE (Edgar 2004b; Edgar 2004a) es un programa para la obtención de alineamientos de aminoácidos o nucleótidos. Se basa en un método iterativo que mejora, respecto a los métodos progresivos, la precisión de las medidas de similitud empleadas para calcular el parentesco entre dos secuencias. El algoritmo que utiliza tiene tres etapas, en las que al final de cada una de ellas se obtiene un alineamiento múltiple y puede terminarse el cálculo.

En la primera etapa se calcula la similitud entre cada par de secuencias con las que se obtiene una matriz de similitudes. A continuación, se obtiene un árbol guía con el que se construye un alineamiento progresivo.

En la segunda etapa se mejora el árbol calculando las similitudes entre pares partiendo del alineamiento obtenido en la etapa anterior y se vuelve a construir un alineamiento progresivo basado en este nuevo árbol. Esta etapa puede repetirse cuantas veces se estime oportuno. Se comparan los árboles de la etapa uno y dos de forma que si en la etapa dos se ha obtenido más de un árbol y el número de diferencias con el primero no disminuye, se termina la iteración. A continuación, se vuelve a calcular el alineamiento progresivo donde se realinean los grupos de secuencias que hayan sufrido cambios de orden en el nuevo árbol.

En la tercera etapa MUSCLE pasa a un nuevo procedimiento para mejorar el alineamiento final. Las secuencias se dividen en dos subgrupos y se construye un ‘perfil’ (alineamiento múltiple) de cada subgrupo basado en el alineamiento múltiple actual. Estos perfiles se realinean uno respecto al otro empleando el mismo algoritmo que en la etapa progresiva del método. Si se obtiene una mejor medida de la calidad del alineamiento en este procedimiento, se mantiene como resultado el alineamiento actual y se descarta el anterior. Este procedimiento se repite con cada par de subgrupos posibles del árbol fijado. Estas iteraciones continúan hasta llegar a la convergencia o a un número máximo de iteraciones.

Se implementó con lenguaje PHP un formulario para integrar esta aplicación en la plataforma. En la sección *Program*

options se incluyeron varias opciones específicas de MUSCLE. Los resultados del análisis son un fichero con el alineamiento y otro fichero con información sobre el desarrollo del análisis.

3.6.3.2 T-Coffee

T-COFFEE (Notredame *et al.* 2000) es un programa que emplea un método progresivo con la peculiaridad de que tiene en cuenta la información de todas las secuencias en cada alineamiento de pares. Es más lento aunque más preciso que la estrategia del alineamiento progresivo (Feng *et al.* 1987) para conjuntos de secuencias lejanas.

En el primer paso del algoritmo se alinean los pares de secuencias de dos formas. Por un lado se realiza un alineamiento global de los pares con CLUSTALW (Thompson *et al.* 1994) y por otro lado se obtiene un alineamiento local de cada par utilizando LALIGN (Huang *et al.* 1991). Posteriormente, se asigna una puntuación a cada par alineado en cada método y se suman ambas para cada par.

Después se alinean las secuencias con un método progresivo y se obtiene una matriz de distancias con las que se obtiene un árbol filogenético guía. Las dos secuencias más cercanas obtenidas en éste árbol son las que se alinean primero. Para realizar este alineamiento se emplean las puntuaciones obtenidas previamente. A continuación, se alinean de la misma forma las siguientes dos secuencias más próximas según el árbol guía. Este proceso continúa hasta que todas las secuencias son alineadas.

Se implementó T-COFFEE en la plataforma mediante un formulario programado en lenguaje PHP. Los ficheros de

resultados son uno con el alineamiento y otro con el árbol filogenético empleado en el análisis.

3.6.3.3 MAFFT

MAFFT (*Katoh et al.* 2008; *Katoh et al.* 2009) es un programa de alineamiento de secuencias que ofrece diferentes métodos de MSA clasificados en tres tipos: (1) método progresivo, (2) método de refinamiento iterativo con la puntuación de la suma ponderada de pares (WSP) (Gotoh 1995) y (3) método de refinamiento iterativo con las puntuaciones WSP y de consistencia.

Este programa fue desarrollado, en un principio, para obtener un MSA con un gran número de secuencias de forma rápida (*Katoh et al.* 2002). Emplea un algoritmo de alineamiento (1) basado en la transformada rápida de Fourier (FFT) y en un método de cálculo de la distancia aproximada entre pares de secuencias basado en el método de las 6 subsecuencias compartidas (*Dayhoff et al.* 1978). En esta primera versión, el algoritmo calcula primero la matriz de distancias basándose en las 6 sub-secuencias compartidas entre cada par. Con esta matriz construye el árbol guía con el método UPGMA y a continuación, alinea progresivamente las secuencias siguiendo el orden de las ramas del árbol obtenido. A continuación MAFFT da la opción de repetir este análisis para obtener un segundo alineamiento.

Posteriormente, MAFFT mejoró la precisión del algoritmo anterior empleando un refinamiento iterativo (2) con las puntuaciones WSP (*Katoh et al.* 2005). En esta mejora, el algoritmo continúa de forma que se divide el alineamiento

obtenido previamente en dos grupos con los que se obtiene un nuevo alineamiento. Se reemplaza el anterior alineamiento si con este último se obtiene una puntuación mejor. Este paso se repite hasta que no se obtienen mejoras en la puntuación. Además, se añadió el criterio de consistencia (Notredame *et al.* 1998) al método de refinamiento iterativo (3). Este criterio refleja la consistencia entre un MSA y una librería de *pairwise alignments* con las mismas secuencias. De esta forma, la mejor puntuación se obtiene de la unión de la puntuación WSP junto con la puntuación del criterio de consistencia.

Por último, MAFFT incluye en su versión 6 el algoritmo PARTTREE (Katoh *et al.* 2007) con el que consigue mejorar la escalabilidad (capacidad de mejorar el algoritmo para poder trabajar con un mayor número de secuencias sin que se vea afectada su eficiencia) a la hora de generar el árbol guía. Con este algoritmo puede procesar unas 60.000 secuencias sin afectar excesivamente su eficiencia. Además, incluye dos métodos para obtener alineamientos de RNA en los que tiene en cuenta la información estructural de las secuencias.

Este programa se integró en la plataforma mediante un formulario desarrollado en PHP. En las opciones específicas de MAFFT se puede elegir, entre otros, el algoritmo y el formato de resultados. El resultado es un fichero con el alineamiento.

3.6.3.4 ProbCons

PROBCONS (Do *et al.* 2005) es un programa que emplea un algoritmo de alineamiento progresivo basado en pares de cadenas de Markov ocultas y que difiere de otros métodos en el empleo de la consistencia probabilística. La consistencia

probabilística es un nuevo parámetro que permite al programa predecir la probabilidad con la que el alineamiento realizado se ajusta a la realidad.

El primer paso del algoritmo consiste en el cálculo de las matrices de probabilidad posterior para cada par de secuencias. En el segundo paso se calcula la precisión esperada en cada par de secuencias y, a continuación, aplica la transformación de consistencia probabilística para re-estimar la puntuación en el par de secuencias comparadas. Posteriormente construye el árbol filogenético guía en el que se define la similitud entre dos agrupaciones por la media ponderada de las similitudes entre los pares del grupo. Después, se alinean progresivamente los grupos de secuencias definidos en el árbol guía. Por último, se realiza un refinamiento iterativo en el que se divide el alineamiento obtenido en dos grupos al azar para realinearlos. Este paso se repite las veces que se estime oportunas.

PROBCONS se integró en la plataforma de forma similar al resto de programas MSA. El formulario se desarrolló con lenguaje PHP. El resultado que se obtiene es un fichero con el alineamiento.

3.6.4 epiPATH-tools

Dentro de este grupo de herramientas se encuentran aquellas que se desarrollaron a medida en esta tesis para realizar análisis poblacionales y evolutivos desde la plataforma bioinformática. La principal característica y novedad de estas aplicaciones es que permiten calcular parámetros relativos a un gran número de secuencias de forma sencilla y relativamente

rápida. Estos programas se van a explicar con más detalle en los siguientes apartados e incluyen el cálculo de parámetros relevantes dentro de los análisis de polimorfismos (EPIPATH-POLYMORPHISM), divergencia entre secuencias (EPIPATH-DIVERGENCE), cálculo de sustituciones sinónimas y no-sinónimas (EPIPATH-SYNONYMOUS) y tests de neutralidad (EPIPATH-NEUTRALITY). Todos ellos se programaron en un sistema operativo Linux Fedora Core 4 con el entorno de desarrollo integrado (IDE) CODE::BLOCKS. Un IDE es un programa compuesto por un grupo de herramientas que facilitan el trabajo de programación y normalmente incluye un editor de código, un compilador, un depurador y un constructor de interfaces gráficas. El lenguaje de programación que se utilizó fue C/C++ junto con la versión 1.5 de las librerías Bio++ (Dutheil *et al.* 2006).

3.6.4.1 epiPATH-Polymorphism

Los polimorfismos de DNA reflejan la variabilidad que existe en una población determinada y su análisis ayuda a comprender los eventos evolutivos que han ocurrido recientemente en regiones genómicas concretas (Nordborg *et al.* 2002; Nielsen 2005).

Este programa ofrece información sobre cada secuencia del alineamiento de entrada y del conjunto de secuencias que componen el alineamiento (Tabla 3). A continuación se detalla el cálculo de algunos de los parámetros más relevantes.

El estimador de diversidad theta de Watterson (Watterson 1975) se calcula mediante las siguientes ecuaciones (Ecs.1 y 2):

$$\theta_w = \frac{K}{a_n}, \text{ donde } K \text{ es el número de sitios segregantes.} \quad (1)$$

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}, \text{ donde } n \text{ es el número de secuencias.} \quad (2)$$

El estimador de diversidad theta de Tajima (Tajima 1983) se calcula con la Ec. 3.

$$\theta_T = 1 - \sum_{i=1}^S \sum_{j=1}^4 \frac{k_{j,i} \times (k_{j,i} - 1)}{n_i \times (n_i - 1)} \quad (3)$$

donde $k_{j,i} > 0$ es el contaje del estado j en el sitio i , n_i es el número de nucleótidos y S es el número de sitios segregantes.

La diversidad haplotípica (H) es un índice que evalúa la diversidad genética mediante haplotipos y se calcula con la Ec. 4 que corresponde con la Ec. 8.4 de Nei (Nei 1987).

$$H = \frac{n}{(n-1)} \left(1 - \sum_{i=1}^k p_i^2 \right) \quad (4)$$

donde n es el número de secuencias de la muestra, k es el número de haplotipos y p_i es la frecuencia de muestreo del haplotipo i -ésimo.

La diversidad nucleotídica (π) es la probabilidad de que al tomar dos secuencias al azar difieran en una posición determinada y se calcula con la Ec. 5 que corresponde con la Ec. 10.5 de Nei (Nei 1987).

$$\pi = \sum_{i=1}^k \sum_{j=1}^k p_i p_j d_{ij} \quad (5)$$

donde d_{ij} es una estima del número de mutaciones ocurridas desde la divergencia entre los haplotipos i y j , k es el número de haplotipos y p_i es la frecuencia del haplotipo i -ésimo.

Tabla 3. Parámetros calculados en EPIPATH-POLYMORPHISM.

Parámetros de cada secuencia	
Identificador de la secuencia	Número de sitios
Número de sitios sin huecos	Contenido GC
Parámetros del alineamiento	
Número de sitios sin huecos	Número de huecos
Número de sitios polimórficos	Número de sitios monomórficos
Número de sitios informativos de parsimonia	Número de sitios no informativos de parsimonia
Número de <i>singletons</i>	Número de mutaciones
Contenido promedio de GC	Número promedio de diferencias nucleotídicas (κ)
Heterocigosidad por sitio	Heterocigosidad por sitio al cuadrado
Theta de Watterson (θ_w)	Theta de Tajima (θ_T)
Número de haplotipos	Diversidad haplotípica (H)
Diversidad nucleotídica (π)	Número de transiciones
Número de transversiones	Tasa de transiciones/transversiones
Número de codones de parada	Número de codones polimórficos con una única posición mutada

Se utilizaron las librerías Bio++ para obtener todos los parámetros excepto la diversidad nucleotídica y haplotípica, que se programaron siguiendo las ecuaciones de Nei. Una característica de esta aplicación es que tiene en cuenta las indeterminaciones, por lo que todos los estados diferentes son considerados como polimorfismos.

EPIPATH-POLYMORPHISM se integró en la plataforma de forma similar a los programas descritos hasta ahora. El formulario se desarrolló con lenguaje PHP e incluye la opción

de realizar un test estadístico basado en coalescencia (esta herramienta se explica en el apartado 3.5.6.1) y la opción *gaps*, que indica al programa cómo debe manejar los huecos y que tiene tres valores: *none* (todas las posiciones con al menos un hueco se eliminan del análisis, valor por defecto), *fifty* (aquellas posiciones con un 50% o más de huecos en el alineamiento no son consideradas en el análisis) y *all* (todas las posiciones son consideradas en análisis).

Una vez realizado el análisis, se obtienen dos ficheros de resultados. Uno con los resultados del análisis en formato de lectura fácil y otro que contiene una línea por fichero de entrada con los resultados separados por ‘;’, con el fin de que puedan procesarse con otros programas como hojas de cálculo.

3.6.4.2 epiPATH-Divergence

Los parámetros de divergencia se emplean en Genética de poblaciones para estimar la distancia evolutiva entre dos poblaciones de secuencias que derivan de un ancestro común. Esto es posible debido a la fijación de alelos distintos en ambos grupos durante el proceso evolutivo, lo que contribuye a la diferenciación de ambas poblaciones. Los parámetros de divergencia aportan información de la evolución que ha ocurrido en un intervalo más amplio de tiempo respecto a los parámetros de polimorfismos (Casillas 2008).

Este programa ofrece información sobre la divergencia intra-poblacional de cada población, la divergencia intra-poblacional de la suma de las dos poblaciones y la divergencia inter-poblacional entre las dos poblaciones (Tabla 4). A

continuación se detalla el cálculo de algunos de los parámetros más relevantes.

La diversidad nucleotídica corregida por Jukes y Cantor (Jukes *et al.* 1969) se calcula con la Ec. 5 pero la diversidad se corrige por

$$d_{ij} = -\frac{3}{4} \log_e \left(1 - \frac{4}{3} p_{ij} \right) \quad (6)$$

La desviación típica de la diversidad nucleotídica corregida por Jukes y Cantor es la raíz cuadrada de la varianza de π obtenida con la Ec. 7 que corresponde a la Ec. 10.7 de Nei (Nei 1987).

$$V(\pi) = \frac{4}{n(n-1)} \left[\begin{aligned} & (6-4n) \left(\sum_{i<j} x_i x_j \pi_{ij} \right)^2 + \\ & (n-2) \sum x_i x_j x_z \pi_{ij} \pi_{iz} + \sum_{i<j} x_i x_j \pi_{ij}^2 \end{aligned} \right] \quad (7)$$

La divergencia entre dos poblaciones (D_{xy}) es la diversidad nucleotídica (Ec. 5) aplicada a pares de secuencias en los que una secuencia pertenece a la población x y otra secuencia pertenece a la población y , como se ve en la Ec. 8.

$$D_{xy} = \sum_{ij} x_i y_j d_{ij} \quad (8)$$

donde d_{ij} es una estima del número de mutaciones ocurridas desde la divergencia entre los haplotipos i y j .

La divergencia entre dos poblaciones corregida por Jukes y Cantor es la Ec. 8 pero con la divergencia de los haplotipos i y j corregida por la Ec. 6.

La varianza de la divergencia corregida por Jukes y Cantor se calcula con la Ec. 9 que corresponde a la Ec. 10.24 de Nei (Nei 1987) y con las divergencias corregidas por la Ec. 6.

$$V(D_{xy}) = \frac{1}{n_x n_y} \left[\begin{aligned} & \left(1 - n_x - n_y\right) \left(\sum x_i y_j d_{ij} \right)^2 + \\ & \left(n_x - 1 \right) \left(\sum x_i^2 y_j d_{ij}^2 + \sum y_i x_j x_k d_{ij} d_{ik} \right) + \\ & \left(n_y - 1 \right) \left(\sum x_i y_j^2 d_{ij}^2 + \sum x_i y_j y_k d_{ij} d_{ik} \right) + \\ & \sum x_i y_j d_{ij}^2 \end{aligned} \right] \quad (9)$$

La desviación típica de la divergencia corregida por Jukes y Cantor es la raíz cuadrada de la divergencia calculada con la Ec. 9.

La divergencia neta entre dos poblaciones (D_a) es el número medio de diferencias fijadas entre las dos muestras y se calcula con la Ec. 10.

$$D_a = D_{xy} - \frac{(D_x + D_y)}{2} \quad (10)$$

La divergencia neta entre dos poblaciones corregida por Jukes y Cantor se calcula con la Ec. 10 pero las divergencias se corrigen con la Ec. 6.

La varianza de la divergencia neta corregida por Jukes y Cantor se calcula con la Ec. 11 que corresponde a la Ec. 25 de Nei (Nei *et al.* 1981).

$$V(D_a) = V(D_{xy}) + \frac{1}{4} \left[V(D_x) + V(D_y) \right] - \frac{1}{2} \left[Cov(D_{xy}, D_x) + Cov(D_{xy}, D_y) \right] \quad (11)$$

donde las divergencias se corrigen con la Ec. 6.

La desviación estándar de la varianza de la divergencia neta corregida por Jukes y Cantor se calcula con la raíz cuadrada de la Ec. 11.

Tabla 4. Parámetros calculados en EPIPATH-DIVERGENCE.

Parámetros intra-poblacionales de cada población	
Número de secuencias	Número de haplotipos
Número de sitios	Número de sitios polimórficos
Número de mutaciones	Número promedio de diferencias nucleotídicas (κ)
Diversidad nucleotídica (π)	Diversidad nucleotídica corregida por Jukes y Cantor
Desviación típica de la diversidad nucleotídica corregida por Jukes y Cantor	
Parámetros intra-poblacionales de la suma de las dos poblaciones	
Número de secuencias	Número de sitios
Número de sitios polimórficos	Número de mutaciones
Número promedio de diferencias nucleotídicas (κ)	Diversidad nucleotídica (π)
Parámetros inter-poblacionales	
Número de diferencias fijadas	Número promedio de diferencias nucleotídicas (κ)
Divergencia (D_{xy})	Divergencia corregida por Jukes y Cantor
Varianza de la divergencia corregida por Jukes y Cantor	Desviación típica de la divergencia corregida por Jukes y Cantor
Divergencia neta (D_n)	Divergencia neta corregida por Jukes y Cantor
Varianza de la divergencia neta corregida por Jukes y Cantor	Desviación estándar de la varianza de la divergencia neta corregida por Jukes y Cantor

Se utilizaron las librerías Bio++ para obtener algunos parámetros como el número de secuencias, número de sitios, número de sitios polimórficos y número total de mutaciones. El número de diferencias fijadas se obtiene con el método de Hey (Hey 1991). Al igual que en EPIPATH-POLYMORPHISM, este programa tiene en cuenta las indeterminaciones.

EPIPATH-DIVERGENCE se integró mediante un formulario en lenguaje PHP y los resultados se obtienen de forma similar a la herramienta anterior, un fichero que puede leerse fácilmente y otro fichero con formato procesable por otros programas.

3.6.4.3 epiPATH-Synonymous

Las sustituciones sinónimas son aquellas que no provocan un cambio de aminoácido en la proteína y las sustituciones no-sinónimas son aquellas modificaciones en la secuencia nucleotídica que derivan en un cambio en la proteína. La teoría neutral mantiene que la mayor parte de la variación molecular observada se debe a la fijación al azar de mutaciones selectivamente neutras (Kimura 1983), como lo son las sustituciones sinónimas. En cambio, la selección suele actuar sobre las sustituciones no-sinónimas debido a sus consecuencias a nivel proteico.

Este programa ofrece información sobre los parámetros que miden las sustituciones sinónimas y no-sinónimas en un alineamiento (Tabla 5). Se programó el cálculo de los parámetros empleando las librerías Bio++ siguiendo la metodología de Nei y Gojobori (Nei *et al.* 1986). A continuación se detalla el cálculo de algunos de los parámetros más relevantes.

La diversidad nucleotídica sinónica (P_s) es el número promedio de diferencias sinónimas calculado con la Ec. 12 (Nei *et al.* 1986).

$$P_s = \frac{S_d}{S} \quad (12)$$

La diversidad nucleotídica sinónima corregida por Jukes y Cantor (d_s) es el número promedio de diferencias sinónimas con la corrección de Jukes y Cantor (Jukes *et al.* 1969) y se calcula con la Ec. 13.

$$d_s = -\frac{3}{4} \log_e \left(1 - \frac{4}{3} p_s \right) \quad (13)$$

La diversidad nucleotídica no-sinónima (P_n) es el número promedio de diferencias no-sinónimas calculado con la Ec. 14 (Nei *et al.* 1986).

$$P_n = \frac{N_d}{N} \quad (14)$$

La diversidad nucleotídica no-sinónima con Jukes y Cantor (d_n) es el número promedio de diferencias no-sinónimas con la corrección de Jukes y Cantor (Jukes *et al.* 1969) y se calcula mediante la Ec. 15.

$$d_n = -\frac{3}{4} \log_e \left(1 - \frac{4}{3} p_n \right) \quad (15)$$

Tabla 5. Parámetros calculados en EPIPATH-SYNONYMOUS.

Número de secuencias	Número de sitios totales
Número medio de posiciones potenciales sinónimas (S)	Número medio de posiciones potenciales no-sinónimas (N)
Número promedio de sustituciones sinónimas (S_d)	Número promedio de sustituciones no-sinónimas (N_d)
Número promedio total de sustituciones	Diversidad nucleotídica sinónima (P_s)
Varianza de P_s	Desviación típica de P_s
P_s corregida por Jukes y Cantor (d_s)	Varianza de d_s
Desviación típica de d_s	Diversidad nucleotídica no-sinónima (P_n)
Varianza de P_n	Desviación típica de P_n
P_n corregida por Jukes y Cantor (d_n)	Varianza de d_n
Desviación típica de d_n	Tasa d_n/d_s

Junto a los parámetros calculados, se incluye una tabla de frecuencias de los variantes que aporta información del número de casos de cada frecuencia, su valor S y su valor N . En esta tabla un variante es un nucleótido con frecuencia $x/(número total de secuencias)$ para un sitio dado. Existen tantas frecuencias posibles como número total de secuencias en el alineamiento. Por tanto, el número de casos de cada frecuencia es equivalente al número de veces que un nucleótido aparece con esta frecuencia a lo largo de todas las posiciones del alineamiento. S y N son el sumatorio para cada frecuencia de su valor sinónimo y no-sinónimo, respectivamente.

EPIPATH-SYNONYMOUS se integró mediante un formulario en lenguaje PHP con la opción específica *genetic code* que indica al programa qué código genético emplear como referencia en los cálculos. En este programa únicamente se toma en cuenta el *pathway* con el mínimo cambio. Los

resultados se obtienen con los mismos formatos que en las herramientas de EPIPATH descritas anteriormente.

3.6.4.4 epiPATH-Neutralit

Para detectar las desviaciones de la neutralidad por parte de las mutaciones existen diversas pruebas que se incluyeron en este programa. La prueba de Tajima (Tajima 1989) compara π y S , dos estimadores del parámetro θ que mide la variabilidad esperada bajo el modelo neutral. Si la diferencia entre π y S es mayor que la esperada por el modelo neutral θ , el modelo es rechazado. El estadístico D sirve para detectar desviaciones de la neutralidad de las mutaciones con esta prueba. La prueba de Fu y Li (Fu *et al.* 1993) es similar a la anterior pero realiza estimas del parámetro θ a partir de η_s (número de mutaciones que aparecen una vez en las secuencias) y η (número total de mutaciones). Un valor negativo de D indica un exceso de mutaciones recientes, mientras que un valor positivo indica un defecto de este tipo de mutaciones.

La información que ofrece EPIPATH-NEUTRALITY (Tabla 6) es, por un lado, relativa a los test de neutralidad de la muestra *ingroup* (grupo de secuencias objeto del estudio), y por otro lado, relativa a los test de neutralidad de la muestra *ingroup* respecto a la muestra *outgroup* de 3 formas diferentes. La primera de ellas es la opción por defecto y se llama ‘modo=1’. En este análisis se emplea la primera secuencia del fichero de la muestra 2 como *outgroup* y se excluyen las posiciones ambiguas (aquellas posiciones en las que el nucleótido presente en el *outgroup* no se encuentra ninguna secuencia del *ingroup*) y los

gaps. Esta aproximación es similar a la empleada en el programa DNASP (Librado *et al.* 2009; Rozas 2009). La segunda de ellas se llama ‘modo=2’, donde se emplean todas las secuencias del fichero *outgroup*, cada una de forma independiente para calcular los parámetros, y a continuación se obtiene la media, la varianza y la desviación típica de cada parámetro. Se excluyen las posiciones ambiguas y los *gaps*. La tercera de ellas se llama ‘modo=3’ y es similar al método del ‘modo=2’ aunque con la diferencia de que el filtrado de las posiciones ambiguas y los *gaps* se realiza en todo el alineamiento *outgroup* en conjunto, y antes de calcular los parámetros con cada secuencia del *outgroup*. De este modo, las posiciones analizadas son las mismas en todas las secuencias, tanto en el *ingroup* como en el *outgroup*. Después, se calculan los parámetros con cada secuencia del *outgroup* de forma independiente, y se obtiene la media, la varianza y la desviación típica de cada parámetro.

A continuación se detalla el cálculo de algunos de los parámetros más relevantes.

El estadístico *D* de Tajima calculado con el número de sitios polimórficos mediante la Ec. 16.

$$D = \frac{\theta_\pi - \theta_s}{\sqrt{Var(\theta_\pi - \theta_s)}} \quad (16)$$

El estadístico *D* de Tajima calculado con el número total de mutaciones es similar al de la Ec. 16, pero en vez de emplear *S* en la comparación se utiliza η .

El estadístico *D** de Fu y Li se calcula con la Ec. 17.

$$D^* = \frac{\left(\frac{n}{n+1} \eta - a_s \eta_s \right)}{\sqrt{Var(\eta - \eta_s)}} \quad (17)$$

El estadístico F^* de Fu y Li se calcula de forma similar al anterior pero realiza estimas del parámetro θ a partir de η_s y π mediante la Ec. 18.

$$F^* = \frac{\pi - \left(\frac{n}{n-1} \right) \eta_s}{\sqrt{Var(\pi - \eta_s)}} \quad (18)$$

Con el índice de neutralidad (NI) se comprueba si la evolución es puramente neutral ($P_n/P_s = D_n/D_s$), si existe selección purificadora ($P_n/P_s > D_n/D_s$) o selección positiva ($P_n/P_s < D_n/D_s$) (McDonald *et al.* 1991). Este parámetro se calcula con la Ec. 19.

$$NI = \frac{(P_n/P_s)}{(D_n/D_s)} \quad (19)$$

donde P_n es el número de polimorfismos no-sinónimos, P_s es el número de polimorfismos sinónimos, D_n es el número de sustituciones no-sinónimas y D_s es el número de sustituciones sinónimas.

El estadístico D de Fu y Li se calcula con la Ec. 20.

$$D = \frac{\eta - a_n \eta_e}{\sqrt{u_D \eta + v_D \eta^2}} \quad (20)$$

El estadístico F de Fu y Li se calcula con la Ec. 21.

$$F = \frac{\Pi_n - \eta_e}{\sqrt{u_F \eta + v_F \eta^2}} \quad (21)$$

donde Π_n es el número promedio de diferencias entre todas las secuencias.

Tabla 6. Parámetros calculados en EPIPATH-NEUTRALITY.

Parámetros de la muestra <i>ingroup</i>	
Número de secuencias	Número de sitios
Número de sitios polimórficos	Número de mutaciones (η)
Theta de Tajima (θ_T)	D de Tajima con el número de sitios polimórficos
D de Tajima con el número total de mutaciones	D* de Fu y Li
F* de Fu y Li	
Parámetros de la muestra <i>outgroup</i>	
Número de secuencias	Número de sitios
Número de sitios polimórficos	Número de mutaciones (η)
Parámetros de la muestra <i>ingroup</i> respecto al <i>outgroup</i> (modo=1)	
Número de sitios del <i>ingroup</i> incluidos	Número de sitios excluidos
Sitios excluidos	Número de sitios polimórficos en el <i>ingroup</i>
Número total de mutaciones en el <i>ingroup</i>	Número total de mutaciones en las ramas externas
Número de sitios polimórficos no-sinónimos (P_a)	Número de sitios polimórficos sinónimos (P_s)
Número de sustituciones fijadas no-sinónimas (D_a)	Número de sustituciones fijadas sinónimas (D_s)
Índice de neutralidad (NI)	D de Fu y Li
F de Fu y Li	
Parámetros de la muestra <i>ingroup</i> respecto al <i>outgroup</i> (modo=2)	
Parámetros del <i>ingroup</i> con cada secuencia del <i>outgroup</i>	
Número de sitios del <i>ingroup</i> incluidos	Número de sitios excluidos
Sitios excluidos	Número de sitios polimórficos en el <i>ingroup</i>
Número total de mutaciones en el <i>ingroup</i>	Número total de mutaciones en las ramas externas
Índice de neutralidad (NI)	D de Fu y Li
F de Fu y Li	
Valores promedio con cada secuencia del <i>outgroup</i>	
Mutaciones en las ramas externas	Varianza del número total de mutaciones en las ramas externas
Índice de neutralidad (NI)	Varianza del NI
Desviación estándar del NI	D de Fu y Li
Varianza de D de Fu y Li	Desviación estándar de D de Fu y Li
F de Fu y Li	Varianza de F de Fu y Li
Desviación estándar de F de Fu y Li	

Tabla 6 (continuación). Parámetros calculados en EPIPATH-NEUTRALITY.

Parámetros de la muestra <i>ingroup</i> respecto al <i>outgroup</i> (modo=3)	
Parámetros de los sitios analizados respecto al conjunto del <i>outgroup</i>	
Número de sitios del <i>ingroup</i> incluidos	Número de sitios excluidos
Sitios excluidos	Número de sitios polimórficos en el <i>ingroup</i>
Número total de mutaciones en el <i>ingroup</i>	
Parámetros del <i>ingroup</i> con cada secuencia del <i>outgroup</i>	
Número total de mutaciones en las ramas externas	Índice de neutralidad (NI)
D de Fu y Li	F de Fu y Li
Valores promedio con cada secuencia del <i>outgroup</i>	
Mutaciones en las ramas externas	Varianza del número total de mutaciones en las ramas externas
Índice de neutralidad (NI)	Varianza del NI
Desviación estándar del NI	D de Fu y Li
Varianza de D de Fu y Li	Desviación estándar de D de Fu y Li
F de Fu y Li	Varianza de F de Fu y Li
Desviación estándar de F de Fu y Li	

EPIPATH-NEUTRALITY se integró mediante un formulario en lenguaje PHP similar a los anteriores donde se incluyó la opción de realizar un test estadístico basado en coalescencia (ver apartado 3.5.6.1) y la opción *mode*, que indica al programa cómo debe analizar el *ingroup* con el *outgroup*. Los resultados se obtienen con los mismos formatos que en las herramientas de EPIPATH descritas anteriormente.

3.6.5 Libsequence

LIBLEQUENCE es una librería de C++ diseñada para ayudar en la programación de software para Genómica y Genética de poblaciones (Thornton 2003). Existen varios programas ya compilados dentro de este proyecto de los cuales

COMPUTE se integró en la plataforma bioinformática por su interés en Genética molecular de poblaciones.

3.6.5.1 Compute

Este programa calcula una serie de parámetros relevantes en los estudios de genética molecular poblacional, como los parámetros B y Q de Wall (Wall 1999), theta de Watterson por sitio, diversidad nucleotídica por sitio, D de Tajima, D* y F* de Fu y Li y la C de Hudson (Hudson 1987), un estimador de la tasa de recombinación en poblaciones.

COMPUTE se integró mediante un formulario en lenguaje PHP con las opciones específicas de COMPUTE. Entre estas últimas están la opción de elegir la secuencia que se emplea como *outgroup* en el análisis y la opción de utilizar los sitios segregantes en vez de las mutaciones inferidas en los cálculos, entre otras.

3.6.6 Análisis estadísticos

En el apartado STATISTICS de las herramientas de análisis implementadas en la plataforma se encuentran las siguientes aplicaciones.

3.6.6.1 Tests estadísticos basados en coalescencia

La coalescencia es la generación en la que se todos los descendientes de una genealogía comparten un único ancestro común. Los métodos basados en coalescencia describen las relaciones genealógicas entre una muestra al azar de alelos bajo el modelo neutral de Wright-Fisher (Wright 1931). Los tests estadísticos basados en coalescencia emplean una muestra poblacional de secuencias generadas bajo un modelo neutral

para obtener la distribución muestral del parámetro de interés sobre la que testar el valor observado del parámetro.

En esta herramienta las muestras de la población se generan con el programa MS (Hudson 2002) que emplea la aproximación de la coalescencia de forma que primero genera una genealogía, con tamaño poblacional constante, sin recombinación y panmixia. Posteriormente, añade las mutaciones aleatoriamente bajo un modelo de sitios infinitos. Los parámetros se calculan sobre esta población con el programa MSSTATS (Thornton 2003) para obtener su distribución.

En la plataforma se implementó esta metodología de dos formas diferentes. Por un lado, se integró el formulario en los lenguajes PHP y PERL para lanzar el análisis estadístico de forma independiente desde el menú en el apartado STATISTICS - COALESCENCE. En el formulario se incluyó una zona para insertar las variables necesarias en el análisis (número de secuencias, número de réplicas, parámetro de mutación, tipo de parámetro sobre el que se realiza el análisis, valor del parámetro de interés e intervalo de confianza). Los ficheros de resultados son uno con el resultado del análisis estadístico y otro que incluye los valores de inicialización del programa MS.

Por otro lado, este test basado en coalescencia se integró como opción en los formularios de las herramientas EPIPATH-POLYMORPHISM y EPIPATH-NEUTRALITY, donde hay una zona en la que especificar el número de réplicas, el parámetro de mutación y el intervalo de confianza. Esta opción se programó con los lenguajes PHP y PERL. El resultado se integra dentro de los ficheros de resultados del correspondiente

análisis y además, se obtiene un fichero llamado *coalsimstats* que incluye el resultado de los programas MS y MSSTATS.

3.6.7 Utils

En este apartado del menú de la plataforma se encuentran otros programas que resultaron útiles para el análisis de secuencias.

3.6.7.1 RevTrans

REVTRANS es un programa que realiza la traducción reversa de un alineamiento de aminoácidos, es decir, a partir de un alineamiento de secuencias proteicas y sus correspondientes secuencias aminoacídicas, traduce el alineamiento en aminoácidos (Wernersson 2003). Este paso es importante en el análisis de secuencias puesto que una vez se dispone de los alineamientos de proteínas es necesario obtener su correspondiente alineamiento en nucleótidos para poder calcular los parámetros de interés (Figura 14).

Esta aplicación se integró mediante un formulario en lenguaje PHP con las opciones específicas del formato de los resultados y el código genético, entre otras.

3.7 Seguridad de la información

En la plataforma que se desarrolló durante esta tesis se implementaron varios sistemas de seguridad.

La primera restricción de uso y acceso a la plataforma es la identificación del usuario, tanto a través de la interfaz *web* como desde la línea de comandos, se requiere un nombre de usuario y una contraseña. Estos datos los genera el

administrador de sistemas y son los mismos para acceder a la base de datos local. Una vez identificado el usuario, su sesión se mantiene hasta que pulsa *logout* o bien hasta que se cumpla el tiempo programado de inactividad, que actualmente son 30 minutos aunque aparece un aviso a los 25 minutos.

Por otro lado, las vistas en la base de datos permiten restringir el acceso a un subconjunto de información almacenada. En esta plataforma se filtró la información para diferentes usuarios según el proyecto al que se encuentran asociadas las secuencias. Estos requisitos pueden ser modificados o ampliados según las necesidades de cada momento y grupo de usuarios. Con las vistas aumenta el nivel de seguridad de la aplicación conforme se generan más niveles de tablas virtuales. El sistema de vistas de MYSQL apareció en su versión 5.0.1 por lo que no pudo ser utilizado en la primera versión de la plataforma bioinformática. Además, se diseñó e implementó una segunda base de datos en la que quedan registradas las modificaciones realizadas en la base de datos principal. Únicamente el administrador puede acceder a esta base de datos secundaria.

En cuanto a los ficheros de resultados, se almacenan temporalmente en el servidor. De esta forma, cada usuario se ve obligado a guardar en su ordenador estos ficheros. Con esto se consigue aligerar el espacio disponible en el servidor y que cada usuario almacene los ficheros que considere útiles en un lugar distinto, aumentando la seguridad en el acceso a los mismos.

Por último, es una buena práctica realizar copias de seguridad regularmente. En este caso, el administrador local es el encargado de realizar las copias de ambas bases de datos.

4. APLICACIONES

4 Aplicaciones

En este apartado se incluyen dos casos reales en los que se utilizó la plataforma bioinformática desarrollada. Ambos estudios podrían haberse realizado sin el empleo de la plataforma pero incrementando el tiempo y la dificultad en el procesado y análisis de la información.

4.1 Predicción de la respuesta al tratamiento en pacientes con hepatitis C.

4.1.1 Introducción

Se han establecido distintos tipos de respuesta al tratamiento de la hepatitis C dependiendo del número de semanas que transcurren hasta el momento en que los niveles de RNA-VHC en suero o plasma del paciente pasan a ser indetectables (Ghany *et al.* 2009). Una respuesta rápida (RVR) se alcanza a las 4 semanas de tratamiento, una respuesta temprana (EVR) a las 12 semanas de tratamiento, una respuesta al final del tratamiento (ETR) transcurridas 24 o 48 semanas de tratamiento y una respuesta sostenida (SVR) a las 24 semanas desde el fin del tratamiento. Se ha visto que este último tipo de respuesta depende especialmente del genotipo viral (Manns *et al.* 2001; Fried *et al.* 2002; Hadziyannis *et al.* 2004). En el caso del genotipo 1, la tasa de SVR en pacientes tratados durante 48 semanas es de 38-48%. Por otro lado, si a las 24 semanas de tratamiento siguen encontrándose partículas virales en el suero del paciente, se considera como ‘no respondedor’ y si aparecen

de nuevo partículas virales tras la interrupción del tratamiento se considera al paciente como ‘recidiva’.

Dado que la progresión de la enfermedad es heterogénea, la individualización y optimización de la terapia para cada tipo de paciente es un proceso habitual y muy estudiado (Brown 2007; Lee *et al.* 2008b; Mallet *et al.* 2010). Para conseguir más efectividad en la acción terapéutica, sería aconsejable identificar aquellos pacientes que responderán al tratamiento antes de comenzarlo. Se sabe que los factores virales, ambientales, del hospedador y del tratamiento juegan un importante papel en el resultado de la infección y en la respuesta al tratamiento (Gao *et al.* 2004; Hofmann *et al.* 2005; Hayashi *et al.* 2006; Lee *et al.* 2008a; Mallat *et al.* 2008; Pang *et al.* 2009). Existen numerosos estudios sobre factores de predicción de la respuesta que toman variables medidas antes del tratamiento (Jensen *et al.* 2006; Martínez-Bauer *et al.* 2006; Martinot-Peignoux *et al.* 2006; Akuta *et al.* 2007; Backus *et al.* 2007; Shirakawa *et al.* 2008; Trapero-Marugan *et al.* 2008; Okanoue *et al.* 2009; Kurosaki *et al.* 2010; Lukasiewicz *et al.* 2010; Saito *et al.* 2010; Saludes *et al.* 2010). En general, estos estudios tienen en cuenta aquellas variables que describen el estado clínico del paciente (niveles de ALT, concentración de virus en suero, biopsia del hígado, etc.), así como datos demográficos y epidemiológicos del paciente (edad, sexo, hábitos, etc.) y algunas variables del virus como el genotipo, sustituciones en las regiones *core* e ISDR y varios parámetros descriptores de la variabilidad genética en la región E1E2. En la bibliografía encontramos que las variables pre-tratamiento más consistentes con la respuesta favorable a la terapia basada en

IFN son el genotipo y la carga viral (Lee 2003; Feld *et al.* 2005; Kau *et al.* 2008).

En un trabajo de Szmaragd y colaboradores (Szmaragd *et al.* 2006) se propuso el empleo de clados filogenéticos del virus como predictores de la respuesta del paciente en vez de los polimorfismos individuales del virus. Esto aumenta el poder estadístico para detectar polimorfismos candidatos de una secuencia viral asociada al resultado de la infección. En esta aproximación se emplea un árbol filogenético en el que las diferencias entre clados se detectan a nivel de genotipos. Por esta razón, este método no puede ser aplicado cuando se quiere distinguir entre secuencias agrupadas en un mismo subtipo, ya que no se dispone de suficiente evidencia estadística para separar significativamente unos clados de otros de un mismo subtipo viral.

En nuestro estudio se emplea una medida alternativa para superar esta falta de evidencia estadística dentro de un subtipo en un árbol filogenético. Además, se incluye una mayor cantidad de factores virales que en los estudios previos, junto con factores relativos al paciente y al tratamiento, para desarrollar un modelo de predicción de la respuesta al tratamiento antes de comenzar la terapia en pacientes con genotipos 1a y 1b del HCV.

4.1.2 Material y métodos

Se obtuvieron secuencias virales de 72 pacientes con HCV tratados con terapia antiviral combinada durante 48 semanas. El resultado del tratamiento se conocía para 67 de los

72 pacientes; además se conocían completamente los factores clínicos en 49 de los 67 pacientes (Jiménez-Hernández 2004; Torres-Puente 2004). En este estudio se emplearon estos 49 pacientes de los cuales 17 estaban infectados por el genotipo viral 1a y 32 por el genotipo viral 1b. Se recuperaron los datos clínicos y epidemiológicos de los pacientes junto con las secuencias virales (regiones NS5A y E1E2) de la base de datos local (Amadoz *et al.* 2007).

4.1.2.1 Variables del paciente

La variable de interés es la respuesta al tratamiento y se quiere conocer el tipo de respuesta que tendrá un paciente antes de administrarlo. La respuesta al tratamiento es una variable categórica con dos valores: positiva (hay respuesta) o negativa (no hay respuesta). La clasificación de los pacientes como respondedores o no respondedores al tratamiento antiviral fue realizada por los hospitales que proporcionaron las muestras. El criterio empleado para considerar a un paciente como ‘respuesta positiva’ fue la ausencia de RNA viral detectable en suero o una caída superior a 2 log en la carga viral respecto de la basal en la semana 12 tras el inicio del tratamiento. Se consideró ‘respuesta negativa’ cuando no se cumplía la condición anterior o se obtenía un resultado positivo de RNA-HCV más allá de la semana 12 desde el inicio del tratamiento.

Las variables relacionadas con el paciente que se han incluido en este estudio son la edad, sexo, índice de Knodell (clasificación histológica que mide la lesión en el tejido hepático) (Knodell *et al.* 1981; Desmet 2003), niveles de

GOT/GPT en suero, niveles de ALT en suero, duración del tratamiento, tratamiento completado, número de tratamiento, dosis de IFN y dosis de RBV. Las características de las variables del paciente se resumen en la Tabla 7.

Tabla 7. Variables de paciente y respuesta al tratamiento.

Variables	Pacientes (n=49)
Resultado	
Respuesta: R/n (porcentaje)	23/49 (46,9%)
No respuesta: NR/n (porcentaje)	26/49 (53,1%)
Factores del paciente	
Edad (años)	43,61 ± 12,122 (23;73)
Sexo	
Hombre: H/n (porcentaje)	34/49 (69,4%)
Mujer: M/n (porcentaje)	15/49 (30,6%)
Índice Knodell	8 ± 3,403 (1;17)
GOT/GPT	0,601 ± 0,323 (0,3;2,3)
ALT (U/L)	122,67 ± 74,463 (24;361)
Duración tratamiento (meses)	11,27 ± 1,987 (6;12)
Tratamiento completado	
Si: S/n (porcentaje)	43/49 (87,8%)
No: N/n (porcentaje)	6/49 (12,2%)
Número de tratamiento	
1: 1/n (porcentaje)	28/49 (57,1%)
2: 2/n (porcentaje)	20/49 (40,8%)
3: 3/n (porcentaje)	1/49 (2%)
Dosis IFN	
3 mU/3vps	40/49 (81,6%)
5 mU/3vps	3/49 (6,1%)
90 µg/día	1/49 (2%)
100 µg/día	3/49 (6,1%)
120 µg/día	2/49 (4,1%)
Dosis RBV (mg/día)	1040,82 ± 122,336 (800;1200)

Nota: mU/3vps: millones de unidades 3 veces por semana.

Los datos se indican como media ± desviación típica si no se especifica otro criterio.

4.1.2.2 Variables del virus

Los variantes que conforman las poblaciones del HCV se encuentran sometidos a la presión inmune del hospedador, aunque esta presión no afecta por igual a todas las regiones del genoma del virus (Torres-Puente 2004). Dos de las regiones más estudiadas en la relación entre variabilidad genética y resistencias a tratamientos son las regiones E1E2 y NS5A del genoma viral. Aparte del genotipo, los factores virales incluidos en este estudio se han calculado para cada paciente a partir de secuencias parciales de ambas regiones, de 472 nucleótidos en el caso de la E1E2 (nucleótidos 1310-1781 en el genoma del HCV, secuencia referencia con número de acceso D50481, que incluye las tres regiones hipervariables pero no la región E2-PePHD) y de 743 nucleótidos en el de la NS5A (nucleótidos 6742-7484 en el genoma HCV, secuencia referencia con número de acceso D50481, que incluye las regiones ISDR, PKR-BD y V3).

Las variables relativas al virus que se han incluido en este estudio son el genotipo, el número de sustituciones no-sinónimas (dN), el número de sustituciones sinónimas (dS), la tasa dN/dS, el número total de mutaciones (η), el número de sitios segregantes (S), el contenido GC (GC), la diversidad haplotípica (H), la H^2 , la H^3 , la diversidad nucleotídica (π), la π^2 , la π^3 , el número medio de diferencias nucleotídicas (κ), los valores de los estadísticos D de Tajima, D* de Fu y Li, F* de Fu y Li, Fs de Fu y el número de sitios bajo selección positiva (pts). Las variables de diversidad haplotípica y nucleotídica se transformaron para obtener la cantidad de información que

pudiera detectar una asociación con la respuesta al tratamiento. Las características de las variables del virus se resumen en la Tabla 8. Los datos del genotipo se obtuvieron del sistema de información, los parámetros de polimorfismos se calcularon con DNAsP (Rozas *et al.* 2003) y los sitios bajo selección positiva se identificaron con CODEML (Yang 1997).

Tabla 8. Variables del virus.

Genotipo	Pacientes (n=49)	
1a: 1a/n (porcentaje)	17/49 (34,7%)	
1b: 1b/n (porcentaje)	32/49 (65,3%)	
Región E1E2		Región NS5A
dN	0,017 ± 0,013 (0;0,044)	0,005 ± 0,005 (0;0,022)
dS	0,041 ± 0,034 (0;0,122)	0,0379 ± 0,032 (0;0,128)
dN/dS	0,671 ± 1,238 (0;8,277)	0,147 ± 0,103 (0;0,479)
η	77,224 ± 43,394 (2;153)	77,449 ± 53,659 (2;177)
S	70,939 ± 38,698 (2;134)	74,224 ± 50,569 (2;163)
GC	0,594 ± 0,011 (0,569;0,621)	0,603 ± 0,008 (0,588;0,621)
H	0,863 ± 0,23 (0,039;0,999)	0,833 ± 0,285 (0,066;1)
H ²	0,797 ± 0,285 (0,002;0,998)	0,774 ± 0,332 (0,004;1)
H ³	0,75 ± 0,315 (0;0,997)	0,736 ± 0,351 (3E-04;1)
π	0,022 ± 0,016 (0;0,058)	0,013 ± 0,011 (9E-05;0,04)
π ²	0,001 ± 0,001 (0;0,003)	2,93E-04 ± 4,02E-04 (8,1E-09;0,002)
π ³	3,01E-05 ± 4,45E-05 (0,2E-04)	7,95E-06 ± 1,51E-05 (0;7,33E-05)
κ	10,458 ± 7,74 (0,039;27,33)	9,782 ± 8,183 (0,07;31,099)
D	-0,993 ± 0,985 (-2,58;0,89)	-1,418 ± 0,678 (-2,43;0,07)
D*	-1,908 ± 1,627 (-5,45;1,88)	-2,567 ± 1,363 (-5,24;0,27)
F*	-1,847 ± 1,507 (-5,21;1,76)	-2,546 ± 1,218 (-4,92;-0,17)
Fs	-46,341 ± 32,161 (-124,662;1,256)	-26,647 ± 19,965 (-85,549;1,662)
pts	3,388 ± 3,523 (0;13)	0,878 ± 1,409 (0-5)

Nota: Los datos se indican como media ± desviación típica (rango) si no se especifica otro criterio.

Además de los parámetros anteriores, se obtuvieron otras variables que agrupan ambas regiones virales por su

similitud. Este método difiere de los árboles filogenéticos en que la medida de la similitud no se basa en la distancia evolutiva entre dos secuencias, sino en la comparación de las secuencias como tales, es decir, a lo largo de todas las posiciones de la región se ven aquellas moléculas que caracterizan a una secuencia y son compartidas por otras secuencias.

Se emplearon 72 pacientes de cada uno de los cuales se obtuvo una secuencia proteica consenso de cada región viral. Esta secuencia es una representación de todas las variantes de un paciente antes de que se le administrase el tratamiento. A continuación, se obtuvo un alineamiento de todas las secuencias consenso y se filtraron aquellas posiciones en las que existía variación entre los pacientes. Este proceso se realizó para cada genotipo (1a y 1b) en cada región (E1E2 y NS5A). Posteriormente se realizó un análisis de componentes principales (PCA) con cada matriz de datos. En el análisis PCA cada posición se considera como una variable y su valor es el tipo de aminoácido presente. Por tanto, se tiene un número de estados posibles igual al número de polimorfismos presentes en la muestra de pacientes.

El método de PCA consiste en la transformación de un grupo de caracteres observados en un nuevo grupo de variables, donde cada nueva variable es una combinación lineal de los caracteres observados. Las primeras componentes principales recogen la mayor parte de la variación entre las secuencias. Este análisis se realizó con la opción de correspondencias múltiples (MCA) del programa SPSS para datos categóricos y el número de dimensiones se eligió según el

criterio de Cattell (Cattell 1966) dado que no hay un acuerdo sobre el criterio más correcto y éste es el más empleado. En este criterio se representan las dimensiones en el eje de abcisas y en el eje de ordenadas los valores propios, y se seleccionan aquellas dimensiones hasta el punto en el que comienza la inflexión de la curva. Aplicando este criterio se obtuvieron 7 dimensiones en la región E1E2 y 12 dimensiones en la región NS5A.

Se empleó el valor obtenido en cada dimensión para cada paciente como una nueva variable en el desarrollo del modelo predictivo. El análisis MCA proporciona también las coordenadas de cada aminoácido en cada posición y la contribución de ese aminoácido a la variación de cada dimensión. Con esta información se evaluaron aquellos aminoácidos que pudieran modular en mayor medida la agrupación o el parecido entre las secuencias.

4.1.2.3 Obtención del modelo predictivo

Se emplearon 49 pacientes y 66 variables, categóricas y numéricas, para modelar la respuesta al tratamiento mediante una regresión logística. Se utilizaron dos metodologías para la obtención del modelo:

- Estudio con subgrupos de variables
- Estudio con todas las variables juntas

Los coeficientes estimados en la regresión representan la pendiente o tasa de cambio de la función de la variable dependiente (respuesta a tratamiento) por unidad de cambio de

la variable independiente (variables del virus y del paciente). En un contexto epidemiológico, un coeficiente obtenido en la regresión equivale al *log odds* o logaritmo del *odds ratio* (razón de ventajas) (OR). El OR es una medida de asociación que indica cuán probable es tener una respuesta positiva al tratamiento entre aquellos pacientes que presenten un carácter sobre aquellos que no lo presenten.

Estudio con subgrupos de variables

Debido a que el número de variables era mayor al de casos, se dividieron las variables en 4 subgrupos: variables relativas al paciente, a la región E1E1 del virus, a la región NS5A del virus y a las dimensiones del análisis MCA. La división de las variables en estos subgrupos tuvo en cuenta el ámbito que caracteriza a cada variable y de esta forma se generó un modelo de predicción equilibrado en los distintos aspectos implicados en la respuesta al tratamiento.

Para la obtención del modelo con esta metodología se empleó la función GLM del programa R (R Development Core Team 2011), que permite desarrollar modelos a partir de datos observados bajo la teoría estadística de los modelos lineares generalizados (GLM). Para seleccionar el mejor modelo se sigue un proceso de selección de variables que simultáneamente verifica su importancia (Figura 16).

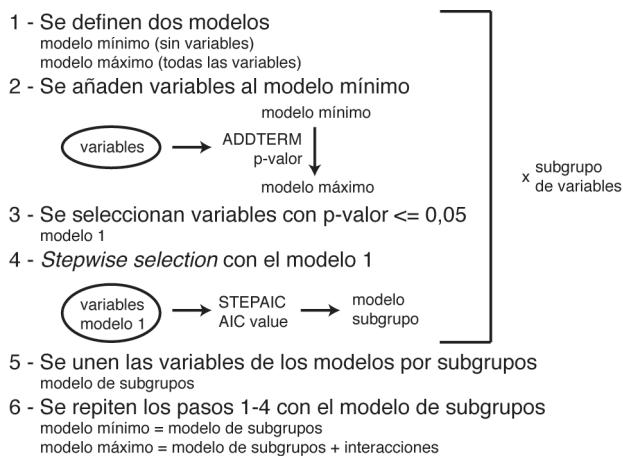


Figura 16. Esquema de obtención del modelo por subgrupos.

Primero se generó un modelo mínimo que no contenía ninguna variable predictora y un modelo máximo que contenía todas las variables. Con la función ADDTERM se fueron añadiendo variables desde el modelo mínimo hasta llegar al más complicado mientras se iba calculando el test de Chi-cuadrado para obtener la significación de la aportación de cada variable. Con este resultado se seleccionaron las variables con un nivel de significación menor o igual a 0.05 y con ellas se generó un modelo (modelo 1). Posteriormente, se empleó el método de *stepwise selection* (selección paso a paso), en el que se incluían y eliminaban variables del modelo 1 por máxima verosimilitud en ambas direcciones: selección hacia delante (partiendo de una variable y añadiendo variables sucesivamente) y selección hacia atrás (partiendo de todas las variables y eliminando variables sucesivamente). Este método añade y elimina variables de un modelo según la significación del estadístico de la razón de verosimilitudes, es decir, se

comprueba si una variable es significativa para el modelo, tomando como nivel de significación 0.05, y la mantiene o elimina dependiendo de si es significativa o no. El método de selección paso a paso se realizó con la función STEPAIC, que calcula para cada modelo generado el valor del criterio de información de Akaike (AIC) (Akaike 1974), que permite comparar distintos modelos. Cuanto menor es el valor de AIC, mejor es el modelo.

Las variables obtenidas en los modelos resultantes para cada subgrupo se unieron en un nuevo modelo y se estudiaron todas las posibles interacciones entre pares mediante la metodología explicada anteriormente, tomando como modelo mínimo el modelo sin variables, como modelo máximo el que incluye todas las variables obtenidas para los subgrupos más sus interacciones, y como modelo de partida el modelo obtenido de los subgrupos.

Se aplicó esta metodología a todos los pacientes (genotipos 1a + 1b) y a ambos subtipos por separado (1a y 1b), obteniendo en total 3 modelos de predicción por subgrupos de variables.

Por último, se evaluó la bondad de ajuste del modelo final comparando las respuestas al tratamiento observadas respecto a las predichas por el modelo. Los grupos de pacientes sobre los que se realizaron estas comparaciones fueron, por un lado, los 49 pacientes empleados en la obtención del modelo y, por otro lado, dos grupos de nuevos pacientes: un grupo de 8 pacientes con información clínica completa y otro grupo de 10 pacientes con datos incompletos. En este último grupo se

estimaron los datos que faltaban con diferentes métodos: el algoritmo Expectación-Maximización (EM) implementado en R y SPSS, la sustitución por las medias y la estimación por regresión en SPSS. El algoritmo EM es un método que estima los valores desconocidos por máxima verosimilitud, en la sustitución por la media se calcula ésta sobre los valores disponibles del resto de los pacientes y en la regresión de SPSS se estima cada valor por regresión de todos datos de los pacientes. Con la función PREDICT de R se realizaron las predicciones en los distintos grupos de pacientes. En esta función se emplea el modelo de regresión logística obtenido y los nuevos datos para estimar la respuesta al tratamiento.

Estudio con todas las variables juntas

Por otro lado, para evaluar todas las variables juntas se obtuvieron modelos de predicción, tanto para los genotipos 1a y 1b juntos como por separado, aplicando la metodología *lasso* mediante la función GLMNET (Friedman *et al.* 2010) de R. Esta metodología selecciona las variables realizando una penalización de los coeficientes de regresión, de forma que si los coeficientes no son mayores que un valor dado, las variables correspondientes no se incluyen en el modelo. En este estudio se utilizaron el mínimo del valor estimado lambda y su desviación típica como penalización. Con esta metodología puede verse cómo las variables de un subgrupo pueden estar afectando a los otros subgrupos.

4.1.2.4 Resumen del análisis

Las herramientas y procesos informáticos empleados en cada paso del análisis se muestran en la Tabla 9.

Tabla 9. Herramientas/procesos empleados en el análisis.

Paso del análisis	Herramienta/proceso informático
Guardar información clínica del paciente (formularios en papel) y las secuencias virales en la plataforma (ficheros de texto).	Carga masiva de datos en la base de datos local a través de la plataforma.
Recuperar información de pacientes seleccionados para el análisis y sus secuencias virales en formato FASTA.	Búsqueda en la base de datos local a través de la plataforma.
Cálculo de parámetros para el estudio de polimorfismos.	DNASP (Rozas <i>et al.</i> 2003), <i>input</i> obtenido con MEGA3 (Kumar <i>et al.</i> 2004).
Cálculo de sitios bajo selección positiva.	CODEML (Yang 1997), <i>input</i> obtenido con MEGA3 (Kumar <i>et al.</i> 2004), PHYML (Guindon <i>et al.</i> 2003) y TREEEDIT.
Análisis de componentes principales de los pacientes en base a sus secuencias virales.	MCA de SPSS, <i>input</i> obtenido con TRANSEQ de EMBOSS (Rice <i>et al.</i> 2000), CLUSTALW (Thompson <i>et al.</i> 1994), CONS de EMBOSS (Rice <i>et al.</i> 2000), MEGA3 (Kumar <i>et al.</i> 2004).
Estudio con subtipos de variables	
Regresión logística.	GLM de R.
Predicción con el modelo obtenido.	PREDICT de R, estimación de parámetros por E-M de R y SPSS, por la media en SPSS y por regresión en SPSS.
Estudio con todas las variables	
Regresión logística aplicando la metodología <i>lasso</i> .	GLMNET de R.

4.1.3 Resultados

Estudio con subgrupos de variables

El mejor modelo estadístico, siguiendo el criterio AIC, de la predicción de la respuesta antes del tratamiento obtenido con 49 pacientes incluye la duración del tratamiento, los niveles de ALT, la dimensión 11 de la región NS5A, la dimensión 7 de la región E1E2, el valor dS de la región E1E2, la H³ de la región E1E2, la π² de la región E1E2, la H de la región NS5A, el número de sitios bajo selección positiva en la región NS5A, la interacción entre la dimensión 11 de la región NS5A y la dimensión 7 de la región E1E2 y la interacción entre el valor dS de la región E1E2 y la H³ de la región E1E2 (Tabla 10).

Tabla 10. Modelo de predicción con 49 pacientes (1a+1b).

Predictor	Coeficiente	Error estándar	Odds ratio
Intersección	-3,19E+03	3,66E+05	1,49E+28
Duración tratamiento	5,89E+02	6,14E+04	7,96E+18
ALT	3,526	3,88E+02	33,98
D11 NS5A	-1,62E+02	1,75E+04	6,33E-06
D7 E1E2	2,68E+01	4,68E+03	1458,25
dS E1E2	-1,43E+05	1,46E+07	6,75E-93
H ³ E1E2	-9,20E+02	1,09E+05	2,99E-30
π ² E1E2	1,94E+05	2,33E+07	inf
H NS5A	-3,01E+03	3,08E+05	5,54E-27
pts NS5A	8,57E+01	9,83E+03	1,30E+10
D11 NS5A : D7 E1E2	3,14E+02	3,29E+04	1,19E+10
dS E1E2 : H ³ E1E2	1,32E+05	1,35E+07	1,20E+85
AIC = 24			

En el grupo de los pacientes con genotipo 1a, el mejor modelo de predicción obtenido incluye la H² de la región E1E2, el contenido GC de la región E1E2 y el número de sitios segregantes de la región NS5A (Tabla 11). Considerando

únicamente los pacientes del genotipo 1b, el mejor modelo obtenido incluye la duración del tratamiento, la dimensión 11 de la región NS5A, el valor dS de la región NS5A, el η de la región NS5A, el κ de la región NS5A y el valor dN de la región NS5A (Tabla 12).

Tabla 11. Modelo de predicción con 17 pacientes (1a).

Predictor	Coeficiente	Error estándar	Odds ratio
Intersección	-1,50E+04	7,11E+06	2,70E-36
H ² E1E2	-1,86E+03	8,60E+05	5,95E-17
GC E1E2	2,76E+04	1,30E+07	2,78E+65
S NS5A	2,406	1,19E+03	11,089
AIC = 8			

Tabla 12. Modelo de predicción con 32 pacientes (1b)

Predictor	Coeficiente	Error estándar	Odds ratio
Intersección	-2,91E+02	2,23E+05	4,58E-10
Duración tratamiento	2,90E+01	1,93E+04	2651,86
D11 NS5A	-3,77E+02	9,65E+04	7,97E-13
dS NS5A	-2,13E+05	5,44E+07	0
η NS5A	-7,627	1,95E+03	4,87E-04
κ NS5A	1,22E+03	3,10E+05	4,38E+10
dN NS5A	-6,44E+05	1,64E+08	0
AIC = 14			

La aplicación de los modelos de predicción a los mismos pacientes que se emplearon para obtenerlos mostró un excelente ajuste entre los datos predichos y observados y un buen ajuste cuando se emplearon nuevos pacientes (Tabla 13). Los resultados de la predicción fueron los mismos empleando distintos métodos de estimación y no hay una relación clara entre el tipo de variable estimada y el fallo en la predicción.

Tabla 13. Bondad de ajuste de las predicciones de respuesta.

Pacientes			
Genotipo	Incluidos	No incluidos	Datos estimados
1a	17/17 (100%)	4/6 (66,66%)	-
1b	32/32 (100%)	6/9 (66,66%)	1/3 (33,33%)
1a+1b	49/49 (100%)	7/10 (70%)	6/8 (75%)

Incluidos: pacientes incluidos en la obtención del modelo.

No incluidos: pacientes no incluidos.

Datos estimados: pacientes en los que algunos datos son estimados.

Resultados expresados en correctos/total.

En la Figura 17 se muestran los porcentajes de la sensibilidad (proporción de los positivos reales correctamente identificados), la especificidad (proporción de negativos reales correctamente identificados), el valor predictivo positivo (proporción de los resultados positivos del test correctamente identificados) (PPV) y el valor predictivo negativo (proporción de los resultados negativos del test correctamente identificados) (NPV) de los modelos obtenidos aplicados a los distintos grupos de pacientes empleados en este estudio.

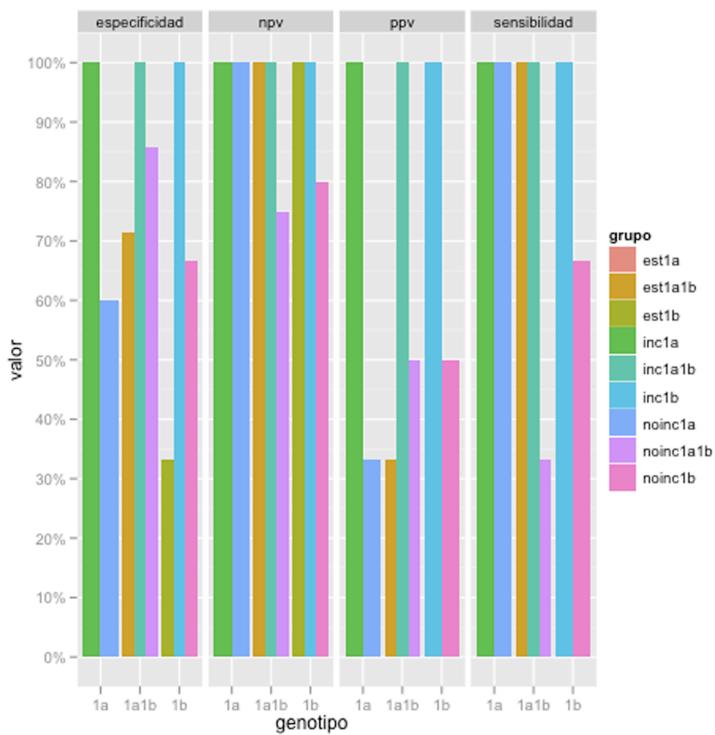


Figura 17. Sensibilidad, especificidad, PPV y NPV de los modelos de predicción.

Debido a que en el modelo de predicción final aparecieron como resultado algunas variables dimensionales, se estudiaron estas dimensiones para encontrar aquellas posiciones que modulan la variabilidad entre pacientes. Se encontró que más del 95% de la variabilidad en la dimensión 11 de la región NS5A se encuentra retenida en 10 de las 69 posiciones del genotipo 1a y en 100 de las 127 posiciones del genotipo 1b. Respecto a la dimensión 7 de la región E1E2, los resultados fueron que 27 de las 65 posiciones en el genotipo 1a y 44 de 93 posiciones en el genotipo 1b retienen más del 95% de la

variabilidad total en cada genotipo. En la Tabla 14 se encuentra un resumen de las posiciones que contribuyen individualmente con más del 3% al total de variabilidad. No se encontró una asociación clara entre el tipo de aminoácido en cada una de estas posiciones y la respuesta al tratamiento.

Tabla 14. Número de posiciones que contribuyen individualmente con más del 3% al total de variabilidad.

Región	Genotipo	Número de posiciones	Contribución total
NS5A	1a	11	98,39 %
NS5A	1b	8	42,41 %
E1E2	1a	16	75,33 %
E1E2	1b	8	39,33 %

Estudio con todas las variables juntas

Los modelos de predicción obtenidos con GLMNET pueden verse en las Tablas 15, 16 y 17 y en las Figuras 18, 19 y 20, la estimación del valor lambda correspondiente. Los coeficientes empleados como punto de corte son el mínimo del valor estimado lambda (min) y su desviación típica (1se).

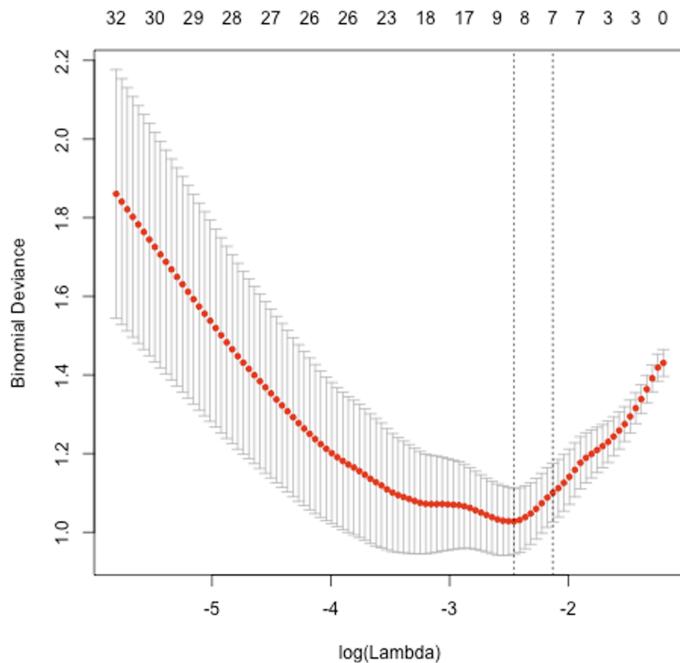


Figura 18. Estimación de lambda para el modelo de genotipos 1a+1b.

Tabla 15. Modelos de predicción de los genotipos 1a+1b.

1se		min	
Variable	Coeficiente	Variable	Coeficiente
(Intercept)	0.8932377276	(Intercept)	0.7146758802
H NS5A	-0.0574104431	D Tajima E1E2	0.0013150119
H2 NS5A	-0.0034528953	S NS5A	-0.0027114319
H3 NS5A	-0.0003951666	H NS5A	-0.0585361814
D11 NS5A	-0.0098538693	H2 NS5A	-0.0035997987
Duración tratamiento	0.0291290617	H3 NS5A	-0.0003970471
Dosis IFN	-0.1112508592	D11 NS5A	-0.0193521583
Nº tratamiento	0.4115901120	Duración tratamiento	0.0525284206
		Dosis IFN	-0.2501826982
		Nº tratamiento	0.8715472925

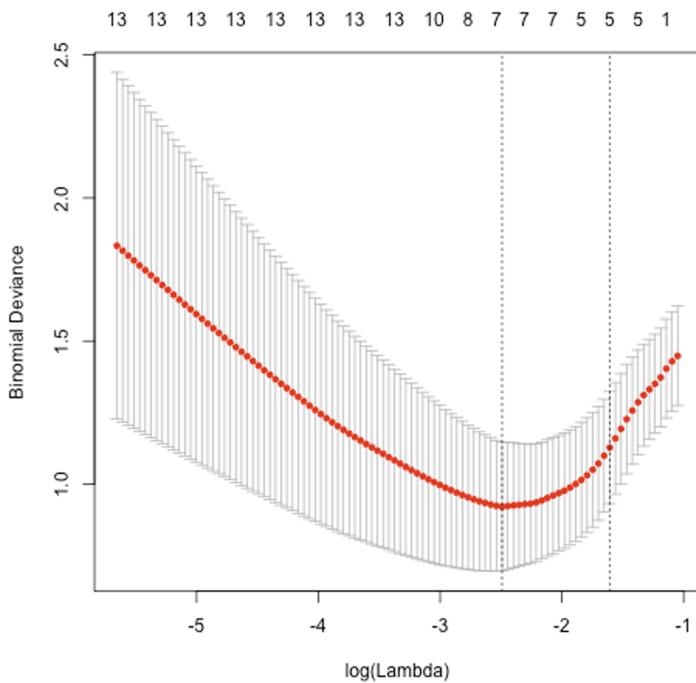


Figura 19. Estimación de lambda para el modelo del genotipo 1a.

Tabla 16. Modelos de predicción del genotipo 1a.

Variable	1se	Variable	min
	Coeficiente		Coeficiente
(Intercept)	1.6629146457	(Intercept)	4.357408358
D1 E1E2	-0.0358974608	D1 E1E2	-0.139248512
H NS5A	-0.0433025243	H NS5A	-0.082866627
H2 NS5A	-0.0057233733	H2 NS5A	-0.018300252
H3 NS5A	-0.0008050908	H3 NS5A	-0.005117889
Dosis IFN	-0.2417936593	D8 NS5A	0.044515440
		Edad	-0.035799333
		Dosis IFN	-0.444569786

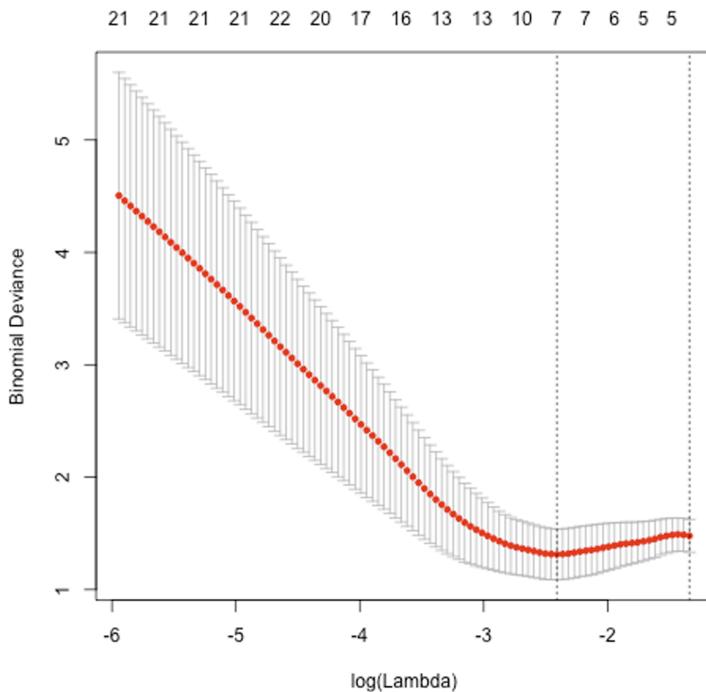


Figura 20. Estimación de lambda para el modelo del genotipo 1b.

Tabla 17. Modelos de predicción del genotipo 1b.

1se		min	
Variable	Coeficiente	Variable	Coeficiente
(Intercept)	-0.5108256	(Intercept)	-0.34297397
		D Tajima E1E2	0.00776326
		S NS5A	-0.01354700
		D7 NS5A	0.01325065
		D9 NS5A	0.01485201
		D11 NS5A	-0.06691847
		Duración tratamiento	0.15503516
		GOT/GPT	-0.06377769

4.1.4 Discusión

El principal objetivo de este estudio es obtener un modelo predictivo para la identificación de pacientes, con HCV de genotipo 1, susceptibles de la terapia combinada empleando variables tomadas antes del tratamiento. Para ello, además de variables del paciente y del tratamiento, se han incluido en este estudio retrospectivo más factores que caracterizan al virus antes del tratamiento que en otros trabajos. Un novedoso factor viral introducido en este trabajo son las coordenadas en el espacio multidimensional, que agrupan las secuencias en base a su similitud en regiones específicas del genoma frente a la globalidad de la región considerada, como resulta del análisis filogenético de las mismas. La información de los pacientes infectados con los subtipos 1a y 1b de HCV y tratados con la terapia combinada (IFN y RBV) se recuperó de la plataforma bioinformática previamente desarrollada.

Se obtuvieron distintos modelos de predicción antes del tratamiento (Tablas 10, 11 y 12). El mejor modelo estadístico para ambos subtipos virales juntos retiene la duración del tratamiento y el nivel ALT como variables del paciente. La recomendación actual sobre la duración del tratamiento en el caso del genotipo 1 son 48 semanas (Ghany *et al.* 2009), aunque se ha propuesto que la ampliación a 72 semanas puede ser recomendable en aquellos pacientes que no hayan presentado una RVR (Berg *et al.* 2006; Sánchez-Tapias *et al.* 2006; Pearlman *et al.* 2007; Mangia *et al.* 2008) y la reducción a 24 semanas en aquellos que sí presenten RVR (Zeuzem *et al.* 2006). El modelo obtenido es coherente con estos estudios previos ya que, según la interpretación del *odds ratio*, conforme aumenta un mes la

duración del tratamiento la probabilidad de obtener una respuesta positiva, en un contexto global junto con todos los parámetros del modelo, aumenta en 7,96E+18. Sin embargo, no se puede hacer una valoración más precisa con este resultado ya que el estudio del que se tomaron los datos de partida queda restringido a un tratamiento de 48 semanas.

En cuanto a los niveles de ALT, se sabe que los mecanismos de defensa inmune del paciente incluyen la eliminación de las células infectadas mediante los linfocitos *natural killer* (NK) (Golden-Mason *et al.* 2006) y los linfocitos T citotóxicos CD8⁺ (CTL) (Neumann-Haefelin *et al.* 2005), entre otros (Guidotti *et al.* 2006). La inclusión de los niveles de ALT previos al tratamiento en nuestro modelo podría indicar la presencia de actividad de las defensas del paciente. En el modelo propuesto, conforme aumenta una unidad por litro de ALT en suero, la probabilidad de obtener una respuesta positiva al tratamiento aumenta un 33,98 en un contexto global junto con todos los parámetros del modelo. Con estos resultados, podría afirmarse que cuanta mayor sea la actividad citotóxica del sistema inmune del paciente antes del tratamiento, más efectividad tendrá el tratamiento antiviral. En esta línea, el nivel de ALT también ha sido obtenido en otros modelos de predicción (Berg *et al.* 2003; Backus *et al.* 2007; Trapero-Marugan *et al.* 2008; Saito *et al.* 2010; Saludes *et al.* 2010), lo que sostiene nuestros resultados. Sin embargo, se ha visto que la tasa de SVR en pacientes con niveles normales es comparable a la alcanzada por los pacientes con niveles elevados (Zeuzem *et al.* 2004) y que existen otros factores que pueden influenciar los niveles de ALT previos al tratamiento

como el desajuste del metabolismo de grasas y carbohidratos (Prati *et al.* 2006), el abuso de alcohol (Sorbi *et al.* 1999) y otras drogas (Kaplowitz 2004). Por lo tanto, para verificar la influencia de los niveles de ALT en nuestro modelo, sería necesario conocer estos factores en los pacientes empleados para la obtención del modelo. Dado que no disponemos de esta información, habría que repetir la obtención del modelo con un grupo nuevo de pacientes en los que se conociesen estas variables.

Respecto a las coordenadas en el espacio multidimensional de la variación genética viral en las regiones E1E2 y NS5A, este modelo incluye la dimensión 11 de la región NS5A y la dimensión 7 de la región E1E2, que refieren a la presencia de determinados aminoácidos en ciertas posiciones. En la literatura existe un amplio debate sobre las mutaciones que aparecen en ambas regiones y su asociación con la respuesta al tratamiento. Estudios realizados con cohortes de pacientes japoneses han encontrado una correlación entre la respuesta al tratamiento y la aparición de sustituciones en la región NS5A (Enomoto *et al.* 1996; Chayama *et al.* 1997; Kurosaki *et al.* 1997; Fukuda *et al.* 1998; Arase *et al.* 1999; Murashima *et al.* 1999; Nakano *et al.* 1999; Chayama *et al.* 2000; Watanabe *et al.* 2001; Watanabe *et al.* 2005; Murayama *et al.* 2007); en cambio, estudios realizados con pacientes europeos (Khorsi *et al.* 1997; Zeuzem *et al.* 1997; Duverlie *et al.* 1998; Odeberg *et al.* 1998; Rispeter *et al.* 1998; Ibarrola *et al.* 1999; Squadrito *et al.* 1999; Gerotto *et al.* 2000; Sarrazin *et al.* 2000a) y americanos (Hofgärtner *et al.* 1997; Chung *et al.* 1999; Nousbaum *et al.* 2000; Murphy *et al.* 2002) no han visto esta

correlación. Sin embargo, algunos estudios europeos han encontrado una correlación entre la secuencia ISDR del genotipo 1b y la respuesta al IFN (Sáiz *et al.* 1998; Sarrazin *et al.* 1999; Berg *et al.* 2000; Puig-Basagoiti *et al.* 2001; Stratidaki *et al.* 2001; Torres-Puente *et al.* 2008a).

Tras una larga controversia, se ha visto una correlación clara entre la respuesta al tratamiento y la región ISDR de la región NS5A en tres meta-análisis diferentes (Witherell *et al.* 2001; Pascu *et al.* 2004; Schinkel *et al.* 2004) y en estudios más recientes realizados con pacientes de otros países (Yen *et al.* 2008; Kumthip *et al.* 2010). Además, sustituciones en otras zonas de la región NS5A también parecen estar relacionadas con la respuesta al tratamiento (Duverlie *et al.* 1998; Berg *et al.* 2000; Nousbaum *et al.* 2000; Sarrazin *et al.* 2000a; Murphy *et al.* 2002; Sarrazin *et al.* 2002; Layden-Almer *et al.* 2005; Puig-Basagoiti *et al.* 2005; Veillon *et al.* 2007; Wohnsland *et al.* 2007; El-Shamy *et al.* 2008; Muñoz de Rueda *et al.* 2008; Jenke *et al.* 2009; Mori *et al.* 2009; de Queiróz *et al.* 2010; ElHefnawi *et al.* 2010; Kumthip *et al.* 2010).

Recientemente se ha visto una asociación entre una SVR y la aparición de mutaciones conjuntas en las regiones NS5A y *core* (Hayashi *et al.* 2010). Respecto a las sustituciones en la región E1E2, en algunos estudios no se ha encontrado una asociación de la región E2-PePHD con la respuesta al tratamiento (Berg *et al.* 2000; Chayama *et al.* 2000; Gerotto *et al.* 2000; Polyak *et al.* 2000; Puig-Basagoiti *et al.* 2001; Sarrazin *et al.* 2001; Murphy *et al.* 2002; Hung *et al.* 2003; Yang *et al.* 2003; Gaudy *et al.* 2005; Muñoz de Rueda *et al.* 2008; Jenke *et al.*

2009); en la mayoría de ellos debido a que esta región se encuentra muy conservada y su variabilidad no permite diferenciar entre pacientes con respuesta y no respuesta. Sin embargo, otros estudios si han encontrado esta asociación (Sarrazin *et al.* 2000b; Lo *et al.* 2001; Saito *et al.* 2003; Gupta *et al.* 2006; Ukai *et al.* 2006).

Los resultados de nuestro modelo de predicción apoyan que la presencia de determinados variantes de la región NS5A disminuye la probabilidad de obtener una respuesta positiva (*odds ratio* D11 NS5A = 6,33E-06) y la presencia de determinados variantes en la región E1E2 aumentan la probabilidad de obtener una respuesta positiva (*odds ratio* D7 E1E2 = 1458,25), en un contexto global junto con todos los parámetros del modelo. Un aspecto interesante de este resultado es que, además de aparecer estas variables por separado, en el modelo obtenido también aparece la interacción entre ambos parámetros (*odds ratio* D11 NS5A:D7 E1E2 = 1,19E+10), lo que indica que un perfil de sustituciones conjunto entre ambas regiones aumentaría la probabilidad de una respuesta al tratamiento. En este sentido, Ukai y colaboradores (Ukai *et al.* 2006) han visto una correlación entre el número de mutaciones de la región E2 y la región ISDR, y por otro lado, en un trabajo de Torres-Puente y colaboradores (Torres-Puente *et al.* 2008b) se obtuvieron resultados significativos en la correlación de la diversidad nucleotídica entre las regiones E1E2 y NS5A.

Una posible interpretación de estos resultados podría ser que las variantes de ambas regiones puedan interaccionar de

forma epistática debido a la relación funcional o estructural entre ellas, como se ha sugerido previamente entre las regiones E2, NS2 y NS5A en Xu y colaboradores (Xu *et al.* 2008). A pesar de los resultados del modelo y al igual que en un estudio previo (Torres-Puente *et al.* 2008b), con los análisis realizados para esta tesis no se pudo identificar una combinación de aminoácidos asociada claramente con la respuesta.

En cuanto a las variables relativas a la selección viral, el modelo incluye el valor dS de la región E1E2 y el número de sitios bajo selección positiva en la región NS5A. Según nuestros resultados, un aumento del valor dS en la región E1E2 disminuye la probabilidad de obtener una respuesta positiva (*odds ratio* dS E1E2 = 6,75E-93) y un aumento en el número de sitios bajo selección positiva de la región NS5A aumenta la probabilidad de obtener una respuesta positiva (*odds ratio* pts NS5A = 1,30E+10), todas ellas en un contexto global junto con todos los parámetros del modelo. Diversos trabajos han estudiado la relación de la selección positiva que actúa sobre el virus con el escape al sistema inmune (Farci *et al.* 2002; Sheridan *et al.* 2004; Chen *et al.* 2007; Cuevas *et al.* 2008b; Jiménez-Hernández *et al.* 2008; Torres-Puente *et al.* 2008c; Cuevas *et al.* 2009a), que tiene como consecuencia la no respuesta al tratamiento. La región E1E2 incluye varias zonas hipervariables que tienden a acumular cambios aminoacídicos durante la evolución del virus y se sabe que la principal fuerza evolutiva que actúa sobre esta región es la selección purificadora debido a restricciones funcionales (Brown *et al.* 2005; Brown *et al.* 2007). En este sentido, la dS obtenida en nuestro modelo podría indicar la actuación de esta fuerza

evolutiva en la respuesta al tratamiento de forma que cuanto menor sea la ω debido a una alta dS, la probabilidad de respuesta disminuye. En el caso de la región NS5A, un elevado número de sitios bajo selección positiva podría provocar un fallo en la inhibición de la PKR y disminuir la replicación viral. Sin embargo, otros estudios no han encontrado una evidencia de selección positiva en pacientes con SVR (Jiménez-Hernández *et al.* 2008; de Queiróz *et al.* 2010).

Los factores virales referentes a la variabilidad viral intrapaciente que se incluyen en este modelo son la H³ de la región E1E2, la π^2 de la región E1E2 y la H de la región NS5A. Según los resultados obtenidos, el aumento de la H en estas regiones disminuye la probabilidad de obtener una respuesta positiva (*odds ratio* H³ E1E2 = 2,99E-30; *odds ratio* H NS5A = 5,54E-27) y el aumento de la π^2 de la región E1E2 aumenta enormemente la probabilidad de obtener una respuesta positiva (*odds ratio* π^2 E1E2 = inf), todas ellas en un contexto global junto con todos los parámetros del modelo. Los resultados respecto a la H concuerdan con los obtenidos en otros estudios donde la diversidad genética antes del tratamiento es mayor en pacientes sin respuesta que en pacientes con respuesta al tratamiento (Chambers *et al.* 2005; Puig-Basagoiti *et al.* 2005; Cuevas *et al.* 2008a; Torres-Puente *et al.* 2008b; Saludes *et al.* 2010). Una alta diversidad de haplotipos podría promover la aparición de variantes resistentes al tratamiento que permitan al virus escapar de la respuesta inmune, como se ha visto anteriormente (Xu *et al.* 2008). En cambio, los resultados de la π^2 de la región E1E2 indican la

existencia de una diversidad de posiciones nucleotídicas concretas entre pacientes que incapacitaría al virus en su entrada a la célula, en el ensamblaje de nuevas partículas virales, en la neutralización de los anticuerpos y/o en su interacción con la PKR, a pesar de que se sabe que su alta variabilidad podría dificultar la respuesta inmune del paciente (Taylor *et al.* 1999). En este sentido, recientemente se ha visto que una alta variabilidad en la región NS5A está correlacionada con una respuesta positiva al tratamiento (Donlin *et al.* 2010). A pesar de que la diversidad genética no se suele considerar como buen predictor de una SVR (Cuevas *et al.* 2009b), una posible explicación a la aparición de estas variables en el modelo podría ser que al transformar el valor de los parámetros de diversidad genética al elevarlos al cuadrado y al cubo, se obtiene la cantidad de información que retiene estas variables en el modelo predictivo.

En el modelo se incluye también la interacción entre la dS de la región E1E2 y la H³ de la región E1E2. Estos dos parámetros se incluyen, además, por separado aunque con un efecto distinto en la probabilidad de respuesta que el de su interacción (*odds ratio* dS E1E2:H³ E1E2 = 1,20E+85). Este resultado podría indicar que una determinada dS junto con la presencia de algunos haplotipos en la región E1E2 favorecería la respuesta al tratamiento.

El modelo resultante se encuentra equilibrado respecto a las entidades involucradas en los procesos patogénicos y del tratamiento. Los modelos obtenidos para los subtipos 1a y 1b por separado son completamente diferentes. Las variables

incluidas en el mejor modelo para los pacientes infectados con el genotipo 1a son 3 relativas a los parámetros de polimorfismo del virus y las variables incluidas en el modelo para los pacientes con el subtipo 1b son una relativa al tratamiento, 4 relativas a los polimorfismos virales, y una relativa a las coordenadas en el espacio multidimensional. Esta última variable aparece también en el modelo obtenido para ambos subtipos juntos, pero no en el modelo del subtipo 1a. En este sentido, se ha visto una correlación significativa entre las mutaciones de la región ISDR y la respuesta al tratamiento en el genotipo 1b pero no en el genotipo 1a (Torres-Puente *et al.* 2008a). Del modelo de predicción obtenido para el subtipo 1b resulta interesante destacar que todas las variables propias del virus se refieren a la región NS5A.

Las predicciones obtenidas con los modelos (Tabla 13) son muy buenas a pesar de que la muestra es pequeña y se necesitan nuevas pruebas con una muestra mayor de casos para verificar la efectividad de los modelos. Los resultados de las predicciones con los modelos de subtipos separados son correctos para 60 de los 67 pacientes totales. Sin embargo, los resultados de las predicciones con el modelo de subtipos juntos son correctos para 62 de los 67 pacientes totales. Estos resultados muestran que las predicciones con los modelos de subtipos separados son muy similares a las del modelo conjunto de subtipos, incluso algo mejores en este último.

Respecto a los modelos obtenidos con la metodología GLMNET, los resultados con valor lambda 1se reflejan que las variables que se retinen en el modelo conjunto de los genotipos

1a y 1b son las que diferencian a los pacientes con genotipo 1a puesto que no sale ningún resultado para los pacientes con genotipo 1b. Esto indica que los pacientes con genotipo 1b son muy heterogéneos y la regresión no encuentra variables que los separen por tipo de respuesta. Los resultados con el valor mínimo de lambda son menos conservadores que el criterio de la desviación típica y se obtienen modelos de predicción con más variables. En cuanto a las variables comunes con el modelo obtenido por subgrupos de variables de los genotipos 1a+1b, únicamente la diversidad haplotípica de la región NS5A, la duración del tratamiento y la dimensión 11 de la región NS5A lo son. Una observación interesante es que las variables obtenidas en todos los modelos por GLMNET pertenecen a distintos subgrupos, lo que podría indicar que no existe una influencia relevante entre los distintos subgrupos.

No siempre un buen modelo obtenido bajo criterio estadístico tiene una interpretación biológica sencilla o directa para cada variable. Se ha intentado dar una interpretación biológica a cada variable por separado incluida en el modelo estadístico obtenido con subgrupos de variables; sin embargo, para interpretarlo en su totalidad deben considerarse todas las variables. Con las metodologías empleadas, las variables retenidas en los modelos fueron incluidas dependiendo de su significación estadística y hay que resaltar que, tanto por el insuficiente tamaño muestral detectado con GLMNET como por la obtención de errores estándar superiores a los coeficientes en los modelos por subgrupos de variables, son necesarias más comprobaciones con una muestra de pacientes mayor para

poder aplicar estos resultados en la individualización del tratamiento.

4.2 Análisis evolutivo y poblacional de un brote causado por el virus de la hepatitis C.

4.2.1 Introducción

Un brote epidémico hace referencia a una mayor ocurrencia de lo esperado de una enfermedad en un sitio y tiempo concretos. Cuando se produce un brote, uno de los pasos importantes en su investigación es la identificación de la fuente del mismo para adoptar las medidas oportunas de control y eliminación de la misma y, en ocasiones, debido a sus posibles implicaciones penales (González-Candelas 2010). Además, la definición de qué o quienes están incluidos en el brote es fundamental para la prevención y control de la salud pública.

La dificultad para hacer frente a las infecciones provocadas por el HCV se encuentra muy relacionada con los distintos fenómenos que rigen las dinámicas evolutivas de los virus de RNA. Las características de este tipo de virus son bien conocidas (Moya *et al.* 2004):

- Una alta tasa de mutación por la falta de un sistema eficaz de reparación de errores en su RNA polimerasa.
- Un genoma pequeño, entre 3 y 30 kilobases (Kb).
- Grandes tamaños poblacionales, ya que en un mismo individuo pueden llegar hasta 10^{12} las partículas virales presentes en un instante dado.

- Rapidez de replicación (mecanismo por el cual un virus genera nuevos virus), ya que una única partícula infecciosa puede producir una media de 10^5 copias virales en 10 horas.

Entre los mecanismos moleculares que provocan cambios genéticos en los virus, la recombinación no parece ser un mecanismo frecuente en el caso del HCV aunque se han descrito algunos casos (Kalinina *et al.* 2002; Moreno *et al.* 2006; Sentandreu *et al.* 2008; Moreno *et al.* 2009) y el reordenamiento genómico no tiene repercusión alguna.

En los análisis de epidemiología molecular se obtiene una mejor resolución de la relación evolutiva y epidemiológica empleando información de regiones con alta variabilidad en los casos en que ha ocurrido un período corto de tiempo entre la transmisión y el muestreo para el análisis, como en muchos brotes nosocomiales (Bracho *et al.* 2005).

El conocimiento de los procesos y patrones que conducen la evolución de las secuencias virales permite realizar inferencias en su evolución futura y, en el caso de la infección del HCV, esto tiene importantes consecuencias en políticas de salud pública (Jiménez-Hernández *et al.* 2007).

En este estudio se comparan los parámetros de interés evolutivo y genético-poblacional entre las secuencias virales de un brote epidémico del HCV y dos grupos ajenos, uno control y otro externo al brote.

4.2.2 Material y métodos

En Marzo de 1998 fue comunicado un brote de hepatitis C en dos hospitales de la ciudad de Valencia y el grupo de Genética Evolutiva del Instituto Cavanilles de la Universidad de Valencia fue designado para estudiar la relación entre los virus aislados de los pacientes afectados y su relación con la presunta fuente del brote.

Las muestras fueron procesadas y secuenciadas en un estudio, independiente a esta tesis, en el que se emplearon muestras susceptibles de pertenecer al brote y muestras control tomadas de serotecas. Se obtuvo un mínimo de 10 clones en cada muestra estudiada. Las muestras fueron proporcionadas por los hospitales Clínico Universitario, Universitario La Fe y Peset Aleixandre, todos ellos de la ciudad de Valencia.

La información sobre el brote de hepatitis C se almacenó en la base de datos local mediante las herramientas de gestión de datos de la plataforma bioinformática. En total, se incorporó información relativa a 485 pacientes, 517 muestras y 4707 secuencias del genotipo 1a, 4200 pertenecientes a la región E1E2 y 507 a la región NS5B.

En este estudio se utilizaron 361 pacientes: 272 pertenecientes al brote, 47 ajenos al mismo, aunque inicialmente considerados como potencialmente incluidos en él, y 42 pacientes controles completamente ajenos al brote. Una vez almacenados los datos, se recuperaron a través de la plataforma las secuencias de la región E1E2 de cada paciente en pauta y en formato FASTA.

A continuación, se analizaron con las herramientas EPIPATH-TOOLS en un servidor Linux. El flujo de análisis para cada paciente fue el siguiente: se tradujeron las secuencias a proteínas con TRANSEQ, se obtuvo un alineamiento con MAFFT y se volvieron a traducir a secuencias nucleotídicas con REVTRANS. Por último, se lanzaron cada uno de los análisis disponibles en EPIPATH-TOOLS con todos los pacientes.

El análisis estadístico de los resultados se realizó con el programa R. Dado que el tamaño muestral es mayor a 30 individuos en cada grupo, se aplicó el Teorema central del límite y se consideró que los datos siguen una distribución normal. Se realizó el test de Levene para ver si existía homogeneidad de varianzas y se representó en un gráfico el nivel del parámetro de cada individuo para comprobar que las muestras eran independientes. Todas ellas resultaron ser independientes aunque no todas presentaron homogeneidad de varianzas, por tanto se realizó el test de Kruskal-Wallis y sus pruebas *post-hoc* para ver si existían diferencias entre los grupos estudiados. Únicamente en aquellos parámetros con homocedasticidad se realizó también el análisis de la varianza (ANOVA) pero, al no encontrar diferencias entre los resultados del análisis paramétrico y no paramétrico, únicamente se muestran en esta tesis los resultados de los análisis no paramétricos.

4.2.2.1 Resumen del análisis

Las herramientas y procesos informáticos empleados en cada paso del análisis se muestran en la Tabla 18.

Tabla 18. Herramientas/procesos empleados en el análisis.

Paso del análisis	Herramienta/proceso informático
Guardar información del paciente (ficheros en EXCEL) y las secuencias virales en la plataforma (ficheros de texto).	Carga masiva de datos en la base de datos local a través de la plataforma.
Recuperar información de pacientes seleccionados para el análisis y sus secuencias virales en formato FASTA.	Búsqueda en la base de datos local a través de la plataforma.
Script lanzado en linux (*)	
Traducción a secuencias proteicas.	TRANSEQ de EMBOSS (Rice <i>et al.</i> 2000) integrado en la plataforma.
Alineamiento de secuencias.	MAFFT (Katoh <i>et al.</i> 2009) integrado en la plataforma.
Traducción a secuencias nucleotídicas.	REVTRANS (Wernersson 2003), integrado en la plataforma.
Cálculo de parámetros de análisis de polimorfismos, divergencia entre secuencias, cálculo de sustituciones sinónimas y no-sinónimas y tests de neutralidad.	EPIPATH-TOOLS integradas en la plataforma.
Análisis estadístico	
Análisis estadístico de los parámetros calculados y generación de gráficos.	Test de Levene, ANOVA, Kruskal-Wallis y pruebas <i>post-hoc</i> en R. Los gráficos se generaron con la librería GGPlot.

(*) Se generó un script por herramienta de EPIPATH-TOOLS con todos los pasos incluidos en esta fase.

4.2.3 Resultados

Se han comparado todos los grupos entre sí para estudiar las posibles diferencias propias del brote epidemiológico pero también para ver que ocurre en los grupos control y externo. Todos los parámetros mostrados son relativos al número de clones.

En las Figuras 21, 22 y en la Tabla 19 se muestra un resumen de los resultados de los parámetros más relevantes en el estudio de polimorfismos.

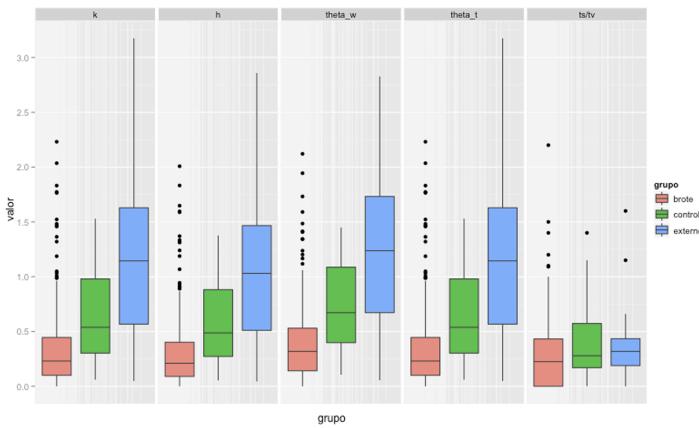


Figura 21. Box-plots del número promedio de diferencias nucleotídicas (k), la heterocigosidad (h), el valor theta de Waterson (theta_w), el valor theta de Tajima (theta_t) y el cociente transiciones/transversiones (ts/tv). En los box-plots, el límite superior de la caja indica el tercer cuartil (percentil 75), el límite inferior de la caja indica el primer cuartil (percentil 25) y la línea que aparece dentro de la caja es la mediana (percentil 50).

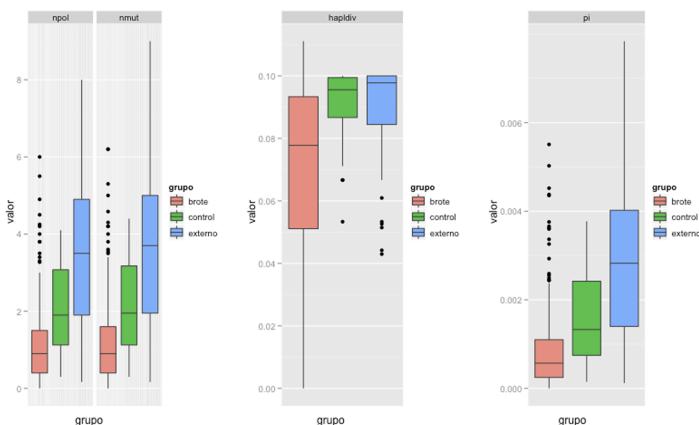


Figura 22. Box-plots del número de polimorfismos (npol), el número de mutaciones (nmut), la diversidad haplotípica (hapldiv) y la diversidad nucleotídica (pi).

Tabla 19. Resumen de las pruebas *post-hoc* de Kruskal-Wallis para los parámetros de polimorfismo (*p*-valor=0,05).

Grupo	npol	μ	κ	h	θ_w	θ_t	H	π	ts/tv
Brote-Control	X	X	X	X	X	X	X	X	-
Brote-Externo	X	X	X	X	X	X	X	X	-
Control-Externo	-	-	-	-	-	-	-	-	-

X: existen diferencias significativas.

-: no existen diferencias significativas.

En las Figuras 23, 24 y la Tabla 20 se muestra un resumen de los resultados de los parámetros más relevantes en el estudio de cambios sinónimos y no-sinónimos.

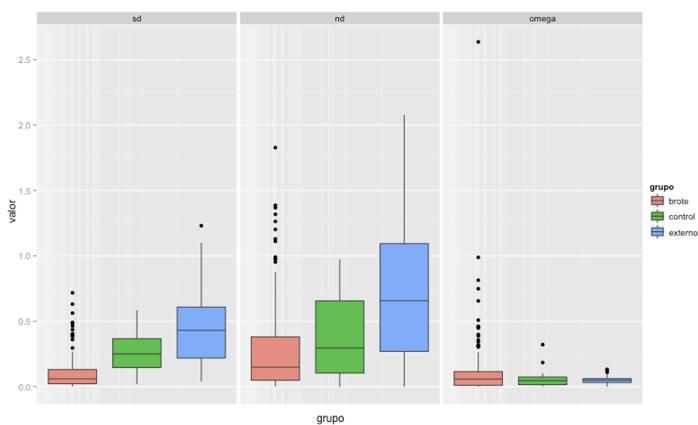


Figura 23. Box-plots del número promedio de sustituciones sinónimas (sd), el número promedio de sustituciones no-sinónimas (nd) y la tasa dN/dS (omega).

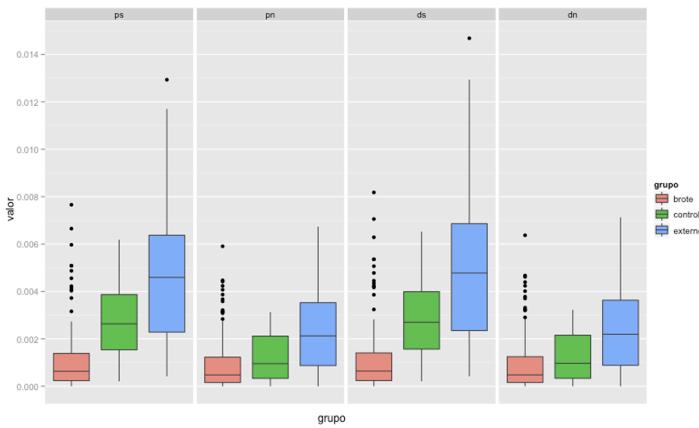


Figura 24. Box-plots de la diversidad nucleotídica sinónima (ps), la diversidad nucleotídica no-sinónima (pn), la diversidad nucleotídica sinónima con Jukes y Cantor (ds) y la diversidad nucleotídica no-sinónima con Jukes y Cantor (dn).

Tabla 20. Resumen de las pruebas *post-hoc* de Kruskal-Wallis para los parámetros sinónimos y no-sinónimos ($p\text{-valor}=0,05$).

Grupo	Sd	Nd	Ps	Pn	dS	dN	Ω
Brote-Control	X	X	X	X	X	X	-
Brote-Externo	X	X	X	X	X	X	-
Control-Externo	-	X	-	X	-	X	-

X: existen diferencias significativas.

-: no existen diferencias significativas.

En las Figuras 25, 26 y la Tabla 21 se muestra un resumen de los resultados de los parámetros más relevantes en el estudio de neutralidad. En este grupo de parámetros, el tamaño muestral del grupo Brote es de 263 pacientes debido a que en las secuencias virales de 9 pacientes no existen sitios polimórficos y mutaciones, por lo que no pueden calcularse los parámetros de neutralidad.

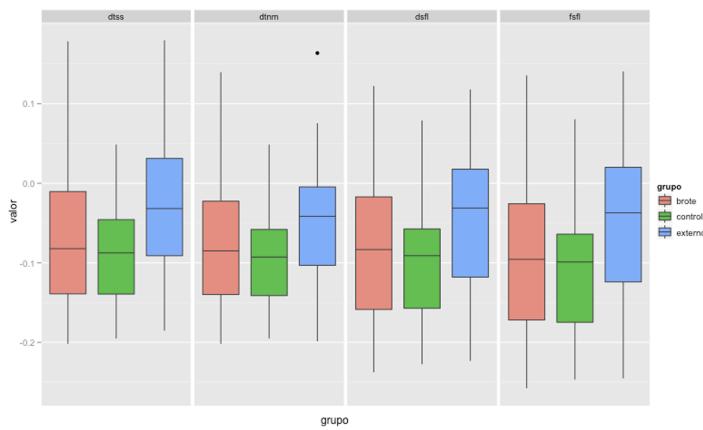


Figura 25. Box-plots de la D de Tajima por sitios segregantes (dtss), la D de Tajima por número mutaciones totales (dtnm), la D* de Fu y Li (dsfl) y la F* de Fu y Li (fsfl).

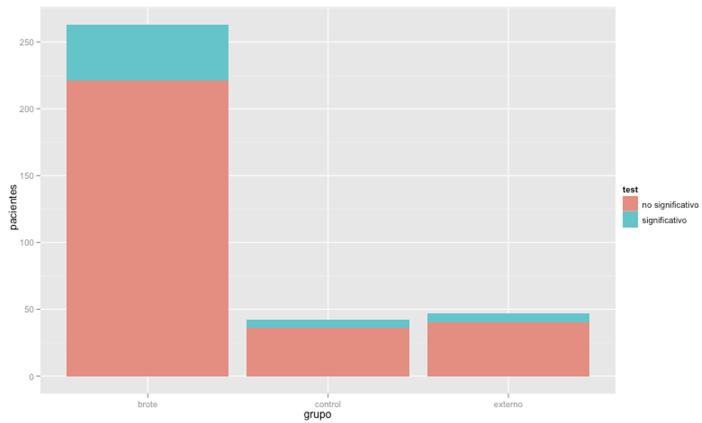


Figura 26. Número de pacientes con alguna prueba de neutralidad significativa.

Tabla 21. Resumen de las pruebas *post-hoc* de Kruskal-Wallis para los parámetros de neutralidad (p -valor=0,05).

Grupo	DTSS	DTNM	DSFL	FSFL
Brote–Control	-	-	-	-
Brote–Externo	X	-	X	X
Control–Externo	X	X	X	X

X: existen diferencias significativas.

-: no existen diferencias significativas.

En las Figuras 27, 28 y la Tabla 22 se muestra un resumen de los resultados de los parámetros más relevantes en el estudio de divergencia entre los distintos grupos.

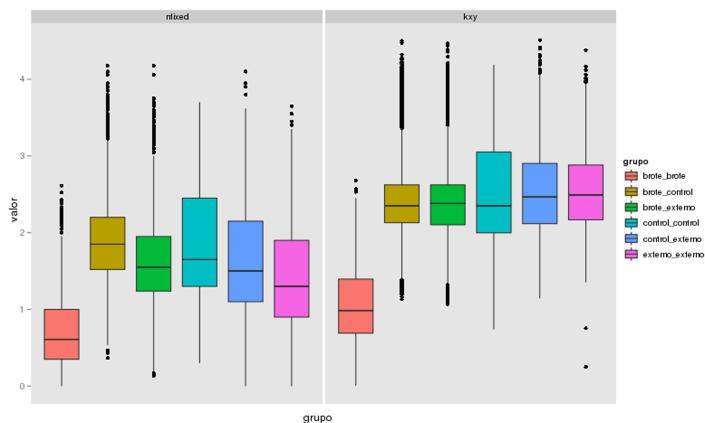


Figura 27. Box-plots del número de diferencias fijadas entre poblaciones (nfixed) y el número promedio de diferencias nucleotídicas entre poblaciones (kxy).

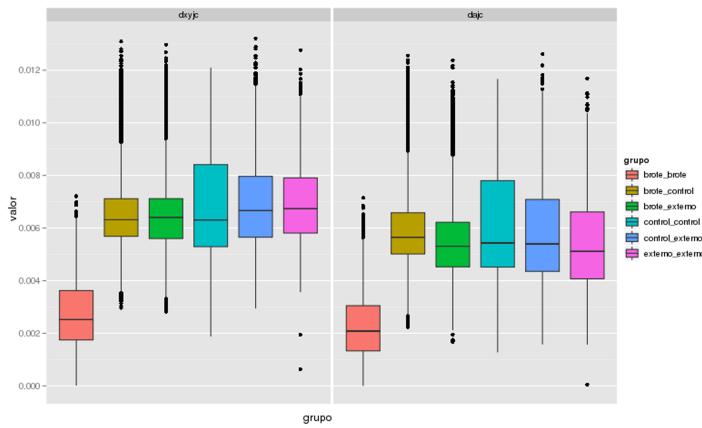


Figura 28. Box-plots de la divergencia entre poblaciones con Jukes y Cantor (dxyjc) y la divergencia neta entre poblaciones con Jukes y Cantor (dajc).

Tabla 22. Resumen de las pruebas *post-hoc* de Kruskal-Wallis para los parámetros de divergencia (p -valor=0,05).

Grupo	nfixed	κ	$D_{xy} JC$	$D_a JC$
Brote Brote – Brote Control	X	X	X	X
Brote Brote – Brote Externo	X	X	X	X
Brote Brote – Control Control	X	X	X	X
Brote Brote – Control Externo	X	X	X	X
Brote Brote – Externo Externo	X	X	X	X
Brote Control – Brote Externo	X	-	-	X
Brote Control – Control Control	X	-	-	X
Brote Control – Control Externo	X	X	X	X
Brote Control – Externo Externo	X	X	X	X
Brote Externo – Control Control	X	-	-	-
Brote Externo – Control Externo	-	X	X	-
Brote Externo – Externo Externo	X	X	X	-
Control Control – Control Externo	X	-	-	-
Control Control – Externo Externo	X	-	X	X
Control Externo – Externo Externo	X	-	-	-

X: existen diferencias significativas.

-: no existen diferencias significativas.

4.2.4 Discusión

Los virus pertenecientes al grupo Brote son significativamente diferentes a los de los grupos Control y Externo en todos los parámetros que describen su polimorfismo salvo en la tasa de transiciones/transversiones (Tabla 19). En general, los virus del grupo Brote tienen una diversidad relativa al número de clones menor que el resto de grupos comparados y su diversidad haplotípica se distribuye en un amplio rango de valores (Figuras 21 y 22). Estos resultados indican un origen relativamente reciente de la población viral incluida en el grupo Brote respecto a las de los otros grupos.

Los parámetros estadísticos que estudian las posiciones sinónimas y no-sinónimas tienen valores relativos al número de clones más pequeños en el grupo Brote que en los grupos Control y Externo. Sin embargo, *Omega* es el doble en el grupo Brote respecto a los otros dos grupos (Figura 23) pero no existen diferencias significativas en las comparaciones *post-hoc* entre ellos (Tabla 20). *Omega* es un indicador de la presión selectiva a nivel de genes codificantes de proteínas (Yang *et al.* 2000). En los resultados de la Figura 23, el valor *Omega* indica la presencia de selección negativa o purificadora característica de regiones génicas conservadas (Lu *et al.* 2001; Alfonso *et al.* 2004). Sin embargo, dada la menor antigüedad de la infección de los pacientes del grupo Brote, el mayor valor relativo de *Omega* podría explicarse por la insuficiente acción de la selección purificadora en el conjunto de estas poblaciones virales junto con el efecto de los cuellos de botella (con un importante efecto de la deriva genética) lo que llevaría a un comportamiento más

próximo a la neutralidad (correspondiente a $\Omega\omega=1$) que en las muestras procedentes de infecciones más antiguas.

En cuanto a las pruebas de neutralidad, al menos uno de los parámetros presenta valor significativamente menor que cero en 55 individuos (42 casos en el grupo Brote, 6 casos en el grupo Control y 7 casos en el grupo Externo) (Figura 26). Todos los grupos presentan una mediana relativa al número de clones negativa (Figura 25), lo que indica una selección direccional, en la que se favorece la fijación de mutaciones ventajosas o eliminación de las deletéreas. Sin embargo, dado que el rango intercuartílico incluye al cero en la mayoría de los parámetros estadísticos del grupo Externo, se podría contemplar otro tipo de selección para este grupo de virus. En general, se obtienen diferencias significativas entre el grupo Externo con los grupos Brote y Control (Tabla 21).

La divergencia dentro del grupo Brote es bastante menor que la existente entre el resto de comparaciones (Figuras 27 y 28) y esta diferencia es significativa en todos los parámetros relevantes (Tabla 22). Este hecho muestra que el grupo Brote se encuentra bien definido respecto a los grupos de comparación ya que sus secuencias virales han evolucionado en un período de tiempo menor en comparación con los otros grupos estudiados. Respecto a si las muestras del grupo Brote son diferentes de las del grupo Control, la divergencia neta es significativamente diferente entre la comparación Brote-Control y las divergencias internas de ambos grupos por separado. En cuanto al grupo Externo, la divergencia entre poblaciones con Jukes y Cantor es significativamente diferente

entre la comparación Brote-Externo y sus respectivas divergencias internas por separado, pero no la divergencia neta entre el grupo Brote-Externo y el grupo Externo.

5. DISCUSIÓN GENERAL

5 Discusión general

Son muchos los datos que se manejan en la investigación genética y epidemiológica de las poblaciones de patógenos y, normalmente en un laboratorio de tamaño medio, se encuentran almacenados de forma independiente en distintos formatos, lo que dificulta su manejo y análisis. Esta organización de los datos conlleva que el trabajo de investigación se realice de forma más lenta y desordenada en comparación con la eficiencia de emplear un SI sistematizado.

El objetivo principal de este trabajo ha consistido en el desarrollo de un SI orientado al área de Epidemiología molecular, que permite almacenar y analizar de forma centralizada la información clínica de pacientes junto con la información molecular de los patógenos de interés. La plataforma presentada en esta tesis consta de dos grupos de herramientas diferenciadas pero, a su vez, diseñadas para interactuar entre sí facilitando el manejo y análisis de grandes cantidades de datos. Por un lado se ha diseñado e implementado una base de datos junto con varias aplicaciones que facilitan su gestión y, por otro lado, se han integrado y desarrollado diferentes aplicaciones de análisis que permiten realizar estudios evolutivos y poblacionales de las secuencias almacenadas.

Existen diversas bases de datos públicas que constituyen un repositorio de secuencias e información molecular de patógenos con alguna herramienta de análisis (Kulkarni-Kale *et al.* 2004; Aurrecoechea *et al.* 2007;

Gnaneshan *et al.* 2007; Hirahata *et al.* 2007; McNeil *et al.* 2007; Snyder *et al.* 2007; Yang *et al.* 2008; Belshaw *et al.* 2009), pero no permiten el almacenamiento y la gestión de las secuencias obtenidas en un laboratorio de forma segura. En este sentido, existen algunos trabajos en los que se ha desarrollado un LIMS para facilitar las tareas de laboratorio relativas al genotipado de secuencias (Van Rossum *et al.* 2010), rastreo de mutaciones (Voegele *et al.* 2007) y secuenciación (Wendl *et al.* 2007), pero sin integrar información clínica y otras aplicaciones de análisis.

En cuanto a las herramientas de análisis, una de las nuevas aportaciones de esta plataforma es que permite lanzar diversos análisis por línea de comandos, lo que facilita el análisis de una gran cantidad de datos simultáneamente. Hasta ahora, con algunos programas e implementaciones existentes había que lanzar los análisis uno a uno a través de una *web* o un programa de escritorio, como el caso de DNASP, MEGA (Kumar *et al.* 2008) entre otros, con la consecuente ralentización del estudio. Concretamente, las aplicaciones implementadas en esta plataforma con la opción de lanzarlas por líneas de comandos son BLAST remoto frente al NCBI, frente a EPIPATH y local frente a un fichero; los programas de EMBOSS como TRANALIGN y TRANSEQ; los programas MSA como MUSCLE, T-COFFEE, MAFFT y PROBCONS; los programas desarrollados específicamente para esta plataforma que son EPIPATH-POLYMORPHISM, EPIPATH-DIVERGENCE, EPIPATH-SYNONYMOUS y EPIPATH-NEUTRALITY; COMPUTE, el análisis estadístico por coalescencia y REVTRANS.

Hasta la fecha, son muy pocas las plataformas bioinformáticas que integran simultáneamente información clínica y molecular de patógenos junto con herramientas de análisis.

LANL-HCV y EUHCVDB, dos de las tres grandes bases de datos públicas específicas de HCV, disponen de secuencias públicas del virus y de algunas herramientas de análisis interesantes pero que funcionan a través de un servidor *web* en el que hay que realizar los análisis uno a uno. Otra desventaja es que no permiten almacenar y gestionar las propias secuencias de forma segura.

PATHEMA (Brinkac *et al.* 2010) agrupa información sobre 6 patógenos de interés prioritario en bio-defensa categorizados por el Instituto Nacional sobre Alergias y Enfermedades Infecciosas (NIAID) que son *Bacillus anthracis*, *Clostridium botulinum*, *Burkholderia mallei*, *Burkholderia pseudomallei*, *Clostridium perfringens* y *Entamoeba histolytica*. Esta plataforma integra secuencias moleculares de los patógenos seleccionados junto con datos epidemiológicos y clínicos, además de herramientas de análisis como BLAST y MSA entre otras. Se encuentra orientada a estudios de Genómica comparativa y funcional. A diferencia de la plataforma desarrollada en este trabajo, se encuentra limitada a los 6 patógenos mencionados, no constituye un LIMS para laboratorios de tamaño medio, la información se encuentra disponible públicamente y no permite búsquedas por perfiles clínicos de pacientes.

GEMINA (Schrimal *et al.* 2010) integra información epidemiológica y clínica sobre patógenos de las categorías A-C del NIAID. Se encuentra orientada a la vigilancia de brotes epidemiológicos y obtiene la información de la literatura científica mediante el empleo de ontologías. Es una base de datos pública y no tiene funcionalidades propias de un LIMS. No dispone de herramientas de análisis ni permite la recuperación de secuencias.

SEQHEPB (Yuen *et al.* 2007) integra datos clínicos y demográficos de pacientes junto con secuencias del virus de la hepatitis B. Tiene un programa de análisis de secuencias con el que identifica el genotipo, serotipo y diversas mutaciones. Esta plataforma no es de libre distribución y no se encuentra disponible públicamente. Además, la herramienta de análisis es muy limitada y únicamente almacena información sobre un virus en concreto.

PASSIM (Viksna *et al.* 2007) integra información demográfica de individuos y muestras. Esta plataforma está orientada a estudios de Epidemiología genómica, en los que las secuencias pertenecen a material genético de los pacientes. Por tanto, a pesar de que el LIMS que utiliza pueda ser muy similar al desarrollado aquí, no comparte los objetivos de esta tesis centrada en el estudio de patógenos.

El trabajo de Araújo y colaboradores (Araújo *et al.* 2006) es, hasta la fecha, el SI más similar al presentado en esta tesis. Aunque son varias las características de DBCOLLHIV que limitan la consecución de los objetivos propuestos en este trabajo. Por un lado, se encuentra orientado únicamente al

estudio del VIH, lo que restringe la plataforma a un único patógeno. Por otro lado, la base de datos que utiliza no tiene un módulo donde almacenar información sobre los procesos de laboratorio, por lo que no tiene una utilidad como LIMS, y, además, no contempla el almacenaje de información sobre brotes epidemiológicos. En cuanto a las herramientas de análisis, integra aplicaciones para comprobar contaminaciones en PCRs, interpretar resistencias a fármacos siguiendo las recomendaciones del Ministerio de Salud brasileño, además de incluir un formulario automatizado de recombinantes y una herramienta de sub-tipado del VIH. Todas estas herramientas son específicas para estudios sobre VIH y no se encuentran orientadas a los estudios evolutivos y poblacionales.

Por otro lado, la plataforma desarrollada en este trabajo también facilita la vigilancia epidemiológica basada en señales específicas de la enfermedad. Este proceso se puede realizar con la identificación de regiones patógenas a través de la comparación de secuencias cuyos pacientes tengan un perfil clínico determinado (Sintchenko 2009; Sintchenko *et al.* 2009a; Sintchenko *et al.* 2009b).

Respecto a las ampliaciones y mejoras de la plataforma presentada en este trabajo que podrían llevarse a cabo en un futuro, se encuentra, principalmente, la integración de herramientas de análisis filogenético dada su utilidad en estudios de Epidemiología molecular (Fitch *et al.* 2001; Tibayrenc 2005; Hall *et al.* 2006). Como se mencionó previamente en esta tesis, en la plataforma existe una sección para almacenar información relativa a los análisis filogenéticos

realizados con secuencias disponibles en la base de datos local. Sin embargo, estas herramientas no se han integrado todavía en esta plataforma bioinformática. En PHYLEMON (Tárraga *et al.* 2007) se han implementado diversas herramientas para realizar este tipo de análisis pero todos los datos a analizar los debe enviar el usuario uno a uno, por lo que el proceso de análisis resulta tedioso si se tienen muchos datos. En este sentido, sería interesante integrar las herramientas de análisis filogenético para completar el *workflow* presentado en esta tesis. También podrían integrarse todas aquellas herramientas de análisis que sean de interés como VIROBLAST (Deng *et al.* 2007). Por otro lado, dada la relevancia de las nuevas técnicas de secuenciación masiva (Metzker 2010; Sorek *et al.* 2010) podría añadirse algún módulo que transfiriese esta nueva información a la base de datos local. Además, podría mejorarse el modo de compartir información con otros SI mediante el empleo de ficheros con formatos estándar como el *eXtensible Markup Language* (XML), uno de los formatos más utilizados actualmente para el intercambio y transmisión de información (He *et al.* 2005; Liu *et al.* 2008). Finalmente, existen iniciativas como EPICOLLECT (Aanensen *et al.* 2009) que aplican las tecnologías móviles a la recolección de datos epidemiológicos y que podrían resultar interesantes para futuras ampliaciones.

6. CONCLUSIONES

6 Conclusiones

1. Una correcta gestión de las grandes cantidades de datos generadas en un laboratorio es esencial para poder realizar eficientemente estudios de Epidemiología molecular, Genética de poblaciones y Biología evolutiva.
2. El almacenamiento conjunto de datos de pacientes y patógenos en un SGBD centraliza la búsqueda y la gestión de información, facilitando comparaciones entre ellos y estudios sobre infecciones múltiples.
3. La base de datos desarrollada en esta tesis es, hasta la fecha, la más completa y versátil orientada al estudio de la epidemiología molecular de patógenos infecciosos, ya que permite estudiarlos desde sus secuencias genéticas, en cuanto a la variación que existe dentro de los distintos patógenos y los cambios que registran a lo largo de su evolución, hasta su distribución en distintos pacientes con diferentes afecciones y cuadros clínicos.
4. La integración de herramientas de análisis junto con la información almacenada en la base de datos facilita las tareas del flujo de trabajo habitual en un laboratorio.
5. Los programas de análisis de genética poblacional y evolutiva creados durante esta tesis son de gran utilidad y permiten el análisis masivo de datos.

6. El SI presentado ayuda a los investigadores en sus tareas diarias de laboratorio y en colaboraciones con otras instituciones, preservando la información del acceso público.
7. La visión en conjunto que ofrece esta plataforma permite obtener una mayor comprensión de la interacción entre paciente y patógeno y es una herramienta con aplicación inmediata en la mejora de los tratamientos.
8. Los resultados obtenidos en el modelo de predicción de la respuesta al tratamiento en pacientes con hepatitis C han demostrado que variables del virus no incluidas hasta ahora en otros estudios tienen un papel relevante en la predicción de la respuesta al tratamiento. Sin embargo, es necesario verificar estos resultados debido al tamaño muestral empleado.
9. Un SI eficiente es esencial en situaciones donde la rapidez en la toma de decisiones es un factor crítico, como el caso de los brotes epidemiológicos. El conocimiento de los procesos y patrones que conducen la evolución de las secuencias virales permite realizar inferencias en su evolución futura y, en el caso de la infección del HCV, esto tiene importantes consecuencias en políticas de salud pública.

7. BIBLIOGRAFÍA

7 Bibliografía

Aanensen, D.M., Huntley, D.M., Feil, E.J., al-Owain, F.a. and Spratt, B.G. (2009) EpiCollect: linking smartphones to web applications for epidemiology, ecology and community data collection, *PLoS ONE*, **4**, e6968.

Akaike, H. (1974) A new look at the statistical model identification, *Automatic Control, IEEE Transactions on*, **19**, 716 - 723.

Akuta, N., Suzuki, F., Kawamura, Y., Yatsuji, H., Sezaki, H., Suzuki, Y., Hosaka, T., Kobayashi, M., Kobayashi, M., Arase, Y., Ikeda, K. and Kumada, H. (2007) Predictive factors of early and sustained responses to peginterferon plus ribavirin combination therapy in Japanese patients infected with hepatitis C virus genotype 1b: amino acid substitutions in the core region and low-density lipoprotein cholesterol levels, *Journal of Hepatology*, **46**, 403-410.

Alfonso, V., Flichman, D.M., Sookoian, S., Mbayed, V.A. and Campos, R.H. (2004) Evolutionary study of HVR1 of E2 in chronic hepatitis C virus infection, *J Gen Virol*, **85**, 39-46.

Alter, H. (2006) Viral hepatitis, *Hepatology*, **43**, S230-234.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool, *J Mol Biol*, **215**, 403-410.

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

Amadoz, A. and González-Candelas, F. (2007) epiPATH: an information system for the storage and management of molecular epidemiology data from infectious pathogens, *BMC Infect Dis*, **7**, 32.

Ambrosone, C.B. and Kadlubar, F.F. (1997) Toward an integrated approach to molecular epidemiology, *Am J Epidemiol*, **146**, 912-918.

Ansorge, W.J. (2009) Next-generation DNA sequencing techniques, *N Biotechnol*, **25**, 195-203.

Arase, Y., Ikeda, K., Chayama, K., Murashima, N., Tsubota, A., Suzuki, Y., Saitoh, S., Kobayashi, M., Kobayashi, M., Kobayashi, M. and Kumada, H. (1999) Efficacy and changes of the nonstructural 5A GENE by prolonged interferon therapy for patients with hepatitis C virus genotype 1b and a high level of serum HCV-RNA, *Intern Med*, **38**, 461-466.

Araújo, L.V., Soares, M.A., Oliveira, S.M., Chequer, P., Tanuri, A., Sabino, E.C. and Ferreira, J.E. (2006) DBCollHIV: a database system for collaborative HIV analysis in Brazil, *Genet Mol Res*, **5**, 203-215.

Arias-Coello, A. and Portela-Filgueiras, I. (2006) Sistema de información y sistema de calidad: relación y dependencia en las organizaciones empresariales, *revistas.ucm.es*.

Attwood, T.K. and Parry-Smith, D.J. (1999) Introduction to bioinformatics. *Addison Wesley Longman Limited*, 218 pp.

Aurrecoechea, C., Heiges, M., Wang, H., Wang, Z., Fischer, S., Rhodes, P., Miller, J., Kraemer, E., Stoeckert, C.J., Roos, D.S. and Kissinger, J.C. (2007) ApiDB: integrated resources for the apicomplexan bioinformatics resource center, *Nucleic Acids Research*, **35**, D427-430.

Backus, L.I., Boothroyd, D.B., Phillips, B.R. and Mole, L.A. (2007) Predictors of response of US veterans to treatment for the hepatitis C virus, *Hepatology*, **46**, 37-47.

Barr, J.N. and Fearns, R. (2010) How RNA viruses maintain their genome integrity, *J Gen Virol*, **91**, 1373-1387.

Barreto, M.L., Teixeira, M.G. and Carmo, E.H. (2006) Infectious diseases epidemiology, *Journal of epidemiology and community health*, **60**, 192-195.

Batzoglou, S. (2005) The many faces of sequence alignment, *Briefings in Bioinformatics*, **6**, 6-22.

Belshaw, R., de Oliveira, T., Markowitz, S. and Rambaut, A. (2009) The RNA virus database, *Nucleic Acids Research*, **37**, D431-435.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. (2011) GenBank, *Nucleic Acids Research*, **39**, D32-37.

Berg, T., Mas Marques, A., Höhne, M., Wiedenmann, B., Hopf, U. and Schreier, E. (2000) Mutations in the E2-PePHD and NS5A region of hepatitis C virus type 1 and the dynamics of hepatitis C viremia decline during interferon alfa treatment, *Hepatology*, **32**, 1386-1395.

Berg, T., Sarrazin, C., Herrmann, E., Hinrichsen, H., Gerlach, T., Zachoval, R., Wiedenmann, B., Hopf, U. and Zeuzem, S. (2003) Prediction of treatment outcome in patients with chronic hepatitis C: significance of baseline parameters and viral dynamics during therapy, *Hepatology*, **37**, 600-609.

Berg, T., von Wagner, M., Nasser, S., Sarrazin, C., Heintges, T., Gerlach, T., Buggisch, P., Goeser, T., Rasenack, J., Pape, G.R., Schmidt, W.E., Kallinowski, B., Klinker, H., Spengler, U., Martus, P., Alshuth, U. and Zeuzem, S. (2006) Extended treatment duration for hepatitis C virus type 1: comparing 48 versus 72 weeks of peginterferon-alfa-2a plus ribavirin, *Gastroenterology*, **130**, 1086-1097.

Bittar, C., Jardim, A.C.G., Yamasaki, L.H.T., de Queiróz, A.T.L., Carareto, C.M.A., Pinho, J.R.R., de Carvalho-Mello, I.M.V.G. and Rahal, P. (2010) Genetic diversity of NS5A protein from hepatitis C virus genotype 3a and its relationship to therapy response, *BMC Infect Dis*, **10**, 36.

Bonhoeffer, S., May, R.M., Shaw, G.M. and Nowak, M.A. (1997) Virus dynamics and drug therapy, *Proc Natl Acad Sci USA*, **94**, 6971-6976.

Bracho, M.A., Gosalbes, M.J., Blasco, D., Moya, A. and González-Candelas, F. (2005) Molecular epidemiology of a hepatitis C virus outbreak in a hemodialysis unit, *J Clin Microbiol*, **43**, 2750-2755.

Brazhnik, O. and Jones, J.F. (2007) Anatomy of data integration, *Journal of Biomedical Informatics*, **40**, 252-269.

Brinkac, L.M., Davidsen, T., Beck, E., Ganapathy, A., Caler, E., Dodson, R.J., Durkin, A.S., Harkins, D.M., Lorenzi, H., Madupu, R., Sebastian, Y., Shrivastava, S., Thiagarajan, M., Orvis, J., Sundaram, J.P., Crabtree, J., Galens, K., Zhao, Y., Inman, J.M., Montgomery, R., Schobel, S., Galinsky, K., Tanenbaum, D.M., Resnick, A., Zafar, N., White, O. and Sutton, G. (2010) Pathema: a clade-specific bioinformatics resource center for pathogen research, *Nucleic Acids Research*, **38**, D408-414.

Brown, R.J.P., Juttl, V.S., Tarr, A.W., Finn, R., Irving, W.L., Hemsley, S., Flower, D.R., Borrow, P. and Ball, J.K. (2005) Evolutionary dynamics of hepatitis C virus envelope genes during chronic infection, *J Gen Virol*, **86**, 1931-1942.

Brown, R.J.P., Tarr, A.W., McClure, C.P., Juttl, V.S., Tagiuri, N., Irving, W.L. and Ball, J.K. (2007) Cross-genotype characterization of genetic diversity and molecular adaptation in hepatitis C virus envelope glycoprotein genes, *J Gen Virol*, **88**, 458-469.

Brown, R.S. (2007) Customizing treatment to patient populations, *Nature clinical practice Gastroenterology & hepatology*, **4 Suppl 1**, S3-9.

Campoccia, D., Montanaro, L. and Arciola, C.R. (2009) Current methods for molecular epidemiology studies of implant infections, *Int J Artif Organs*, **32**, 642-654.

Casillas, S. (2008) Development and application of bioinformatic tools for the representation and analysis of genetic diversity, *Tesis doctoral. Universidad Autónoma de Barcelona*.

Cattell, R. (1966) The scree test for the number of factors, *Multivariate Behavioral Research*.

Chambers, T.J., Fan, X., Droll, D.A., Hembrador, E., Slater, T., Nickells, M.W., Dustin, L.B. and Dibisceglie, A.M. (2005) Quasispecies heterogeneity within the E1/E2 region as a pretreatment variable during pegylated interferon therapy of chronic hepatitis C virus infection, *Journal of Virology*, **79**, 3071-3083.

Chanzá, D. (2008) Proceso de creación de la página web personal dvwebz, *Tesis Máster Universitario de Aplicaciones Multimedia para Internet. Escuela Técnica Superior de Ingeniería del Diseño. Universidad Politécnica de Valencia.*

Charles, E.D. and Dustin, L.B. (2009) Hepatitis C virus-induced cryoglobulinemia, *Kidney Int*, **76**, 818-824.

Chayama, K., Suzuki, F., Tsubota, A., Kobayashi, M., Arase, Y., Saitoh, S., Suzuki, Y., Murashima, N., Ikeda, K., Takahashi, N., Kinoshita, M. and Kumada, H. (2000) Association of amino acid sequence in the PKR-eIF2 phosphorylation homology domain and response to interferon therapy, *Hepatology*, **32**, 1138-1144.

Chayama, K., Tsubota, A., Kobayashi, M., Okamoto, K., Hashimoto, M., Miyano, Y., Koike, H., Kobayashi, M., Koida, I., Arase, Y., Saitoh, S., Suzuki, Y., Murashima, N., Ikeda, K. and Kumada, H. (1997) Pretreatment virus load and multiple amino acid substitutions in the interferon sensitivity-determining region predict the outcome of interferon treatment in patients with chronic genotype 1b hepatitis C virus infection, *Hepatology*, **25**, 745-749.

Chen, S. and Wang, Y.-m. (2007) Evolutionary study of hepatitis C virus envelope genes during primary infection, *Chin Med J*, **120**, 2174-2180.

Chung, R.T., Monto, A., Dienstag, J.L. and Kaplan, L.M. (1999) Mutations in the NS5A region do not predict interferon-responsiveness in american patients infected with genotype 1b hepatitis C virus, *J. Med. Virol.*, **58**, 353-358.

Clarke, D.K., Duarte, E.A., Elena, S.F., Moya, A., Domingo, E. and Holland, J. (1994) The red queen reigns in the kingdom of RNA viruses, *Proc Natl Acad Sci USA*, **91**, 4821-4824.

Cochrane, G., Karsch-Mizrachi, I., Nakamura, Y. and Collaboration, I.N.S.D. (2011) The International Nucleotide Sequence Database Collaboration, *Nucleic Acids Research*, **39**, D15-18.

Codd, E. (1970) A relational model of data for large shared data banks, *Communications of the ACM*, **13**.

Combet, C., Garnier, N., Charavay, C., Grando, D., Crisan, D., Lopez, J., Dehne-Garcia, A., Geourjon, C., Bettler, E., Hulo, C., Le Mercier, P., Bartenschlager, R., Diepolder, H., Moradpour, D., Pawlotsky, J.-M., Rice, C.M., Trépo, C., Penin, F. and Deléage, G. (2007) euHCVdb: the European hepatitis C virus database, *Nucleic Acids Research*, **35**, D363-366.

Crick, F. (1970) Central dogma of molecular biology, *Nature*, **227**, 561-563.

Cuevas, J.M., Gonzalez, M., Torres-Puente, M., Jiménez-Hernández, N., Bracho, M.A., García-Robles, I., González-Candelas, F. and Moya, A. (2009a) The role of positive selection in hepatitis C virus, *Infect Genet Evol*, **9**, 860-866.

Cuevas, J.M., Moya, A. and Sanjuán, R. (2005) Following the very initial growth of biological RNA viral clones, *J Gen Virol*, **86**, 435-443.

Cuevas, J.M., Torres-Puente, M., Jiménez-Hernández, N., Bracho, M.A., García-Robles, I., Carnicer, F., Olmo, J.D., Ortega, E., González-Candelas, F. and Moya, A. (2009b) Combined therapy of interferon plus ribavirin promotes multiple adaptive solutions in hepatitis C virus, *J. Med. Virol.*, **81**, 650-656.

Cuevas, J.M., Torres-Puente, M., Jiménez-Hernández, N., Bracho, M.A., García-Robles, I., Carnicer, F., Olmo, J.D., Ortega, E., Moya, A. and González-Candelas, F. (2008a) Refined analysis of genetic variability parameters in hepatitis C virus and the ability to predict antiviral treatment response, *J Viral Hepat*.

Cuevas, J.M., Torres-Puente, M., Jiménez-Hernández, N., Bracho, M.A., García-Robles, I., Wrobel, B., Carnicer, F., del Olmo, J., Ortega, E., Moya, A. and González-Candelas, F. (2008b) Genetic variability of hepatitis C virus before and after combined therapy of interferon plus ribavirin, *PLoS ONE*, **3**, e3058.

Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) A model of evolutionary change in proteins, In Dayhoff, M.O. and Erch, R.V. (Eds), *Atlas of Protein Sequence Structure. National Biomedical Research Foundation, Maryland.*, 345-352.

de Queiróz, A.T.L., Maracaja-Coutinho, V., Jardim, A.C.G., Rahal, P., de Carvalho-Mello, I.M.V.G. and Matioli, S.R. (2010) Relation of pretreatment sequence diversity in NS5A region of HCV genotype 1 with immune response between pegylated-INF/ribavirin therapy outcomes, *J Viral Hepat.*

Dean, A.G., Dean, J.A., Burton, A.H. and Dicker, R.C. (1991) Epi Info: a general-purpose microcomputer program for public health information systems, *Am J Prev Med*, **7**, 178-182.

Deng, W., Nickle, D.C., Learn, G.H., Maust, B. and Mullins, J.I. (2007) ViroBLAST: a stand-alone BLAST web server for flexible queries of multiple databases and user's datasets, *Bioinformatics*, **23**, 2334-2336.

Desmet, V.J. (2003) Knodell RG, Ishak KG, Black WC, Chen TS, Craig R, Kaplowitz N, Kiernan TW, Wollman J. Formulation and application of a numerical scoring system for assessing histological activity in asymptomatic chronic active hepatitis [Hepatology 1981;1:431-435], *J.Hepatol.*, **38**, 382-386.

Dirección General de Salud Pública (2009) Informe Hepatitis, *Generalitat Valenciana. Consellería de Sanitat.*

Do, C.B., Mahabhashyam, M.S.P., Brudno, M. and Batzoglou, S. (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment, *Genome Res.*, **15**, 330-340.

Domingo, E. and Holland, J. (1997) RNA virus mutations and fitness for survival, *Annu. Rev. Microbiol.*, **51**, 151-178.

Donlin, M.J., Cannon, N.A., Aurora, R., Li, J., Wahed, A.S., Di Bisceglie, A.M., Tavis, J.E. and Group, V.-C.S. (2010) Contribution of genome-wide HCV genetic differences to outcome of interferon-based therapy in Caucasian American and African American patients, *PLoS ONE*, **5**, e9032.

Duarte, E., Clarke, D., Moya, A., Domingo, E. and Holland, J. (1992) Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet, *Proc Natl Acad Sci USA*, **89**, 6015-6019.

Dubuisson, J. (2007) Hepatitis C virus proteins, *World J Gastroenterol*, **13**, 2406-2415.

Duffy, S., Shackelton, L.A. and Holmes, E.C. (2008) Rates of evolutionary change in viruses: patterns and determinants, *Nat Rev Genet*, **9**, 267-276.

Dutheil, J., Gaillard, S., Bazin, E., Glémin, S., Ranwez, V., Galtier, N. and Belkhir, K. (2006) Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics, *BMC Bioinformatics*, **7**, 188.

Duverlie, G., Khorsi, H., Castelain, S., Jaillon, O., Izopet, J., Lunel, F., Eb, F., Penin, F. and Wychowski, C. (1998) Sequence analysis of the NS5A protein of European hepatitis C virus 1b isolates and relation to interferon sensitivity, *J Gen Virol*, **79 (Pt 6)**, 1373-1381.

Edgar, R.C. (2004a) MUSCLE: a multiple sequence alignment method with reduced time and space complexity, *BMC Bioinformatics*, **5**, 113.

Edgar, R.C. (2004b) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Research*, **32**, 1792-1797.

Eigen, M., McCaskill, J. and Schuster, P. (1988) Molecular quasi-species, *The Journal of Physical Chemistry*.

El-Shamy, A., Nagano-Fujii, M., Sasase, N., Imoto, S., Kim, S.-R. and Hotta, H. (2008) Sequence variation in hepatitis C virus nonstructural protein 5A predicts clinical outcome of pegylated interferon/ribavirin combination therapy, *Hepatology*, **48**, 38-47.

Elena, S.F., González-Candelas, F., Novella, I.S., Duarte, E.A., Clarke, D.K., Domingo, E., Holland, J.J. and Moya, A. (1996) Evolution of fitness in experimental populations of vesicular stomatitis virus, *Genetics*, **142**, 673-679.

ElHefnawi, M.M., Zada, S. and El-Azab, I.A. (2010) Prediction of prognostic biomarkers for Interferon-based therapy to Hepatitis C Virus patients: a metaanalysis of the NS5A protein in subtypes 1a, 1b, and 3a, *Virol J*, **7**, 130.

- Enjuanes, L., Sola, I., Alonso, S., Escors, D. and Zúñiga, S. (2005) Coronavirus reverse genetics and development of vectors for gene expression, *Curr Top Microbiol Immunol*, **287**, 161-197.
- Enomoto, N., Sakuma, I., Asahina, Y., Kuroski, M., Murakami, T., Yamamoto, C., Ogura, Y., Izumi, N., Marumo, F. and Sato, C. (1996) Mutations in the nonstructural protein 5A gene and response to interferon in patients with chronic hepatitis C virus 1b infection, *N Engl J Med*, **334**, 77-81.
- Evans, D.J. (1999) Reverse genetics of picornaviruses, *Adv Virus Res*, **53**, 209-228.
- Fagin, R. (1977) Multivalued dependencies and a new normal form for relational databases, *ACM Transactions on Database Systems*, **2**, 262-278.
- Fagin, R. (1981) A normal form for relational databases that is based on domains and keys, *ACM Transactions on Database Systems*, **6**, 387-415.
- Farci, P., Alter, H.J., Shimoda, A., Govindarajan, S., Cheung, L.C., Melpolder, J.C., Sacher, R.A., Shih, J.W. and Purcell, R.H. (1996) Hepatitis C virus-associated fulminant hepatic failure, *N Engl J Med*, **335**, 631-634.
- Farci, P., Strazzera, R., Alter, H.J., Farci, S., Degioannis, D., Coiana, A., Peddis, G., Usai, F., Serra, G., Chessa, L., Diaz, G., Balestrieri, A. and Purcell, R.H. (2002) Early changes in hepatitis C viral quasispecies during interferon therapy predict the therapeutic outcome, *Proc Natl Acad Sci USA*, **99**, 3081-3086.
- Feld, J.J. and Hoofnagle, J.H. (2005) Mechanism of action of interferon and ribavirin in treatment of hepatitis C, *Nature*, **436**, 967-972.
- Fellenberg, M. (2003) Developing integrative bioinformatics systems, *BIOSILICO*, **1**, 177-183.
- Feng, D.F. and Doolittle, R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees, *J Mol Evol*, **25**, 351-360.

Ferguson, N.M., Galvani, A.P. and Bush, R.M. (2003) Ecological and immunological determinants of influenza evolution, *Nature*, **422**, 428-433.

Fitch, W., Brisse, S., Stevens, J. and Tibayrenc, M. (2001) Infectious diseases and the golden age of phylogenetics: an E-debate, *Infect Genet Evol*, **1**, 69-74.

Fowler, M. (2002) Information systems architecture, *Software Engineering, 2002. ICSE 2002. Proceedings of the 24rd International Conference on*, 702.

Foxman, B. and Riley, L. (2001) Molecular epidemiology: focus on infection, *Am J Epidemiol*, **153**, 1135-1141.

Fried, M.W., Schiffman, M.L., Reddy, K.R., Smith, C., Marinos, G., Gonçales, F.L., Häussinger, D., Diago, M., Carosi, G., Dhumeaux, D., Craxi, A., Lin, A., Hoffman, J. and Yu, J. (2002) Peginterferon alfa-2a plus ribavirin for chronic hepatitis C virus infection, *N Engl J Med*, **347**, 975-982.

Friedberg, E., Walker, G. and Siede, W. (1995) DNA repair and mutagenesis, *American Society for Microbiology. Washington, DC*.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent, *J Stat Softw*, **33**, 1-22.

Fu, Y.X. and Li, W.H. (1993) Statistical tests of neutrality of mutations, *Genetics*, **133**, 693-709.

Fukuda, M., Chayama, K., Tsubota, A., Kobayashi, M., Hashimoto, M., Miyano, Y., Koike, H., Kobayashi, M., Koida, I., Arase, Y., Saitoh, S., Murashima, N., Ikeda, K. and Kumada, H. (1998) Predictive factors in eradicating hepatitis C virus using a relatively small dose of interferon, *J Gastroenterol Hepatol*, **13**, 412-418.

Gale, M., Kwieciszewski, B., Dossett, M., Nakao, H. and Katze, M.G. (1999) Antiapoptotic and oncogenic potentials of hepatitis C virus are linked to interferon resistance by viral repression of the PKR protein kinase, *Journal of Virology*, **73**, 6506-6516.

- Galperin, M.Y. and Cochrane, G.R. (2011) The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection, *Nucleic Acids Research*, **39**, D1-6.
- Gao, B., Hong, F. and Radaeva, S. (2004) Host factors and failure of interferon-alpha treatment in hepatitis C virus, *Hepatology*, **39**, 880-890.
- Gao, R., Yu, J. and Zhang, M. (2010) Systems Theoretic Analysis of the Central Dogma of Molecular Biology: Some Recent Results, *IEEE transactions on nanobioscience*.
- Gaudy, C., Lambelé, M., Moreau, A., Veillon, P., Lunel, F. and Goudeau, A. (2005) Mutations within the hepatitis C virus genotype 1b E2-PePHD domain do not correlate with treatment outcome, *J Clin Microbiol*, **43**, 750-754.
- Gause, G.F. (1934) Experimental analysis of Vito Volterra's mathematical theory of the struggle for existence, *Science*, **79**, 16-17.
- Gerotto, M., Dal Pero, F., Pontisso, P., Noventa, F., Gatta, A. and Alberti, A. (2000) Two PKR inhibitor HCV proteins correlate with early but not sustained response to interferon, *Gastroenterology*, **119**, 1649-1655.
- Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S. and Snyder, M. (2007) What is a gene, post-ENCODE? History and updated definition, *Genome Res.*, **17**, 669-681.
- Ghany, M.G., Strader, D.B., Thomas, D.L., Seeff, L.B. and Diseases, A.A.f.t.S.o.L. (2009) Diagnosis, management, and treatment of hepatitis C: an update, *Hepatology*, **49**, 1335-1374.
- Gnaneshan, S., Ijaz, S., Moran, J., Ramsay, M. and Green, J. (2007) HepSEQ: International Public Health Repository for Hepatitis B, *Nucleic Acids Research*, **35**, D367-370.
- Golden-Mason, L. and Rosen, H.R. (2006) Natural killer cells: primary target for hepatitis C virus immune evasion strategies?, *Liver Transpl*, **12**, 363-372.

- González-Candelas, F. (2010) Molecular Phylogenetic Analyses in Court Trials, *Encyclopedia of Life Sciences (ELS)*, 6.
- González-Candelas, F. and López-Labrador, F. (2010) Clinical relevance of genetic heterogeneity in HCV, *Future Virology*.
- Gotoh, O. (1995) A weighting system and algorithm for aligning many phylogenetically related sequences, *Comput Appl Biosci*, **11**, 543-551.
- Guidotti, L.G. and Chisari, F.V. (2006) Immunobiology and pathogenesis of viral hepatitis, *Annual review of pathology*, **1**, 23-61.
- Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, *Systematic Biology* **52**, 696-704.
- Guo, H.-Z., Yin, Y., Wang, W.-L., Zhang, C.-S., Wang, T., Wang, Z., Zhang, J., Cheng, H. and Wang, H.-T. (2004) Sequence evolution of putative cytotoxic T cell epitopes in NS3 region of hepatitis C virus, *World J Gastroenterol*, **10**, 847-851.
- Gupta, R., Subramani, M., Khaja, M.N., Madhavi, C., Roy, S., Habibullah, C.M. and Das, S. (2006) Analysis of mutations within the 5' untranslated region, interferon sensitivity region, and PePHD region as a function of response to interferon therapy in hepatitis C virus-infected patients in India, *J Clin Microbiol*, **44**, 709-715.
- Guttmacher, A.E. (2001) Human genetics on the web, *Annu Rev Genomics Hum Genet*, **2**, 213-233.
- Haas, L., Schwarz, P., Kodali, P., Kotlar, E. and Rice, J. (2001) DiscoveryLink: A system for integrated access to life sciences data sources, *Ibm Systems Journal*.
- Hadziyannis, S.J., Sette, H., Morgan, T.R., Balan, V., Diago, M., Marcellin, P., Ramadori, G., Bodenheimer, H., Bernstein, D., Rizzetto, M., Zeuzem, S., Pockros, P.J., Lin, A., Ackrill, A.M. and Group, P.I.S. (2004) Peginterferon-alpha2a and ribavirin combination therapy in chronic hepatitis C: a randomized study of treatment duration and ribavirin dose, *Ann Intern Med*, **140**, 346-355.

Hall, B.G. and Barlow, M. (2006) Phylogenetic analysis as a tool in molecular epidemiology of infectious diseases, *Ann Epidemiol*, **16**, 157-169.

Hayashi, K., Katano, Y., Ishigami, M., Itoh, A., Hirooka, Y., Nakano, I., Urano, F., Yoshioka, K., Toyoda, H., Kumada, T. and Goto, H. (2010) Mutations in the core and NS5A region of hepatitis C virus genotype 1b and correlation with response to pegylated-interferon-alpha 2b and ribavirin combination therapy, *J Viral Hepat*.

Hayashi, N. and Takehara, T. (2006) Antiviral therapy for chronic hepatitis C: past, present, and future, *J Gastroenterol*, **41**, 17-27.

He, Y., Vines, R.R., Wattam, A.R., Abramochkin, G.V., Dickerman, A.W., Eckart, J.D. and Sobral, B.W.S. (2005) PIML: the Pathogen Information Markup Language, *Bioinformatics*, **21**, 116-121.

Hey, J. (1991) The structure of genealogies and the distribution of fixed differences between DNA sequence samples from natural populations, *Genetics*, **128**, 831-840.

Hill, K.R., Hajjou, M., Hu, J.Y. and Raju, R. (1997) RNA-RNA recombination in Sindbis virus: roles of the 3' conserved motif, poly(A) tail, and nonviral sequences of template RNAs in polymerase recognition and template switching, *Journal of Virology*, **71**, 2693-2704.

Hirahata, M., Abe, T., Tanaka, N., Kuwana, Y., Shigemoto, Y., Miyazaki, S., Suzuki, Y. and Sugawara, H. (2007) Genome Information Broker for Viruses (GIB-V): database for comparative analysis of virus genomes, *Nucleic Acids Research*, **35**, D339-342.

Hofgärtner, W.T., Polyak, S.J., Sullivan, D.G., Carithers, R.L. and Gretch, D.R. (1997) Mutations in the NS5A gene of hepatitis C virus in North American patients infected with HCV genotype 1a or 1b, *J. Med. Virol.*, **53**, 118-126.

Hofmann, W.P., Zeuzem, S. and Sarrazin, C. (2005) Hepatitis C virus-related resistance mechanisms to interferon alpha-based antiviral therapy, *J Clin Virol*, **32**, 86-91.

- Holmes, E.C. (1998) Molecular epidemiology and evolution of emerging infectious diseases, *Br Med Bull*, **54**, 533-543.
- Hoofnagle, J.H. (2002) Course and outcome of hepatitis C, *Hepatology*, **36**, S21-29.
- Huang, X. and Miller, W. (1991) A time-efficient linear-space local similarity algorithm, *Advances in Applied Mathematics*.
- Hudson, R.R. (1987) Estimating the recombination parameter of a finite population model without selection, *Genet Res*, **50**, 245-250.
- Hudson, R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation, *Bioinformatics*, **18**, 337-338.
- Hung, C.-H., Lee, C.-M., Lu, S.-N., Lee, J.-F., Wang, J.-H., Tung, H.-D., Chen, T.-M., Hu, T.-H., Chen, W.-J. and Changchien, C.-S. (2003) Mutations in the NS5A and E2-PePHD region of hepatitis C virus type 1b and correlation with the response to combination therapy with interferon and ribavirin, *J Viral Hepat*, **10**, 87-94.
- Ibarrola, N., Moreno-Monteagudo, J.A., Sáiz, M., García-Monzón, C., Sobrino, F., García-Buey, L., Lo Iacono, O., Moreno-Otero, R. and Martínez-Salas, E. (1999) Response to retreatment with interferon-alpha plus ribavirin in chronic hepatitis C patients is independent of the NS5A gene nucleotide sequence, *Am J Gastroenterol*, **94**, 2487-2495.
- Jackson, D., Healy, M. and Davison, D. (2003) Binformatics: not just for sequences anymore, *BIOSILICO*, **1**, 103-111.
- Jansen, A. and Yu, J. (2006) Differential gene expression of pathogens inside infected hosts, *Current Opinion in Microbiology*, **9**, 138-142.
- Jenke, A.C.W., Moser, S., Orth, V., Zilbauer, M., Gerner, P. and Wirth, S. (2009) Mutation frequency of NS5A in patients vertically infected with HCV genotype 1 predicts sustained virological response to peginterferon alfa-2b and ribavirin combination therapy, *J Viral Hepat*, **16**, 853-859.

Jensen, D.M., Morgan, T.R., Marcellin, P., Pockros, P.J., Reddy, K.R., Hadziyannis, S.J., Ferenci, P., Ackrill, A.M. and Willems, B. (2006) Early identification of HCV genotype 1 patients responding to 24 weeks peginterferon alpha-2a (40 kd)/ribavirin therapy, *Hepatology*, **43**, 954-960.

Jiménez-Hernández, N. (2004) Evolución del virus de la hepatitis C en muestras hospitalarias de la Comunidad Valenciana, *Tesis doctoral. Universidad de Valencia*.

Jiménez-Hernández, N., Sentandreu, V., Castro, J.A., Torres-Puente, M., Bracho, A., García-Robles, I., Ortega, E., del Olmo, J., Carnicer, F., González-Candelas, F. and Moya, A. (2008) Effect of antiviral treatment and host susceptibility on positive selection in hepatitis C virus (HCV), *Virus Res*, **131**, 224-232.

Jiménez-Hernández, N., Torres-Puente, M., Bracho, M.A., García-Robles, I., Ortega, E., del Olmo, J., Carnicer, F., González-Candelas, F. and Moya, A. (2007) Epidemic dynamics of two coexisting hepatitis C virus subtypes, *J Gen Virol*, **88**, 123-133.

Jukes, T. and Cantor, C. (1969) Evolution of protein molecules, *Mammalian protein metabolism*.

Kalinina, O., Norder, H., Mukomolov, S. and Magnius, L.O. (2002) A natural intergenotypic recombinant of hepatitis C virus identified in St. Petersburg, *Journal of Virology*, **76**, 4034-4043.

Kaminuma, E., Kosuge, T., Kodama, Y., Aono, H., Mashima, J., Gojobori, T., Sugawara, H., Ogasawara, O., Takagi, T., Okubo, K. and Nakamura, Y. (2011) DDBJ progress report, *Nucleic Acids Research*, **39**, D22-27.

Kaplowitz, N. (2004) Drug-induced liver injury, *Clin Infect Dis*, **38 Suppl 2**, S44-48.

Karasavvas, K.A., Baldock, R. and Burger, A. (2004) Bioinformatics integration and agent technology, *Journal of Biomedical Informatics*, **37**, 205-219.

- Kato, N., Ootsuyama, Y., Tanaka, T., Nakagawa, M., Nakazawa, T., Muraiso, K., Ohkoshi, S., Hijikata, M. and Shimotohno, K. (1992) Marked sequence diversity in the putative envelope proteins of hepatitis C viruses, *Virus Res*, **22**, 107-123.
- Katoh, K., Asimenos, G. and Toh, H. (2009) Multiple Alignment of DNA Sequences with MAFFT, *Methods Mol Biol*, **537**, 39-64.
- Katoh, K., Kuma, K.-i., Toh, H. and Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment, *Nucleic Acids Research*, **33**, 511-518.
- Katoh, K., Misawa, K., Kuma, K.-i. and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Research*, **30**, 3059-3066.
- Katoh, K. and Toh, H. (2007) PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences, *Bioinformatics*, **23**, 372-374.
- Katoh, K. and Toh, H. (2008) Recent developments in the MAFFT multiple sequence alignment program, *Briefings in Bioinformatics*, **9**, 286-298.
- Kau, A., Vermehren, J. and Sarrazin, C. (2008) Treatment predictors of a sustained virologic response in hepatitis B and C, *Journal of Hepatology*, **49**, 634-651.
- Khorsi, H., Castelain, S., Wyseur, A., Izopet, J., Canva, V., Rombout, A., Capron, D., Capron, J.P., Lunel, F., Stuyver, L. and Duverlie, G. (1997) Mutations of hepatitis C virus 1b NS5A 2209-2248 amino acid sequence do not predict the response to recombinant interferon-alfa therapy in French patients, *Journal of Hepatology*, **27**, 72-77.
- Kimura, M. (1983) The neutral theory of molecular evolution. *Cambridge University Press*, 384 pp.
- Knodell, R.G., Ishak, K.G., Black, W.C., Chen, T.S., Craig, R., Kaplowitz, N., Kiernan, T.W. and Wollman, J. (1981) Formulation and application of a numerical scoring system for assessing

histological activity in asymptomatic chronic active hepatitis, *Hepatology*, **1**, 431-435.

Köhler, J. (2004) Integration of life science databases, *Drug Discovery Today: BIOSILICO*, **2**, 61-69.

Kuiken, C., Hraber, P., Thurmond, J. and Yusim, K. (2008) The hepatitis C sequence database in Los Alamos, *Nucleic Acids Research*, **36**, D512-516.

Kuiken, C., Mizokami, M., Deleage, G., Yusim, K., Penin, F., Shin-I, T., Charavay, C., Tao, N., Crisan, D., Grando, D., Dalwani, A., Geourjon, C., Agrawal, A. and Combet, C. (2006) Hepatitis C databases, principles and utility to researchers, *Hepatology*, **43**, 1157-1165.

Kulkarni-Kale, U., Bhosle, S., Manjari, G.S. and Kolaskar, A.S. (2004) VirGen: a comprehensive viral genome resource, *Nucleic Acids Research*, **32**, D289-292.

Kumar, S., Tamura, K. and Nei, M. (2004) MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment, *Briefings in Bioinformatics*, **5**, 150-163.

Kumar, S. and Dudley, J. (2007a) Bioinformatics software for biologists in the genomics era, *Bioinformatics*, **23**, 1713-1717.

Kumar, S. and Filipski, A. (2007b) Multiple sequence alignment: in pursuit of homologous DNA positions, *Genome Res.*, **17**, 127-135.

Kumar, S., Nei, M., Dudley, J. and Tamura, K. (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences, *Briefings in Bioinformatics*, **9**, 299-306.

Kumthip, K., Pantip, C., Chusri, P., Thongsawat, S., O'Brien, A., Nelson, K.E. and Maneekarn, N. (2010) Correlation between mutations in the core and NS5A genes of hepatitis C virus genotypes 1a, 1b, 3a, 3b, 6f and the response to pegylated interferon and ribavirin combination therapy, *J Viral Hepat.*

Kurosaki, M., Enomoto, N., Murakami, T., Sakuma, I., Asahina, Y., Yamamoto, C., Ikeda, T., Tozuka, S., Izumi, N., Marumo, F. and Sato, C. (1997) Analysis of genotypes and amino acid residues 2209 to 2248 of the NS5A region of hepatitis C virus in relation to the response to interferon-beta therapy, *Hepatology*, **25**, 750-753.

Kurosaki, M., Matsunaga, K., Hirayama, I., Tanaka, T., Sato, M., Yasui, Y., Tamaki, N., Hosokawa, T., Ueda, K., Tsuchiya, K., Nakanishi, H., Ikeda, H., Itakura, J., Takahashi, Y., Asahina, Y., Higaki, M., Enomoto, N. and Izumi, N. (2010) A predictive model of response to peginterferon ribavirin in chronic hepatitis C using classification and regression tree analysis, *Hepatol Res*, **40**, 251-260.

Lacroix, Z. (2002) Biological data integration: wrapping data and tools, *Information Technology in Biomedicine, IEEE Transactions on*, **6**, 123 - 128.

Lai, M.M. (1992) RNA recombination in animal and plant viruses, *Microbiol Rev*, **56**, 61-79.

Lappalainen, M., Chen, R.W., Maunula, L., von Bonsdorff, C., Plyusnin, A. and Vaheri, A. (2001) Molecular epidemiology of viral pathogens and tracing of transmission routes: hepatitis-, calici- and hantaviruses, *J Clin Virol*, **21**, 177-185.

Lauritsen, J.M. (2000) EpiData Data Entry, Data Management and basic Statistical Analysis System, *Odense Denmark, EpiData Association*.

Layden-Almer, J.E., Kuiken, C., Ribeiro, R.M., Kunstman, K.J., Perelson, A.S., Layden, T.J. and Wolinsky, S.M. (2005) Hepatitis C virus genotype 1a NS5A pretreatment sequence variation and viral kinetics in African American and white patients, *J Infect Dis*, **192**, 1078-1087.

Lee, C.-M., Hung, C.-H., Lu, S.-N. and Changchien, C.-S. (2008a) Hepatitis C virus genotypes: clinical relevance and therapeutic implications, *Chang Gung Med J*, **31**, 16-25.

Lee, S.S. (2003) Review article: indicators and predictors of response to anti-viral therapy in chronic hepatitis C, *Aliment Pharmacol Ther*, **17**, 611-621.

Lee, S.S. and Ferenci, P. (2008b) Optimizing outcomes in patients with hepatitis C virus genotype 1 or 4, *Antivir Ther (Lond)*, **13 Suppl 1**, 9-16.

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., Zalunin, V. and Cochrane, G. (2011) The European Nucleotide Archive, *Nucleic Acids Research*, **39**, D28-31.

Librado, P. and Rozas, J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data, *Bioinformatics*, **25**, 1451-1452.

Lindsay, K.L., Trepo, C., Heintges, T., Shiffman, M.L., Gordon, S.C., Hoefs, J.C., Schiff, E.R., Goodman, Z.D., Laughlin, M., Yao, R., Albrecht, J.K. (2001) A randomized, double-blind trial comparing pegylated interferon alfa-2b to interferon alfa-2b as initial treatment for chronic hepatitis C, *Hepatology*, **34**, 395-403.

Lipman, D.J. and Pearson, W.R. (1985) Rapid and sensitive protein similarity searches, *Science*, **227**, 1435-1441.

Liu, D., Wang, X., Pan, F., Xu, Y., Yang, P. and Rao, K. (2008) Web-based infectious disease reporting using XML forms, *International Journal of Medical Informatics*, **77**, 630-640.

Lo, S. and Lin, H.H. (2001) Variations within hepatitis C virus E2 protein and response to interferon treatment, *Virus Res*, **75**, 107-112.

López-Hernández, J. (1990) La gestión de la información en las organizaciones: una disciplina emergente, *Documentación de las Ciencias de la Información*.

López-Labrador, F.X., He, X.-S., Berenguer, M., Cheung, R.C., González-Candelas, F., Wright, T.L. and Greenberg, H.B. (2004)

Genetic variability of hepatitis C virus non-structural protein 3 and virus-specific CD8+ response in patients with chronic hepatitis C, *J. Med. Virol.*, **72**, 575-585.

Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A. and Tarczy-Hornoch, P. (2007) Data integration and genomic medicine, *Journal of Biomedical Informatics*, **40**, 5-16.

Lu, L., Nakano, T., Orito, E., Mizokami, M. and Robertson, B.H. (2001) Evaluation of accumulation of hepatitis C virus mutations in a chronically infected chimpanzee: comparison of the core, E1, HVR1, and NS5b regions, *Journal of Virology*, **75**, 3004-3009.

Lukasiewicz, E., Gorfine, M., Freedman, L.S., Pawlotsky, J.-M., Schalm, S.W., Ferrari, C., Zeuzem, S., Neumann, A.U. and Group, D.-H.S. (2010) Prediction of nonSVR to therapy with pegylated interferon-alpha2a and ribavirin in chronic hepatitis C genotype 1 patients after 4, 8 and 12 weeks of treatment, *J Viral Hepat*, **17**, 345-351.

Ma, J., Otten, M., Kamadjeu, R., Mir, R., Rosencrans, L., McLaughlin, S. and Yoon, S. (2008) New frontiers for health information systems using Epi Info in developing countries: structured application framework for Epi Info (SAFE), *International Journal of Medical Informatics*, **77**, 219-225.

Maekawa, S. and Enomoto, N. (2009) Viral factors influencing the response to the combination therapy of peginterferon plus ribavirin in chronic hepatitis C, *J Gastroenterol*, **44**, 1009-1015.

Maier, D. (1983) The theory of relational databases, *Computer Science Press, Rockville*, 637.

Makino, S., Keck, J.G., Stohlman, S.A. and Lai, M.M. (1986) High-frequency RNA recombination of murine coronaviruses, *Journal of Virology*, **57**, 729-737.

Mallat, A., Hezode, C. and Lotersztajn, S. (2008) Environmental factors as disease accelerators during chronic hepatitis C, *Journal of Hepatology*, **48**, 657-665.

- Mallet, V., Vallet-Pichard, A. and Pol, S. (2010) New trends in hepatitis C management, *Presse Med*, **39**, 446-451.
- Mangia, A., Minerva, N., Bacca, D., Cozzolongo, R., Ricci, G.L., Carretta, V., Vinelli, F., Scotto, G., Montalto, G., Romano, M., Cristofaro, G., Mottola, L., Spirito, F. and Andriulli, A. (2008) Individualized treatment duration for hepatitis C genotype 1 patients: A randomized controlled trial, *Hepatology*, **47**, 43-50.
- Manns, M.P., McHutchison, J.G., Gordon, S.C., Rustgi, V.K., Schiffman, M., Reindollar, R., Goodman, Z.D., Koury, K., Ling, M. and Albrecht, J.K. (2001) Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of chronic hepatitis C: a randomised trial, *Lancet*, **358**, 958-965.
- Manns, M.P., Wedemeyer, H. and Cornberg, M. (2006) Treating viral hepatitis C: efficacy, side effects, and complications, *Gut*, **55**, 1350-1359.
- Maojo, V. and Martin-Sánchez, F. (2004) Bioinformatics: towards new directions for public health, *Methods Inf Med*, **43**, 208-214.
- Martínez-Bauer, E., Crespo, J., Romero-Gómez, M., Moreno-Otero, R., Solá, R., Tesei, N., Pons, F., Forns, X. and Sánchez-Tapias, J.M. (2006) Development and validation of two models for early prediction of response to therapy in genotype 1 chronic hepatitis C, *Hepatology*, **43**, 72-80.
- Martinot-Peignoux, M., Comanor, L., Minor, J.M., Ripault, M.P., Pham, B.-N., Boyer, N., Castelnau, C., Giuly, N., Hendricks, D. and Marcellin, P. (2006) Accurate model predicting sustained response at week 4 of therapy with pegylated interferon with ribavirin in patients with chronic hepatitis C, *J Viral Hepat*, **13**, 701-707.
- McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in *Drosophila*, *Nature*, **351**, 652-654.
- McNeil, L.K., Reich, C., Aziz, R.K., Bartels, D., Cohoon, M., Disz, T., Edwards, R.A., Gerdes, S., Hwang, K., Kubal, M., Margaryan, G.R., Meyer, F., Mihalo, W., Olsen, G.J., Olson, R., Osterman, A., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Rodionov, D.A.,

Shi, X., Vassieva, O., Vonstein, V., Zagnitko, O., Xia, F., Zinner, J., Overbeek, R. and Stevens, R. (2007) The National Microbial Pathogen Database Resource (NMPDR): a genomics platform based on subsystem annotation, *Nucleic Acids Research*, **35**, D347-353.

Metzker, M.L. (2010) Sequencing technologies - the next generation, *Nat Rev Genet*, **11**, 31-46.

Moreno, M.P., Casane, D., López, L. and Cristina, J. (2006) Evidence of recombination in quasispecies populations of a Hepatitis C Virus patient undergoing anti-viral therapy, *Virol J*, **3**, 87.

Moreno, P., Alvarez, M., López, L., Moratorio, G., Casane, D., Castells, M., Castro, S., Cristina, J. and Colina, R. (2009) Evidence of recombination in Hepatitis C Virus populations infecting a hemophiliac patient, *Virol J*, **6**, 203.

Morgan, U., Ochman, H., Renaud, F. and Tibayrenc, M. (2001) Population genetics and population biology: what did they bring to the epidemiology of transmissible diseases? An e-debate, *Infect Genet Evol*, **1**, 161-166.

Mori, N., Imamura, M., Kawakami, Y., Saneto, H., Kawaoka, T., Takaki, S., Aikata, H., Takahashi, S., Chayama, K. and Group, H.L.S. (2009) Randomized trial of high-dose interferon-alpha-2b combined with ribavirin in patients with chronic hepatitis C: Correlation between amino acid substitutions in the core/NS5A region and virological response to interferon therapy, *J. Med. Virol.*, **81**, 640-649.

Morris, J.A., Gayther, S.A., Jacobs, I.J. and Jones, C. (2008) A Perl toolkit for LIMS development, *Source Code Biol Med*, **3**, 4.

Morris, P.J. (2005) Relational database design and implementation for biodiversity informatics, *Phyloinformatics*, **7**, 1-66.

Moya, A., Elena, S.F., Bracho, A., Miralles, R. and Barrio, E. (2000) The evolution of RNA viruses: A population genetics view, *Proc.Natl.Acad.Sci.U.S.A*, **97**, 6967-6973.

- Moya, A., Holmes, E.C. and González-Candelas, F. (2004) The population genetics and evolutionary epidemiology of RNA viruses, *Nat Rev Microbiol*, **2**, 279-288.
- Mullan, L.J. and Bleasby, A.J. (2002) Short EMBOSS User Guide. European Molecular Biology Open Software Suite, *Briefings in Bioinformatics*, **3**, 92-94.
- Muller, H. (1964) The relation of recombination to mutational advance, *Mutat Res*, **106**, 2-9.
- Muñoz de Rueda, P., Casado, J., Patón, R., Quintero, D., Palacios, A., Gila, A., Quiles, R., León, J., Ruiz-Extremera, A. and Salmerón, J. (2008) Mutations in E2-PePHD, NS5A-PKRBD, NS5A-ISDR, and NS5A-V3 of hepatitis C virus genotype 1 and their relationships to pegylated interferon-ribavirin treatment responses, *Journal of Virology*, **82**, 6644-6653.
- Murashima, S., Ide, T., Miyajima, I., Kumashiro, R., Ueno, T., Sakisaka, S. and Sata, M. (1999) Mutations in the NS5A gene predict response to interferon therapy in Japanese patients with chronic hepatitis C and cirrhosis, *Scand J Infect Dis*, **31**, 27-32.
- Murayama, M., Katano, Y., Nakano, I., Ishigami, M., Hayashi, K., Honda, T., Hirooka, Y., Itoh, A. and Goto, H. (2007) A mutation in the interferon sensitivity-determining region is associated with responsiveness to interferon-ribavirin combination therapy in chronic hepatitis patients infected with a Japan-specific subtype of hepatitis C virus genotype 1B, *J. Med. Virol.*, **79**, 35-40.
- Murphy, M.D., Rosen, H.R., Marousek, G.I. and Chou, S. (2002) Analysis of sequence configurations of the ISDR, PKR-binding domain, and V3 region as predictors of response to induction interferon-alpha and ribavirin therapy in chronic hepatitis C infection, *Dig Dis Sci*, **47**, 1195-1205.
- Nájera, R., Delgado, E., Pérez-Alvarez, L. and Thomson, M.M. (2002) Genetic recombination and its role in the development of the HIV-1 pandemic, *AIDS*, **16 Suppl 4**, S3-16.

Nakamura, S., Yang, C.-S., Sakon, N., Ueda, M., Tougan, T., Yamashita, A., Goto, N., Takahashi, K., Yasunaga, T., Ikuta, K., Mizutani, T., Okamoto, Y., Tagami, M., Morita, R., Maeda, N., Kawai, J., Hayashizaki, Y., Nagai, Y., Horii, T., Iida, T. and Nakaya, T. (2009) Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach, *PLoS ONE*, **4**, e4219.

Nakano, I., Fukuda, Y., Katano, Y., Nakano, S., Kumada, T. and Hayakawa, T. (1999) Why is the interferon sensitivity-determining region (ISDR) system useful in Japan?, *Journal of Hepatology*, **30**, 1014-1022.

Nei, M. (1987) Molecular Evolutionary Genetics. *Columbia University Press, New York*, 512 pp.

Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions, *Mol Biol Evol*, **3**, 418-426.

Nei, M. and Tajima, F. (1981) DNA polymorphism detectable by restriction endonucleases, *Genetics*, **97**, 145-163.

Nelson, M., Reisinger, S. and Henry, S. (2003) Designing databases to store biological information, *BIOSILICO*, **1**, 134-142.

Neumann-Haefelin, C., Blum, H.E., Chisari, F.V. and Thimme, R. (2005) T cell response in hepatitis C virus infection, *J Clin Virol*, **32**, 75-85.

Nielsen, R. (2005) Molecular signatures of natural selection, *Annu Rev Genet*, **39**, 197-218.

Nordborg, M. and Innan, H. (2002) Molecular population genetics, *Curr Opin Plant Biol*, **5**, 69-73.

Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment, *J Mol Biol*, **302**, 205-217.

- Notredame, C., Holm, L. and Higgins, D.G. (1998) COFFEE: an objective function for multiple sequence alignments, *Bioinformatics*, **14**, 407-422.
- Nousbaum, J., Polyak, S.J., Ray, S.C., Sullivan, D.G., Larson, A.M., Carithers, R.L. and Gretch, D.R. (2000) Prospective characterization of full-length hepatitis C virus NS5A quasispecies during induction and combination antiviral therapy, *Journal of Virology*, **74**, 9028-9038.
- Novella, I.S. and Ebendick-Corpus, B.E. (2004) Molecular basis of fitness loss and fitness recovery in vesicular stomatitis virus, *J Mol Biol*, **342**, 1423-1430.
- Nuismer, S.L. and Otto, S.P. (2005) Host-parasite interactions and the evolution of gene expression, *PLoS Biol*, **3**, e203.
- Odeberg, J., Yun, Z., Sönnnerborg, A., Weiland, O. and Lundeberg, J. (1998) Variation in the hepatitis C virus NS5a region in relation to hypervariable region 1 heterogeneity during interferon treatment, *J. Med. Virol.*, **56**, 33-38.
- Okanoue, T., Itoh, Y., Hashimoto, H., Yasui, K., Minami, M., Takehara, T., Tanaka, E., Onji, M., Toyota, J., Chayama, K., Yoshioka, K., Izumi, N., Akuta, N. and Kumada, H. (2009) Predictive values of amino acid sequences of the core and NS5A regions in antiviral therapy for hepatitis C: a Japanese multi-center study, *J Gastroenterol*, **44**, 952-963.
- Olson, S.A. (2002) EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite, *Briefings in Bioinformatics*, **3**, 87-91.
- Pang, P.S., Planet, P.J. and Glenn, J.S. (2009) The evolution of the major hepatitis C genotypes correlates with clinical response to interferon therapy, *PLoS ONE*, **4**, e6579.
- Pascu, M., Martus, P., Höhne, M., Wiedenmann, B., Hopf, U., Schreier, E. and Berg, T. (2004) Sustained virological response in hepatitis C virus type 1b infected patients is predicted by the number of mutations within the NS5A-ISDR: a meta-analysis focused on geographical differences, *Gut*, **53**, 1345-1351.

- Patel, K., Muir, A.J. and McHutchison, J.G. (2006) Diagnosis and treatment of chronic hepatitis C infection, *BMJ*, **332**, 1013-1017.
- Pawlotsky, J.M. (2006) Hepatitis C virus population dynamics during infection, *Curr Top Microbiol Immunol*, **299**, 261-284.
- Pearlman, B.L., Ehleben, C. and Saifee, S. (2007) Treatment extension to 72 weeks of peginterferon and ribavirin in hepatitis c genotype 1-infected slow responders, *Hepatology*, **46**, 1688-1694.
- Penin, F., Dubuisson, J., Rey, F.A., Moradpour, D. and Pawlotsky, J.-M. (2004) Structural biology of hepatitis C virus, *Hepatology*, **39**, 5-19.
- Polyak, S.J., Nousbaum, J.B., Larson, A.M., Cotler, S., Carithers, R.L. and Gretch, D.R. (2000) The protein kinase-interacting domain in the hepatitis C virus envelope glycoprotein-2 gene is highly conserved in genotype 1-infected patients treated with interferon, *J Infect Dis*, **182**, 397-404.
- Porta, M. (2008) A dictionary of epidemiology. *Oxford University Press, USA; 5 edition*, 320 pp.
- Prati, D., Schiffman, M.L., Diago, M., Gane, E., Rajender Reddy, K., Pockros, P., Farci, P., O'Brien, C.B., Lardelli, P., Blotner, S. and Zeuzem, S. (2006) Viral and metabolic factors influencing alanine aminotransferase activity in patients with chronic hepatitis C, *Journal of Hepatology*, **44**, 679-685.
- Preston, R.J. (2003) Molecular epidemiology: potential impacts on the assessment of public health, *Mutat Res*, **543**, 121-124.
- Prljic, J., Veljkovic, N. and Veljkovic, V. (2004) Recombination property of the HIV-1 gp120 gene, *Int Rev Immunol*, **23**, 447-454.
- Puig-Basagoiti, F., Forns, X., Furcić, I., Ampurdanés, S., Giménez-Barcons, M., Franco, S., Sánchez-Tapias, J.M. and Saiz, J.-C. (2005) Dynamics of hepatitis C virus NS5A quasispecies during interferon and ribavirin therapy in responder and non-responder patients with genotype 1b chronic hepatitis C, *J Gen Virol*, **86**, 1067-1075.

Puig-Basagoiti, F., Sáiz, J.C., Forns, X., Ampurdanès, S., Giménez-Barcons, M., Franco, S., Sánchez-Fueyo, A., Costa, J., Sánchez-Tapias, J.M. and Rodés, J. (2001) Influence of the genetic heterogeneity of the ISDR and PePHD regions of hepatitis C virus on the response to interferon therapy in chronic hepatitis C, *J. Med. Virol.*, **65**, 35-44.

Quer, J., Huerta, R., Novella, I.S., Tsimring, L., Domingo, E. and Holland, J.J. (1996) Reproducible nonlinear population dynamics and critical points during replicative competitions of RNA virus quasispecies, *J Mol Biol*, **264**, 465-471.

R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.

Reed, K.E. and Rice, C.M. (2000) Overview of hepatitis C virus genome structure, polyprotein processing, and protein properties, *Curr Top Microbiol Immunol*, **242**, 55-84.

Rhodes, T., Wargo, H. and Hu, W.-S. (2003) High rates of human immunodeficiency virus type 1 recombination: near-random segregation of markers one kilobase apart in one round of viral replication, *Journal of Virology*, **77**, 11193-11200.

Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite, *Trends Genet*, **16**, 276-277.

Rispeter, K., Lu, M., Zibert, A., Wiese, M., de Oliveira, J.M. and Roggendorf, M. (1998) The "interferon sensitivity determining region" of hepatitis C virus is a stable sequence element, *Journal of Hepatology*, **29**, 352-361.

Robertson, B., Myers, G., Howard, C., Brettin, T., Bukh, J., Gaschen, B., Gojobori, T., Maertens, G., Mizokami, M., Nainan, O., Netesov, S., Nishioka, K., Shin i, T., Simmonds, P., Smith, D., Stuyver, L. and Weiner, A. (1998) Classification, nomenclature, and database development for hepatitis C virus (HCV) and related viruses: proposals for standardization. International Committee on Virus Taxonomy, *Arch Virol*, **143**, 2493-2503.

- Romano, P. (2008) Automation of in-silico data analysis processes through workflow management systems, *Briefings in Bioinformatics*, **9**, 57-68.
- Rozas, J. (2009) DNA sequence polymorphism analysis using DnaSP, *Methods Mol Biol*, **537**, 337-350.
- Rozas, J., Sánchez-DelBarrio, J., Messeguer, X. and Rozas, R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods, *Bioinformatics*, **19**, 2496-2497.
- Saito, H., Ebinuma, H., Ojiro, K., Wakabayashi, K., Inoue, M., Tada, S. and Hibi, T. (2010) On-treatment predictions of success in peg-interferon/ribavirin treatment using a novel formula, *World J Gastroenterol*, **16**, 89-97.
- Saito, T., Ito, T., Ishiko, H., Yonaha, M., Morikawa, K., Miyokawa, A. and Mitamura, K. (2003) Sequence analysis of PePHD within HCV E2 region and correlation with resistance of interferon therapy in Japanese patients infected with HCV genotypes 2a and 2b, *Am J Gastroenterol*, **98**, 1377-1383.
- Sáiz, J.C., López-Labrador, F.X., Ampurdanés, S., Dopazo, J., Forns, X., Sánchez-Tapias, J.M. and Rodés, J. (1998) The prognostic relevance of the nonstructural 5A gene interferon sensitivity determining region is different in infections with genotype 1b and 3a isolates of hepatitis C virus, *J Infect Dis*, **177**, 839-847.
- Saludes, V., Bracho, M., Valero, O. , Ardèvol M., Planas, R., González-Candelas, F., Ausina, V., and Martró, E. (2010) Baseline Prediction of Combination Therapy Outcome in Hepatitis C Virus 1b Infected Patients by Discriminant Analysis Using Viral and Host Factors, *PLoS ONE* 5(11):e14132.
- Sánchez-Tapias, J.M., Diago, M., Escartín, P., Enríquez, J., Romero-Gómez, M., Bárcena, R., Crespo, J., Andrade, R., Martínez-Bauer, E., Pérez, R., Testillano, M., Planas, R., Solá, R., García-Bengoechea, M., García-Samaniego, J., Muñoz-Sánchez, M., Moreno-Otero, R. and Group, T.-S. (2006) Peginterferon-alfa2a plus ribavirin for 48 versus 72 weeks in patients with detectable hepatitis C virus RNA at week 4 of treatment, *Gastroenterology*, **131**, 451-460.

Sansonno, D. and Dammacco, F. (2005) Hepatitis C virus, cryoglobulinaemia, and vasculitis: immune complex relations, *The Lancet Infectious Diseases*, **5**, 227-236.

Sarrazin, C., Berg, T., Lee, J.H., Rüster, B., Kronenberger, B., Roth, W.K. and Zeuzem, S. (2000a) Mutations in the protein kinase-binding domain of the NS5A protein in patients infected with hepatitis C virus type 1a are associated with treatment response, *J Infect Dis*, **181**, 432-441.

Sarrazin, C., Berg, T., Lee, J.H., Teuber, G., Dietrich, C.F., Roth, W.K. and Zeuzem, S. (1999) Improved correlation between multiple mutations within the NS5A region and virological response in European patients chronically infected with hepatitis C virus type 1b undergoing combination therapy, *Journal of Hepatology*, **30**, 1004-1013.

Sarrazin, C., Bruckner, M., Herrmann, E., Rüster, B., Bruch, K., Roth, W.K. and Zeuzem, S. (2001) Quasispecies heterogeneity of the carboxy-terminal part of the E2 gene including the PePHD and sensitivity of hepatitis C virus 1b isolates to antiviral therapy, *Virology*, **289**, 150-163.

Sarrazin, C., Herrmann, E., Bruch, K. and Zeuzem, S. (2002) Hepatitis C virus nonstructural 5A protein and interferon resistance: a new model for testing the reliability of mutational analyses, *Journal of Virology*, **76**, 11079-11090.

Sarrazin, C., Kornetzky, I., Rüster, B., Lee, J.H., Kronenberger, B., Bruch, K., Roth, W.K. and Zeuzem, S. (2000b) Mutations within the E2 and NS5A protein in patients infected with hepatitis C virus type 3a and correlation with treatment response, *Hepatology*, **31**, 1360-1370.

Schinkel, J., Spaan, W.J.M. and Kroes, A.C.M. (2004) Meta-analysis of mutations in the NS5A gene and hepatitis C virus resistance to interferon therapy: uniting discordant conclusions, *Antivir Ther (Lond)*, **9**, 275-286.

Schrimal, L.M., Arze, C., Nadendla, S., Ganapathy, A., Felix, V., Mahurkar, A., Phillippy, K., Gussman, A., Angiuoli, S., Ghedin, E.,

White, O. and Hall, N. (2010) GeMInA, Genomic Metadata for Infectious Agents, a geospatial surveillance pathogen database, *Nucleic Acids Research*, **38**, D754-764.

Searls, D.B. (2000) Bioinformatics tools for whole genomes, *Annu Rev Genomics Hum Genet*, **1**, 251-279.

Sentandreu, V., Jiménez-Hernández, N., Torres-Puente, M., Bracho, M.A., Valero, A., Gosalbes, M.J., Ortega, E., Moya, A. and González-Candelas, F. (2008) Evidence of recombination in intrapatient populations of hepatitis C virus, *PLoS ONE*, **3**, e3239.

Shapiro, J.A. (2009) Revisiting the central dogma in the 21st century, *Ann NY Acad Sci*, **1178**, 6-28.

Shepard, C.W., Finelli, L. and Alter, M.J. (2005) Global epidemiology of hepatitis C virus infection, *The Lancet Infectious Diseases*, **5**, 558-567.

Shepherd, J., Brodin, H.F.T., Cave, C.B., Waugh, N.R., Price, A. and Gabbay, J. (2005) Clinical- and cost-effectiveness of pegylated interferon alfa in the treatment of chronic hepatitis C: a systematic review and economic evaluation, *Int J Technol Assess Health Care*, **21**, 47-54.

Sheridan, I., Pybus, O.G., Holmes, E.C. and Klenerman, P. (2004) High-resolution phylogenetic analysis of hepatitis C virus adaptation and its relationship to disease progression, *Journal of Virology*, **78**, 3447-3454.

Shin-I, T., Tanaka, Y., Tateno, Y. and Mizokami, M. (2008) Development and public release of a comprehensive hepatitis virus database, *Hepatol Res*, **38**, 234-243.

Shirakawa, H., Matsumoto, A., Joshita, S., Komatsu, M., Tanaka, N., Umemura, T., Ichijo, T., Yoshizawa, K., Kiyosawa, K., Tanaka, E. and Group, N.I.T.R. (2008) Pretreatment prediction of virological response to peginterferon plus ribavirin therapy in chronic hepatitis C patients using viral and host factors, *Hepatology*, **48**, 1753-1760.

Silberschatz, A., F. Korth, H. and Sudarshan, S. (2002) Fundamentos de bases de datos. *McGraw-Hill*, 797 pp.

Simmonds, P., Bukh, J., Combet, C., Deléage, G., Enomoto, N., Feinstone, S., Halfon, P., Inchauspé, G., Kuiken, C., Maertens, G., Mizokami, M., Murphy, D.G., Okamoto, H., Pawlotsky, J.-M., Penin, F., Sablon, E., Shin-I, T., Stuyver, L.J., Thiel, H.-J., Viazov, S., Weiner, A.J. and Widell, A. (2005) Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes, *Hepatology*, **42**, 962-973.

Sintchenko, V. (2009) Infectious Disease Informatics. *Springer*, 434 pp.

Sintchenko, V. and Gallego, B. (2009a) Laboratory-guided detection of disease outbreaks: three generations of surveillance systems, *Arch Pathol Lab Med*, **133**, 916-925.

Sintchenko, V., Gallego, B., Chung, G. and Coiera, E. (2009b) Towards bioinformatics assisted infectious disease control, *BMC Bioinformatics*, **10 Suppl 2**, S10.

Sklan, E.H., Charuworn, P., Pang, P.S. and Glenn, J.S. (2009) Mechanisms of HCV survival in the host, *Nat Rev Gastroenterol Hepatol*, **6**, 217-227.

Snyder, E.E., Kampanya, N., Lu, J., Nordberg, E.K., Karur, H.R., Shukla, M., Soneja, J., Tian, Y., Xue, T., Yoo, H., Zhang, F., Dharmanolla, C., Dongre, N.V., Gillespie, J.J., Hamelius, J., Hance, M., Huntington, K.I., Jukneliene, D., Koziski, J., Mackasmie, L., Mane, S.P., Nguyen, V., Purkayastha, A., Shallom, J., Yu, G., Guo, Y., Gabbard, J., Hix, D., Azad, A.F., Baker, S.C., Boyle, S.M., Khudyakov, Y., Meng, X.J., Rupprecht, C., Vinje, J., Crasta, O.R., Czar, M.J., Dickerman, A., Eckart, J.D., Kenyon, R., Will, R., Setubal, J.C. and Sobral, B.W.S. (2007) PATRIC: the VBI PathoSystems Resource Integration Center, *Nucleic Acids Research*, **35**, D401-406.

Sorbi, D., Boynton, J. and Lindor, K.D. (1999) The ratio of aspartate aminotransferase to alanine aminotransferase: potential value in differentiating nonalcoholic steatohepatitis from alcoholic liver disease, *Am J Gastroenterol*, **94**, 1018-1022.

Sorek, R. and Cossart, P. (2010) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity, *Nat Rev Genet*, **11**, 9-16.

Squadrito, G., Orlando, M.E., Cacciola, I., Rumi, M.G., Artini, M., Picciotto, A., Loiacono, O., Siciliano, R., Levrero, M. and Raimondo, G. (1999) Long-term response to interferon alpha is unrelated to "interferon sensitivity determining region" variability in patients with chronic hepatitis C virus-1b infection, *Journal of Hepatology*, **30**, 1023-1027.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G.R., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D. and Birney, E. (2002) The Bioperl toolkit: Perl modules for the life sciences, *Genome Res.*, **12**, 1611-1618.

Stein, L.D. (2003) Integrating biological databases, *Nat Rev Genet*, **4**, 337-345.

Steinhauer, D.A., Domingo, E. and Holland, J.J. (1992) Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase, *Gene*, **122**, 281-288.

Steinmann, E., Penin, F., Kallis, S., Patel, A.H., Bartenschlager, R. and Pietschmann, T. (2007) Hepatitis C virus p7 protein is crucial for assembly and release of infectious virions, *PLoS Pathog*, **3**, e103.

Stormo, G.D. (2009) An introduction to sequence similarity ("homology") searching, *Current protocols in bioinformatics / editorial board, Andreas D Baxevanis [et al]*, Chapter 3, Unit 3.1.1-7.

Stratidaki, I., Skoulika, E., Kelefiotis, D., Matrella, E., Alexandrakis, G., Economou, A. and Kouroumalis, E. (2001) NS5A mutations predict biochemical but not virological response to interferon-alpha treatment of sporadic hepatitis C virus infection in European patients, *J Viral Hepat*, **8**, 243-248.

Sy, T. and Jamal, M.M. (2006) Epidemiology of hepatitis C virus (HCV) infection, *Int J Med Sci*, **3**, 41-46.

Szmaragd, C., Nichols, R.A. and Balloux, F. (2006) A novel approach to characterise pathogen candidate genetic polymorphisms involved in clinical outcome, *Infect Genet Evol*, **6**, 38-45.

Tajima, F. (1983) Evolutionary relationship of DNA sequences in finite populations, *Genetics*, **105**, 437-460.

Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism, *Genetics*, **123**, 585-595.

Tarasevich, I.V., Shaginyan, I.A. and Mediannikov, O.Y. (2003) Problems and perspectives of molecular epidemiology of infectious diseases, *Ann N Y Acad Sci*, **990**, 751-756.

Tárraga, J., Medina, I., Arbiza, L., Huerta-Cepas, J., Gabaldón, T., Dopazo, J. and Dopazo, H. (2007) Phylemon: a suite of web tools for molecular evolution, phylogenetics and phylogenomics, *Nucleic Acids Research*, **35**, W38-42.

Taylor, D.R., Shi, S.T., Romano, P.R., Barber, G.N. and Lai, M.M. (1999) Inhibition of the interferon-inducible protein kinase PKR by HCV E2 protein, *Science*, **285**, 107-110.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Research*, **22**, 4673-4680.

Thornton, K. (2003) Libsequence: a C++ class library for evolutionary genetic analysis, *Bioinformatics*, **19**, 2325-2327.

Tibayrenc, M. (2005) Bridging the gap between molecular epidemiologists and evolutionists, *Trends Microbiol*, **13**, 575-580.

Torres-Puente, M. (2004) Variabilidad genética y respuesta al tratamiento antiviral en el virus de la Hepatitis C (VHC), *Tesis doctoral. Universidad de Valencia*.

Torres-Puente, M., Cuevas, J.M., Jiménez-Hernández, N., Bracho, M.A., García-Robles, I., Carnicer, F., del Olmo, J., Ortega, E., Moya, A. and González-Candelas, F. (2008a) Hepatitis C virus and the

controversial role of the interferon sensitivity determining region in the response to interferon treatment, *J. Med. Virol.*, **80**, 247-253.

Torres-Puente, M., Cuevas, J.M., Jiménez-Hernández, N., Bracho, M.A., García-Robles, I., Wrobel, B., Carnicer, F., Del Olmo, J., Ortega, E., Moya, A. and González-Candelas, F. (2008b) Genetic variability in hepatitis C virus and its role in antiviral treatment response, *J Viral Hepat.*, **15**, 188-199.

Torres-Puente, M., Cuevas, J.M., Jiménez-Hernández, N., Bracho, M.A., García-Robles, I., Wrobel, B., Carnicer, F., del Olmo, J., Ortega, E., Moya, A. and González-Candelas, F. (2008c) Using evolutionary tools to refine the new hypervariable region 3 within the envelope 2 protein of hepatitis C virus, *Infect Genet Evol*, **8**, 74-82.

Trapero-Marugan, M., Marin, M., Pivel, J.P., Del Rio, J.M., Nunez, O., Clemente, G., Gisbert, J.P. and Moreno-Otero, R. (2008) Predictive graphical model, network-based medical tool for the prognosis of chronic hepatitis C patients treated with peg-interferon plus ribavirin, *Aliment Pharmacol Ther*, **28**, 468-474.

Troesch, M., Meunier, I., Lapierre, P., Lapointe, N., Alvarez, F., Boucher, M. and Soudeyns, H. (2006) Study of a novel hypervariable region in hepatitis C virus (HCV) E2 envelope glycoprotein, *Virology*, **352**, 357-367.

Ukai, K., Ishigami, M., Yoshioka, K., Kawabe, N., Katano, Y., Hayashi, K., Honda, T., Yano, M. and Goto, H. (2006) Mutations in carboxy-terminal part of E2 including PKR/eIF2alpha phosphorylation homology domain and interferon sensitivity determining region of nonstructural 5A of hepatitis C virus 1b: their correlation with response to interferon monotherapy and viral load, *World J Gastroenterol*, **12**, 3722-3728.

van Belkum, A., Struelens, M., de Visser, A., Verbrugh, H. and Tibayrenc, M. (2001) Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology, *Clin Microbiol Rev*, **14**, 547-560.

Van Rossum, T., Tripp, B. and Daley, D. (2010) SLIMS--a user-friendly sample operations and inventory management system for genotyping labs, *Bioinformatics*, **26**, 1808-1810.

Van Valen, L. (1973) A new evolutionary law, *Evolutionary theory*.

Veillon, P., Payan, C., Le Guillou-Guillemette, H., Gaudy, C. and Lunel, F. (2007) Quasispecies evolution in NS5A region of hepatitis C virus genotype 1b during interferon or combined interferon-ribavirin therapy, *World J Gastroenterol*, **13**, 1195-1203.

Viksna, J., Celms, E., Opmanis, M., Podnieks, K., Rucevskis, P., Zarins, A., Barrett, A., Neogi, S.G., Krestyaninova, M., McCarthy, M.I., Brazma, A. and Sarkans, U. (2007) PASSIM--an open source software system for managing information in biomedical studies, *BMC Bioinformatics*, **8**, 52.

Voegele, C., Tavtigian, S.V., De Silva, D., Cuber, S., Thomas, A. and Le Calvez-Kelm, F. (2007) A Laboratory Information Management System (LIMS) for a high throughput genetic platform aimed at candidate gene mutation screening, *Bioinformatics*, **23**, 2504-2506.

Wall, J. (1999) Recombination and the power of statistical tests of neutrality, *Genetics Research*.

Wallace, I.M., Blackshields, G. and Higgins, D.G. (2005) Multiple sequence alignments, *Curr Opin Struct Biol*, **15**, 261-266.

Watanabe, H., Enomoto, N., Nagayama, K., Izumi, N., Marumo, F., Sato, C. and Watanabe, M. (2001) Number and position of mutations in the interferon (IFN) sensitivity-determining region of the gene for nonstructural protein 5A correlate with IFN efficacy in hepatitis C virus genotype 1b infection, *J Infect Dis*, **183**, 1195-1203.

Watanabe, K., Yoshioka, K., Yano, M., Ishigami, M., Ukai, K., Ito, H., Miyata, F., Mizutani, T. and Goto, H. (2005) Mutations in the nonstructural region 5B of hepatitis C virus genotype 1b: their relation to viral load, response to interferon, and the nonstructural region 5A, *J. Med. Virol.*, **75**, 504-512.

- Watterson, G.A. (1975) On the number of segregating sites in genetical models without recombination, *Theoretical population biology*, **7**, 256-276.
- Weiner, A.J., Brauer, M.J., Rosenblatt, J., Richman, K.H., Tung, J., Crawford, K., Bonino, F., Saracco, G., Choo, Q.L. and Houghton, M. (1991) Variable and hypervariable domains are found in the regions of HCV corresponding to the flavivirus envelope and NS1 proteins and the pestivirus envelope glycoproteins, *Virology*, **180**, 842-848.
- Wendl, M.C., Smith, S., Pohl, C.S., Dooling, D.J., Chinwalla, A.T., Crouse, K., Hepler, T., Leong, S., Carmichael, L., Nhan, M., Oberkfell, B.J., Mardis, E.R., Hillier, L.W. and Wilson, R.K. (2007) Design and implementation of a generalized laboratory data model, *BMC Bioinformatics*, **8**, 362.
- Wernersson, R. (2003) RevTrans: multiple alignment of coding DNA from aligned amino acid sequences, *Nucleic Acids Research*, **31**, 3537-3539.
- Witherell, G.W. and Beineke, P. (2001) Statistical analysis of combined substitutions in nonstructural 5A region of hepatitis C virus and interferon response, *J. Med. Virol.*, **63**, 8-16.
- Wohnsland, A., Hofmann, W.P. and Sarrazin, C. (2007) Viral determinants of resistance to treatment in patients with hepatitis C, *Clin Microbiol Rev*, **20**, 23-38.
- Wong, J.B., Davis, G.L., McHutchison, J.G., Manns, M.P., Albrecht, J.K. and Group, I.H.I.T. (2003) Economic and clinical effects of evaluating rapid viral response to peginterferon alfa-2b plus ribavirin for the initial treatment of chronic hepatitis C, *Am J Gastroenterol*, **98**, 2354-2362.
- Wright, S. (1931) Evolution in Mendelian Populations, *Genetics*, **16**, 97-159.
- Xu, Z., Fan, X., Xu, Y. and Di Bisceglie, A.M. (2008) Comparative analysis of nearly full-length hepatitis C virus quasispecies from patients experiencing viral breakthrough during antiviral therapy:

clustered mutations in three functional genes, E2, NS2, and NS5a, *Journal of Virology*, **82**, 9417-9424.

Yang, I.S., Ryu, C., Cho, K.J., Kim, J.K., Ong, S.H., Mitchell, W.P., Kim, B.S., Oh, H.-B. and Kim, K.H. (2008) IDBD: infectious disease biomarker database, *Nucleic Acids Research*, **36**, D455-460.

Yang, S.-S., Lai, M.-Y., Chen, D.-S., Chen, G.-H. and Kao, J.-H. (2003) Mutations in the NS5A and E2-PePHD regions of hepatitis C virus genotype 1b and response to combination therapy of interferon plus ribavirin, *Liver Int*, **23**, 426-433.

Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood, *Comput Appl Biosci*, **13**, 555-556.

Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A.M. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites, *Genetics*, **155**, 431-449.

Yen, Y.-H., Hung, C.-H., Hu, T.-H., Chen, C.-H., Wu, C.-M., Wang, J.-H., Lu, S.-N. and Lee, C.-M. (2008) Mutations in the interferon sensitivity-determining region (nonstructural 5A amino acid 2209-2248) in patients with hepatitis C-1b infection and correlating response to combined therapy of pegylated interferon and ribavirin, *Aliment Pharmacol Ther*, **27**, 72-79.

Yuen, L.K.W., Ayres, A., Littlejohn, M., Colledge, D., Edgely, A., Maskill, W.J., Locarnini, S.A. and Bartholomeusz, A. (2007) SeqHepB: a sequence analysis program and relational database system for chronic hepatitis B, *Antiviral Research*, **75**, 64-74.

Yuste, E., Sánchez-Palomino, S., Casado, C., Domingo, E. and López-Gálvez, C. (1999) Drastic fitness loss in human immunodeficiency virus type 1 upon serial bottleneck events, *Journal of Virology*, **73**, 2745-2751.

Zaidi, N., Konstantinou, K. and Zervos, M. (2003) The role of molecular biology and nucleic acid technology in the study of human infection and epidemiology, *Arch Pathol Lab Med*, **127**, 1098-1105.

Zeuzem, S., Buti, M., Ferenci, P., Sperl, J., Horsmans, Y., Cianciara, J., Ibranyi, E., Weiland, O., Noviello, S., Brass, C. and Albrecht, J. (2006) Efficacy of 24 weeks treatment with peginterferon alfa-2b plus ribavirin in patients with chronic hepatitis C infected with genotype 1 and low pretreatment viremia, *Journal of Hepatology*, **44**, 97-103.

Zeuzem, S., Diago, M., Gane, E., Reddy, K.R., Pockros, P., Prati, D., Schiffman, M., Farci, P., Gitlin, N., O'Brien, C.B., Lamour, F., Lardelli, P. and Group, P.S.N.I. (2004) Peginterferon alfa-2a (40 kilodaltons) and ribavirin in patients with chronic hepatitis C and normal aminotransferase levels, *Gastroenterology*, **127**, 1724-1732.

Zeuzem, S., Feinman, S.V., Rasenack, J., Heathcote, E.J., Lai, M.Y., Gane, E., O'Grady, J., Reichen, J., Diago, M., Lin, A., Hoffman, J., Brunda, M.J. (2000) Peginterferon alfa-2a in patients with chronic hepatitis C, *The New England Journal of Medicine*, **343**, 1666-1672.

Zeuzem, S., Lee, J.H. and Roth, W.K. (1997) Mutations in the nonstructural 5A gene of European hepatitis C virus isolates and response to interferon alfa, *Hepatology*, **25**, 740-744.

8. ANEXOS

8 Anexos

8.1 Especificación de requisitos de la base de datos principal

Módulo bibliografía

- Tabla AUTORES: almacena información relativa a los autores de las referencias bibliográficas.

Atributo	Tipo	Descripción
author_id	varchar(6)	identificador del autor
author_firstname	varchar(50)	nombre del autor
author_lastname1	varchar(100)	primer apellido del autor
author_lastname2	varchar(100)	segundo apellido del autor
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

- Tabla GRUOS DE INVESTIGACIÓN: almacena información relativa a los grupos de investigación.

Atributo	Tipo	Descripción
group_id	char(10)	identificador del grupo
group_name	varchar(200)	nombre del grupo
group_description	varchar(250)	descripción del grupo
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

- Tabla INVESTIGADORES: almacena información relativa a los investigadores.

Atributo	Tipo	Descripción
researcher_id	varchar(7)	identificador del investigador
group_id	char(10)	identificador del grupo
first_name	varchar(20)	nombre
last_name1	varchar(100)	primer apellido
last_name2	varchar(100)	segundo apellido
dni	char(9)	documento de identidad
address_street	varchar(150)	calle
address_postalcode	char(5)	código postal
address_place	varchar(100)	lugar

telephone	char(9)	teléfono
e_mail	varchar(40)	email
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

- Tabla REFERENCIAS BIBLIOGRÁFICAS: almacena información relativa a las referencias bibliográficas.

Atributo	Tipo	Descripción
Referente_id	int(10) unsigned	identificador de la referencia
Title	varchar(250)	título
publication_date	date	fecha de publicación
publication	varchar(100)	nombre de la publicación
acceptance_date	date	fecha de aceptación
link	varchar(100)	enlace
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

- Tabla RELATION_N: almacena información relativa a la relación entre las referencias bibliográficas y los investigadores.

Atributo	Tipo	Descripción
researcher_id	varchar(7)	identificador del investigador
reference_id	int(10) unsigned	identificador de la referencia
project	varchar(20)	identificador de proyecto

- Tabla RELATION_P: almacena información relativa a la relación entre las referencias bibliográficas y los autores.

Atributo	Tipo	Descripción
reference_id	int(10) unsigned	identificador de la referencia
author_id	varchar(6)	identificador del autor
project	varchar(20)	identificador de proyecto

Módulo brotes epidemiológicos

- Tabla BROTES: almacena información relativa a los brotes epidemiológicos.

Atributo	Tipo	Descripción
ob_id	int(10) unsigned	identificador del brote
ob_name	varchar(100)	nombre
ob_city	varchar(70)	localidad

ob_region	varchar(70)	región
ob_country	varchar(70)	país
ob_startdate	date	fecha de inicio
ob_finaldate	date	fecha de finalización
ob_comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

Módulo fuentes de muestra

- Tabla AMBIENTES: almacena información relativa a los lugares ambientales de localización de las muestras.

Atributo	Tipo	Descripción
environment_id	int(10) unsigned	identificador del ambiente
collecting_point	varchar(50)	punto de recogida
company	varchar(50)	empresa donde se encuentra el punto de recogida
place	varchar(50)	localización del centro
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

- Tabla CENTROS DE SALUD: almacena información relativa a los centros de salud.

Atributo	Tipo	Descripción
centre_number	int(10) unsigned	identificador del centro
centre_name	varchar(50)	nombre del centro
place	varchar(50)	localización del centro
project	varchar(20)	identificador de proyecto

- Tabla CENTROS P: almacena información relativa a los pacientes de los centros de salud.

Atributo	Tipo	Descripción
centre_number	int(10) unsigned	identificador del centro
sip	char(8)	identificador del paciente
admission_date	date	fecha de admisión
centre_dossier	varchar(20)	identificador de historial
project	varchar(20)	identificador de proyecto

- Tabla EXTRANJEROS: almacena información relativa a los pacientes extranjeros.

Atributo	Tipo	Descripción
sip	char(8)	identificador del paciente
country	varchar(30)	país de origen
time_in_lcountry	varchar(20)	tiempo en el país actual
administrative_situation	varchar(20)	situación administrativa
health_card	varchar(50)	disposición de la tarjeta sanitaria
language_level	enum(without level,basic level,advanced level,perfect,unknown)	nivel de idioma local
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

- Tabla PACIENTES: almacena información relativa a los pacientes.

Atributo	Tipo	Descripción
sip	char(8)	identificador del paciente
birth_date	date	fecha de nacimiento
age	int(3) unsigned	edad
sex	enum(male,female,unknown)	sexo
risk_group	varchar(50)	grupo de riesgo
present_pregnancy	set(yes,no,unknown)	embarazo actual
stable_couple	enum(yes,no,unknown)	pareja estable
children_number	int(2) unsigned	número de hijos
educational_level	enum(without studies,primary school,secondary school,university,doctor,unknown)	nivel educativo
place_of_living	varchar(50)	lugar de residencia
nationality	varchar(50)	nacionalidad
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

Módulo información clínica

- Tabla ENFERMEDADES: almacena información relativa a las enfermedades.

Atributo	Tipo	Descripción
disease_name	varchar(150)	nombre de la enfermedad
project	varchar(20)	identificador de proyecto

- Tabla FACTORES DE PROTECCIÓN: almacena información relativa a los factores de protección.

Atributo	Tipo	Descripción
protectorfactor_id	smallint(4) unsigned	identificador del factor de protección
protectorfactor_name	varchar(60)	nombre del factor de protección
protectorfactor_description	varchar(200)	descripción del factor de protección
project	varchar(20)	identificador de proyecto

- Tabla FACTORES DE RIESGO: almacena información relativa a los factores de riesgo.

Atributo	Tipo	Descripción
riskfactor_id	smallint(4) unsigned	identificador del factor de riesgo
riskfactor_name	varchar(60)	nombre del factor de riesgo
riskfactor_description	varchar(200)	descripción del factor de riesgo
project	varchar(20)	identificador de proyecto

- Tabla INFORMACIÓN CLÍNICA: almacena información relativa a la información clínica del paciente.

Atributo	Tipo	Descripción
disease_name	varchar(150)	nombre de la enfermedad
sip	char(8)	identificador del paciente
information_date	date	fecha de la información
diagnosis_date	date	fecha de diagnóstico
centre_number	int(10) unsigned	identificador del centro de salud
start_date	date	fecha de inicio de los

		síntomas
pathological_basis	varchar(250)	patología de base
immunosupresor_state	varchar(50)	estado inmunosupresor
vaccinated_state	enum(complete,incomplete,unknown)	estado vacunal
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

- Tabla RELATION_A: almacena información relativa a la relación entre los factores de protección y la información clínica.

Atributo	Tipo	Descripción
protectorfactor_id	smallint(4) unsigned	identificador del factor de protección
sip	char(8)	identificador del paciente
disease_name	varchar(150)	nombre de la enfermedad
information_date	date	fecha de la información clínica
project	varchar(20)	identificador de proyecto

- Tabla RELATION_B: almacena información relativa a la relación entre los factores de riesgo y la información clínica.

Atributo	Tipo	Descripción
riskfactor_id	smallint(4) unsigned	identificador del factor de riesgo
sip	char(8)	identificador del paciente
disease_name	varchar(150)	nombre de la enfermedad
information_date	date	fecha de la información clínica
project	varchar(20)	identificador de proyecto

- Tabla RELATION_C: almacena información relativa a la relación entre los signos y la información clínica.

Atributo	Tipo	Descripción
sign_id	smallint(4) unsigned	identificador del signo
sip	char(8)	identificador del paciente
disease_name	varchar(150)	nombre de la enfermedad
information_date	date	fecha de la información clínica
project	varchar(20)	identificador de proyecto

- Tabla RELATION_D: almacena información relativa a la relación entre los síntomas y la información clínica.

Atributo	Tipo	Descripción
symptom_id	smallint(4) unsigned	identificador del síntoma
sip	char(8)	identificador del paciente
disease_name	varchar(150)	nombre de la enfermedad
information_date	date	fecha de la información clínica
project	varchar(20)	identificador de proyecto

- Tabla SIGNOS: almacena información relativa a los signos de los pacientes.

Atributo	Tipo	Descripción
sign_id	smallint(4) unsigned	identificador de signo
sign_name	varchar(60)	nombre del signo
sign_description	varchar(200)	descripción del signo
project	varchar(20)	identificador de proyecto

- Tabla SÍNTOMAS: almacena información relativa a los síntomas de los pacientes.

Atributo	Tipo	Descripción
symptom_id	smallint(4) unsigned	identificador de síntoma
symptom_name	varchar(60)	nombre del síntoma
symptom_description	varchar(200)	descripción del síntoma
project	varchar(20)	identificador de proyecto

- Tabla TEST DE PATÓGENO: almacena información relativa a los tests de patógeno realizados a pacientes.

Atributo	Tipo	Descripción
sip	char(8)	identificador del paciente
pathogen_id	varchar(10)	identificador del patógeno
last_negative_date	date	última fecha de negativo
first_positive_date	date	primera fecha de positivo
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

- Tabla VACUNACIONES: almacena información relativa a las vacunaciones de pacientes.

Atributo	Tipo	Descripción
vaccine_name	varchar(50)	nombre de la vacuna
sip	char(8)	identificador del paciente
vaccination_date	date	fecha de la vacunación
vaccine_dose	varchar(30)	dosis de la vacuna
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

- Tabla VACUNAS: almacena información relativa a las vacunas.

Atributo	Tipo	Descripción
Vaccine_name	varchar(50)	nombre de la vacuna
Vaccine_type	varchar(100)	tipo de vacuna
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

Módulo muestras

- Tabla ALMACENAMIENTOS: almacena información relativa al almacenamiento de muestras en el laboratorio.

Atributo	Tipo	Descripción
sample_id	varchar(10)	identificador de la muestra
samples_project	varchar(20)	identificador del proyecto de la muestra
storing_date	date	fecha de almacenamiento
quantity	varchar(50)	número de unidades
place	varchar(50)	lugar de almacenamiento
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

- Tabla MUESTRAS: almacena información relativa a las muestras.

Atributo	Tipo	Descripción
sample_id	varchar(10)	identificador de la muestra
project	varchar(20)	identificador del proyecto
sip	char(8)	identificador del paciente
environment_id	int(10) unsigned	identificador del ambiente
centre_number	int(10) unsigned	identificador del centro de salud

date_collected	date	fecha de obtención
date_arrived	date	fecha de llegada
date_extracted	date	fecha de extracción
comments	varchar(250)	comentarios

Módulo patógenos

- Tabla PATÓGENOS: almacena información relativa a los patógenos.

Atributo	Tipo	Descripción
pathogen_id	varchar(10)	identificador del patógeno
scientific_name	varchar(70)	nombre científico
common_name	varchar(70)	nombre común
pathogen_type	varchar(50)	tipo de patógeno
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

Módulo procesos de laboratorio

- Tabla AMPLIFICACIONES: almacena información relativa a las amplificaciones.

Atributo	Tipo	Descripción
amplification_id	int(10) unsigned	identificador de la amplificación
sample_id	varchar(10)	identificador de la muestra
samples_project	varchar(20)	identificador del proyecto de la muestra
amplification_result	enum(+,-, ,unknown)	resultado de la amplificación
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

- Tabla CEBADORES: almacena información relativa a los cebadores.

Atributo	Tipo	Descripción
primer_seq	varchar(250)	secuencia del cebador
primer_name	varchar(30)	nombre del cebador
position	int(10) unsigned	posición del cebador
sense	enum(sense,antise nse,unknown)	sentido del cebador
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

- Tabla EXTRACCIONES: almacena información relativa a las extracciones.

Atributo	Tipo	Descripción
extraction_id	int(10) unsigned	identificador de la extracción
sample_id	varchar(10)	identificador de la muestra
samples_project	varchar(20)	identificador del proyecto de la muestra
extraction_date	date	fecha de extracción
kit	varchar(30)	kit de extracción
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

- Tabla PROCESOS DE LABORATORIO: almacena información relativa a los procesamientos de las muestras realizados en el laboratorio.

Atributo	Tipo	Descripción
process_id	int(10) unsigned	identificador del proceso
sample_id	varchar(10)	identificador de muestra
samples_project	varchar(20)	identificador del proyecto de la muestra
extraction_id	int(10) unsigned	identificador de extracción
extractions_sample_id	varchar(10)	identificador de la muestra de la extracción
extractions_samples_project	varchar(20)	identificador del proyecto de la muestra de la extracción
region_name	varchar(15)	nombre de la región
amplification_id	int(10) unsigned	identificador de la amplificación
amplifications_sample_id	varchar(10)	identificador de la muestra de la amplificación
amplifications_samples_project	varchar(20)	identificador del proyecto de la muestra de la amplificación
sequencing_id	int(10) unsigned	identificador de la secuenciación
sequencing_sample_id	varchar(10)	identificador de la muestra de la secuenciación
sequencing_samples_project	varchar(20)	identificador del proyecto de la muestra de la secuenciación
comments	varchar(250)	comentarios
project	varchar(20)	identificador del proyecto

- Tabla REGIONES: almacena información relativa a las regiones.

Atributo	Tipo	Descripción
region_name	varchar(15)	nombre de la región
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

- Tabla RELATION_E: almacena información relativa a la relación entre cebadores y procesos de laboratorio.

Atributo	Tipo	Descripción
primer_seq	varchar(250)	secuencia del cebador
process_id	int(10) unsigned	identificador del proceso
project	varchar(20)	identificador de proyecto

- Tabla SECUENCIACIONES: almacena información relativa a las secuenciaciones.

Atributo	Tipo	Descripción
sequencing_id	int(10) unsigned	identificador de la secuenciación
sample_id	varchar(10)	identificador de la muestra
samples_project	varchar(20)	identificador del proyecto de la muestra
sequencing_result	enum(+,-, ,unknown)	resultado de la secuenciación
date_sent	date	fecha de envío
date_arrived	date	fecha de recepción
comments	varchar(250)	comentarios
project	varchar(20)	identificador de proyecto

Módulo pruebas y resultados clínicos

- Tabla LABORATORIOS: almacena información relativa a los laboratorios.

Atributo	Tipo	Descripción
lab_id	varchar(10)	identificador del laboratorio
lab_name	varchar(60)	nombre del laboratorio
lab_telephone	char(9)	teléfono
project	varchar(20)	identificador del proyecto

- Tabla PRUEBAS: donde se almacena la información relativa a las pruebas clínicas.

Atributo	Tipo	Descripción
test_id	smallint(4) unsigned	identificador de la prueba
test_name	varchar(60)	nombre de la prueba
spanish_name	varchar(60)	nombre de la prueba en castellano
speciality	smallint(2) unsigned	especialidad
project	varchar(20)	identificador del proyecto

- Tabla RESULTADOS: almacena información relativa al tipo de resultado.

Atributo	Tipo	Descripción
result_id	smallint(2) unsigned	identificador del resultado
result	varchar(15)	resultado
project	varchar(20)	identificador del proyecto

- Tabla RESULTADOS DE PRUEBA: almacena información relativa a los resultados de las pruebas clínicas.

Atributo	Tipo	Descripción
test_result_id	int(10) unsigned	identificador del resultado de prueba
sip	char(8)	identificador del paciente
analysis_date	date	fecha de la prueba
lab_id	varchar(10)	identificador de laboratorio
sample_id	varchar(10)	identificador de la muestra
samples_project	varchar(20)	proyecto de la muestra
result_id	smallint(2) unsigned	identificador del resultado
test_id	smallint(4) unsigned	identificador de la prueba
result	varchar(20)	valor del resultado
comments	varchar(250)	comentarios
project	varchar(20)	identificador del proyecto

Módulo resultados filogenéticos

- Tabla ALINEAMIENTOS: almacena información relativa a los alineamientos de secuencias.

Atributo	Tipo	Descripción
alignment_name	varchar(40)	nombre del alineamiento
date_aobtained	date	fecha de obtención

aarchive_location	varchar(50)	localización del archivo
alignment_type	set(multiple alignment,database search,pairwise comparison,unknown)	tipo de alineamiento
program	varchar(30)	programa utilizado
score	int(10) unsigned	puntuación obtenida
p_value	decimal(10,6)	valor p
e_value	decimal(10,6)	valor e
comments	varchar(250)	comentarios
project	varchar(20)	identificador del proyecto

- Tabla ÁRBOLES FILOGENÉTICOS: almacena información relativa a los árboles filogenéticos.

Atributo	Tipo	Descripción
tree_name	varchar(40)	nombre del árbol
alignment_name	varchar(40)	nombre del alineamiento
date_tobtained	date	fecha de obtención
tarchive_location	varchar(50)	localización del archivo
model	set(JC,F81,K80,HKY,TNef,TN,K81,K81uf,TIMef,TIM,TVMef,TVM,SYM,GTR,unknown)	modelo utilizado
program	varchar(20)	programa utilizado
options_invariants	decimal(5,4)	invariantes
options_gamma	decimal(7,4)	valor gamma
options_comments	varchar(250)	comentarios de las opciones
lrt_value	decimal(18,7)	valor LRT
lrt_pvalue	decimal(8,6)	valor p del LRT
lrt_df	int(3) unsigned	grados de libertad del LRT
lrt_comptree	varchar(40)	nombre del árbol con el que se compara para obtener el valor LRT
aic	decimal(15,4)	valor AIC
comments	varchar(250)	comentarios
project	varchar(20)	identificador del proyecto

Nota sobre los valores que toma el campo ‘model’ en la tabla ÁRBOLES FILOGENÉTICOS:

JC	Jukes and Cantor (Jukes and Cantor, 1969)
F81	Felsenstein 81 (Felsenstein, 1981)
K80	Kimura 80 (=K2P) (Kimura, 1980)
HKY	Hasegawa, Kishino, Yano 85 (Hasegawa, Kishino, Yano, 1985)
TNef	Tamura-Nei equal frequencies
TN	Tamura-Nei (Tamura and Nei, 1993)
K81	Two transversion parameters model 1 (=K81 =K3P) (Kimura, 1981)
K81uf	Two transversion parameters model 1 unequal frequencies
TIMef	Transitional model equal frequencies
TIM	Transitional model
TVMef	Tranversional model equal frequencies
TVM	Tranversional model
SYM	Symmetrical model (Zharkikh, 1994)
GTR	General time reversible (=REV) (Tavaré, 1986)

Módulo secuencias

- Tabla SECUENCIAS: almacena información relativa a las secuencias de material genético.

Atributo	Tipo	Descripción
sequence_id	int(10) unsigned	identificador de la secuencia
process_id	int(10) unsigned	identificador del proceso
alignment_name	varchar(40)	identificador del alineamiento
pathogen_id	varchar(10)	identificador del patógeno
ob_id	int(10) unsigned	identificador del brote
clone_id	varchar(25)	identificador del clon
codon_start	enum('1','2','3')	pauta de lectura
sequence	text	secuencia de material genético
region	varchar(15)	región donde se encuentra la secuencia
start_position	smallint(5) unsigned	posición de inicio
stop_position	smallint(5) unsigned	posición de terminación
date_obtained	date	fecha de obtención
typing_result	varchar(25)	tipado
NCBI_id	varchar(20)	identificador en NCBI
comments	varchar(250)	comentarios

project	varchar(20)	identificador del proyecto
---------	-------------	----------------------------

Módulo transmisiones

- Tabla TRANSMISIONES: almacena información relativa a las transmisiones.

Atributo	Tipo	Descripción
transmission_id	int(10) unsigned	identificador de la transmisión
pathogen_id	varchar(10)	identificador del patógeno
sip	char(8)	identificador del paciente
route	varchar(50)	vía de transmisión
others	varchar(150)	otras vías posibles
situation	varchar(50)	situación de la transmisión
date_probable	date	fecha de transmisión
transmission_country	varchar(30)	país de la transmisión
comments	varchar(250)	comentarios
Project	varchar(20)	identificador del proyecto

Módulo tratamientos

- Tabla FECHAS DE TRATAMIENTO: almacena información relativa a las fechas de tratamiento.

Atributo	Tipo	Descripción
start_date	date	fecha de inicio
sip	char(8)	identificador del paciente
treatment_name	varchar(50)	nombre del tratamiento
duration	varchar(30)	duración del tratamiento
given_dose	varchar(50)	dosis administrada
comments	varchar(250)	comentarios
project	varchar(20)	identificador del proyecto

- Tabla NOMBRES DE TRATAMIENTOS: almacena información relativa a los tratamientos.

Atributo	Tipo	Descripción
treatment_name	varchar(50)	nombre del tratamiento
treatment_description	varchar(200)	descripción del tratamiento
comments	varchar(250)	comentarios
project	varchar(20)	identificador del proyecto

- Tabla TRATAMIENTOS: almacena información relativa a los tratamientos recibidos por los pacientes.

Atributo	Tipo	Descripción
Sip	char(8)	identificador del paciente
treatment_name	varchar(50)	nombre del tratamiento
patient_response	varchar(50)	respuesta del paciente
treatment_number	int(10) unsigned	número de tratamiento
treatment_completed	set(yes,no,unknown)	tratamiento completado
comments	varchar(250)	comentarios
project	varchar(20)	identificador del proyecto

Tablas fuera de módulos

- Tabla RELATION_CC: almacena información relativa a la relación entre las muestras y los patógenos.

Atributo	Tipo	Descripción
sample_id	varchar(10)	identificador de la muestra
samples_project	varchar(20)	identificador del proyecto de la muestra
pathogen_id	varchar(10)	identificador del patógeno
project	varchar(20)	identificador del proyecto

- Tabla RELATION_J: almacena información relativa a la relación entre las secuencias y las referencias bibliográficas.

Atributo	Tipo	Descripción
sequence_id	int(10) unsigned	identificador de la secuencia
reference_id	int(10) unsigned	identificador de la referencia
project	varchar(20)	identificador del proyecto

- Tabla RELATION_K: almacena información relativa a la relación entre las secuencias y los investigadores.

Atributo	Tipo	Descripción
researcher_id	varchar(7)	identificador del investigador
sequence_id	int(10) unsigned	identificador de la secuencia

project	varchar(20)	identificador del proyecto
---------	-------------	----------------------------

- Tabla RELATION_R: almacena información relativa a la relación entre los tratamientos y los patógenos.

Atributo	Tipo	Descripción
treatment_name	varchar(50)	nombre del tratamiento
sip	char(8)	identificador del paciente
pathogen_id	varchar(10)	identificador del patógeno
project	varchar(20)	identificador del proyecto

8.2 Especificación de requisitos de la base de datos secundaria

Módulo fecha de modificación

- Tabla DELETE_DATES: almacena información relativa a las fechas de borrado de registros en la base de datos principal.

Atributo	Tipo	Descripción
date	datetime	fecha de modificación
username	varchar(20)	nombre de usuario
table_name	varchar(25)	nombre de la tabla
field_name	varchar(30)	nombre del campo

- Tabla INSERT_DATES: almacena información relativa a las fechas de inserción de registros en la base de datos principal.

Atributo	Tipo	Descripción
date	datetime	fecha de modificación
username	varchar(20)	nombre de usuario
table_name	varchar(25)	nombre de la tabla
field_name	varchar(30)	nombre del campo

- Tabla UPDATE_DATES: almacena información relativa a las fechas de actualización de registros en la base de datos principal.

Atributo	Tipo	Descripción
date	datetime	fecha de modificación
username	varchar(20)	nombre de usuario
table_name	varchar(25)	nombre de la tabla
field_name	varchar(30)	nombre del campo

Módulo modificación de la información

- Tabla DB_FIELDS: almacena información relativa a los campos de las tablas de la base de datos principal.

Atributo	Tipo	Descripción
field_name	varchar(30)	nombre del campo

- Tabla DB_TABLES: almacena información relativa a las tablas de la base de datos principal.

Atributo	Tipo	Descripción
table_name	varchar(25)	nombre de la tabla

- Tabla FILLED_POINT: almacena información relativa a la relación entre campos y tablas de la base de datos principal.

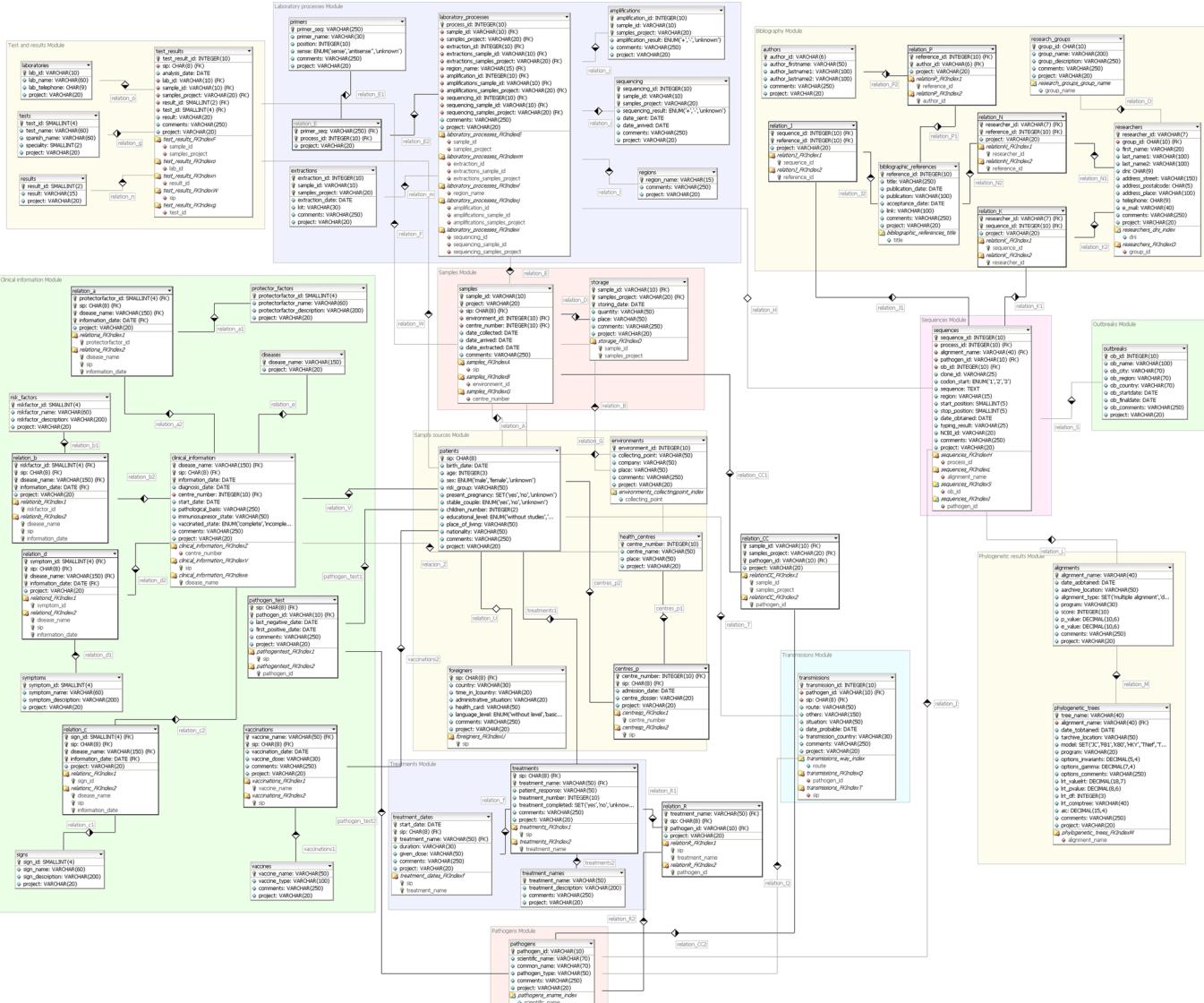
Atributo	Tipo	Descripción
field_name	varchar(30)	nombre del campo
table_name	varchar(25)	nombre de la tabla

Módulo usuarios

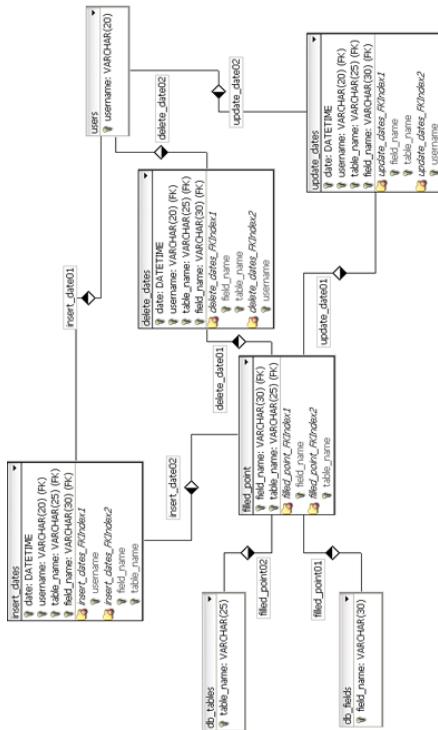
- Tabla USERS: almacena información relativa a los usuarios de la plataforma.

Atributo	Tipo	Descripción
username	varchar(20)	nombre del usuario

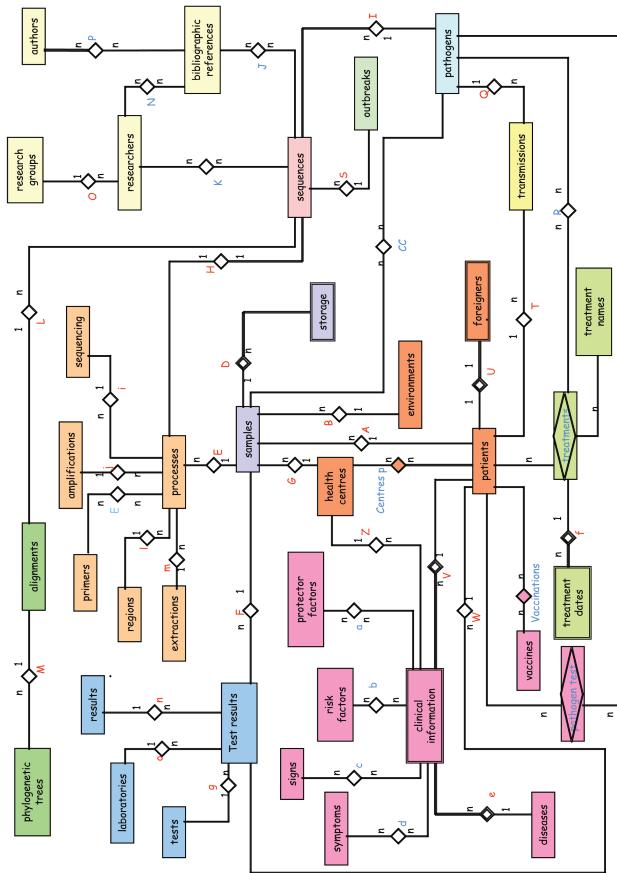
8.3 Esquema de la base de datos principal



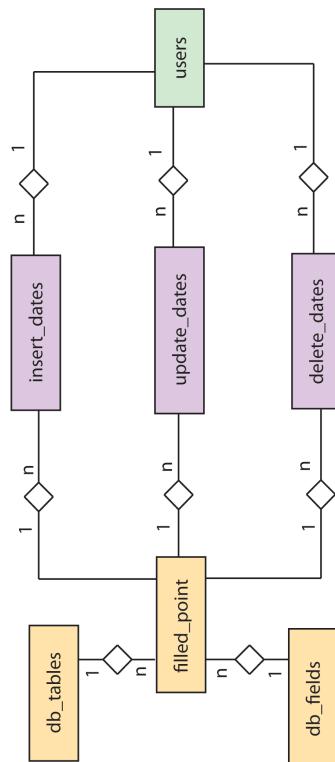
8.4 Esquema de la base de datos secundaria



8.5 Diagrama ER de la base de datos principal



8.6 Diagrama ER de la base de datos secundaria



9. LISTADO DE ABREVIATURAS

9 Listado de abreviaturas

AIC	<i>Akaike information criterion</i>
ALT	Alanina aminotransferasa
ANOVA	<i>Analysis of variance</i>
BLAST	<i>Basic local alignment search tool</i>
CTL	<i>Cytotoxic T lymphocytes</i>
CSS	<i>Cascading style sheets</i>
DDBJ	<i>DNA data bank of Japan</i>
DDL	<i>Data definition language</i>
DML	<i>Data manipulation language</i>
DNA	<i>Deoxyribonucleic acid</i>
EBI	<i>European bioinformatics institute</i>
EM	Expectación-maximización
EMBL	<i>European molecular biology laboratory</i>
EMBOSS	<i>The european molecular biology open software suite</i>
ER	Diagrama entidad-relación
ETR	<i>End-of-treatment response</i>
euHCVdb	<i>European hepatitis C virus database</i>
EVR	<i>Early virological response</i>
FFT	<i>Fast fourier transform</i>
GLM	<i>General linear models</i>
GOT	<i>Glutamic-oxaloacetic transaminase</i>
GPT	<i>Glutamic-pyruvic transaminase</i>
HCV	<i>Hepatitis C virus, VHC</i>
HSP	<i>High scoring pairs</i>
HTML	<i>Hypertext markup language</i>
HVDB	<i>Hepatitis virus database</i>
HVR	<i>Hypervariable region</i>
IDE	<i>Integrated development environment</i>

IDs	Identificadores
IFN	Interferón
INSDC	<i>International nucleotide sequence database collaboration</i>
ISDR	<i>Interferon sensitivity determining region</i>
IUPAC	<i>International union of pure and applied chemistry</i>
Kb	Kilobases
LANL-HCV	<i>Los alamos national laboratory – hepatitis C virus</i>
LIMS	<i>Laboratory information management system</i>
MAFFT	<i>Multiple alignment using fast fourier transform</i>
MCA	<i>Multiple correspondence analysis</i>
MSA	<i>Multiple sequence alignment</i>
MUSCLE	<i>Multiple sequence comparison by log-expectation</i>
NCBI	<i>National center for biotechnology information</i>
NGS	<i>Next-generation sequencing</i>
NIAID	<i>National institute of allergy and infectious diseases</i>
NK	<i>Natural killer</i>
NPV	<i>Negative predictive value</i>
OMS	Organización mundial de la salud, <i>WHO</i>
OR	<i>Odds ratio</i>
PCA	<i>Principal component analysis</i>
PCR	<i>Polymerase chain reaction</i>
PHP	<i>PHP hypertext pre-processor</i>
PKR	<i>Protein kinase R</i>
PPV	<i>Positive predictive value</i>
ProbCons	<i>Probabilistic consistency-based multiple sequence alignment</i>
RBV	Ribavirina
RdRp	<i>RNA-dependent RNA polymerase</i>
RE	Retículo endoplasmático
RNA	<i>Ribonucleic acid</i>

RVR	<i>Rapid virological response</i>
SGBD	Sistema gestor de bases de datos
SI	Sistema de información
SQL	<i>Structured query language</i>
SVR	<i>Sustained virological response</i>
T-COFFEE	<i>Tree based consistency objective function for alignment evaluation</i>
UPGMA	<i>Unweighted pair group method with arithmetic mean</i>
VIH	Virus de la inmunodeficiencia humana, <i>HIV</i>
W3C	<i>World wide web consortium</i>
WSP	<i>Weighted sum-of-pairs</i>
XHTML	<i>eXtensible hypertext markup language</i>
XML	<i>eXtensible markup language</i>

