Tesis doctoral:

Diseño y Desarrollo de un Sistema de Información para la Gestión de Información sobre Cáncer de Mama

Verónica Burriel Coll

Directores: Prof. Dr. Óscar Pastor López
Dra. Gloria Ribas Despuig

Programa de Doctorado en Tecnologías para la Salud y el Bienestar

Julio 2017

Esta tesis ha sido redactada para la obtención del título de Doctor en Tecnologías para la Salud y el Bienestar por la Universitat Politècnica de València y defendida el 6 de julio de 2017.

Autora:
Verónica Burriel Coll, Universitat Politècnica de València, España.

Directores:
Prof. Dr. Óscar Pastor López, Universitat Politècnica de València, España.
Dra. Gloria Ribas Despuig, Fundación de Investigación Incliva, España.

Comité de Evaluadores Externos:
Dr. Diego Álvarez Estévez, Haaglanden Medisch Centrum, Países Bajos.
Prof. Dr. Manuel Noguera García, Universidad de Granada, España.
Prof. Dr. Jesús Peral Cortés, Universidad de Alicante, España.

Tribunal de Defensa:
Presidente: Prof. Dr. Vicente Traver Salcedo, Universitat Politècnica de València, España.
Secretario: Prof. Dr. Juan Carlos Trujillo Mondejar, Universidad de Alicante, España.
Vocal: Dra. Mª Adela Cañete Nieto, Instituto de Investigación Sanitaria La Fe, España.

"Vive como si fueras a morir mañana,
aprende como si fueras a vivir siempre"
Gandhi

Agradecimientos

Llegado este momento en el que ya ves que se acerca el día de recoger el fruto de la semilla que con tanto mimo, paciencia y esfuerzo plantaste y cuidaste durante cuatro años, no puedes evitar echar la vista atrás y ver que ha habido muchas personas a tu lado que también han contribuido de una u otra manera a conseguir lo que estás a punto de lograr. Cuando era niña me enseñaron que hay que estar agradecida con los regalos que te da la vida y, para mí, esta experiencia y cada una de las personas que habéis participado en ella, sois el mejor regalo.

Para empezar, quiero agradecer a mis directores de tesis toda la sabiduría, apoyo y ánimo que me han transmitido durante estos años. Óscar, muchas gracias por estar a mi lado desde el principio. Por confiar en mí desde el día en el que nos conocimos en el aula de informática, en las prácticas del máster, cuando me diste la oportunidad de empezar a trabajar con vosotros hace ya ocho años. Por todo lo que me has enseñado sobre la investigación y la vida. Por enseñarme y ayudarme a mejorar profesionalmente y apoyarme y valorarme como investigadora. Tu carácter valiente, positivo y alegre es una inspiración para mí. Gloria, gracias por creer en mi tesis desde el día que empezamos a trabajar en ella. Por todos los conocimientos sobre oncología y biología del cáncer de mama que me has transmitido. Por abrirme las puertas de vuestro laboratorio y dejarme formar parte de vuestro fascinante proyecto, aportando mi granito de arena en la lucha contra esta enfermedad. Por contagiarme tu pasión por la investigación en medicina y tu alegría en cada visita. Pero sin duda, lo que más valoro de toda esta experiencia a vuestro lado, es que me llevo dos grandes amigos de los que el tiempo no es capaz de borrar.

Además, me gustaría agradecer a los miembros del tribunal y del comité de evaluación el haber aceptado participar en esta última etapa de mi tesis, lo que supone para mí un gran honor.

Mis compañeros de batalla también forman parte de este regalo. Quiero dar las gracias a mis compañeros del PROS. A Ana, José, Carlos, Alberto, Michael, Sipan y Hendrik por vuestro apoyo y ánimos hasta el último día. A Ignacio por ayudarme en la validación y transmitir tu simpático humor cuando apareces por el laboratorio. A José Marín por leerte mi tesis y darme sabios consejos. A los profes, Juan Carlos, Laura, Mariángeles y Mati, por todo el conocimiento compartido y los cientos de horas de reuniones debatiendo el modelo del genoma. A Ana Cidad por estar siempre apoyando en la retaguardia.

Y también a los que ya no estáis en PROS. Al grupito de Acero Azul, Sergio, Marce y Mario, grandes compañeros y mejores amigos, por todas las risas, juergas y salseos que quedarán grabados en nuestra memoria para siempre. A Mariajo, Paco y Caro, por los buenos momentos compartidos y las inolvidables tartas de naranja, chocolate y dulce de leche. A Ana Levin, por ser el primer apoyo cuando llegué y alguien de quien aprendí mucho más que genómica. Al *Dark Side*, Noni, Rupis, Ani, Enro, Costi, Flan, Vicky, George y Darks, por las fiestas temáticas y no temáticas, los días de risas y bromas interminables y nuestra bonita amistad. A Ricardo, por ayudarme con mi tesis y convertirte en un amigo. A Isa, Natalia, Alba y Mireia, por interesaros en mi trabajo y querer participar en él. A Anna, por los buenos ratos y las interesantes noticias de las reuniones.

También forman parte de los compañeros de batalla, aquellas personas con las que he compartido agradables momentos en el Incliva y que también merecen mi agradecimiento. A María por tu paciencia enseñándome como obtener y manejar los datos y tantas horas compartidas. A Maite por enseñarme todo lo que sé sobre el cáncer de mama, revisar la parte médica de mi tesis y ayudarme en todo lo que he necesitado. A Tania porque, aunque llegaste casi al final de esta etapa, tus ánimos me han ayudado mucho. A Ana porque fuimos compañeras del máster y de doctorado y lo hemos logrado.

Tampoco puedo olvidarme en este momento de mis amigos. He recibido ánimos y apoyo de cada uno de vosotros durante toda esta época que está

a punto de acabar. Por lo tanto, quiero daros gracias a todos. A Adrianet, Adrianot, Alberto, Ana, Chris, Clara, Cris, Eva, Guille, Higino, Joan, Laura, Litos, Marín, Marisa, Patri, Sara y Xeni, por ser tan geniales, por estar ahí desde siempre y por todas las alegrías y momentos vividos que hacen cualquier camino más fácil. En especial a Ana por ser tan especial para mí, por estar apoyándome desde el minuto uno, preocupándote por el estado de la tesis y mostrando día a día lo orgullosa que estás de mí. Y también a Eva por regalarme tus palabras bonitas y tener tu apoyo siempre que lo he necesitado. A Sara y Toni por los buenos ratos y los mejores, por vuestra amistad especial, por ser como sois. A Cris e Iván por los momentos vividos en el pisito y fuera de él, por lo molones que sois y porque la vida nos unió por casualidad y fue una casualidad maravillosa. A Rafa y Liza por ser la parejita más peculiar y aventurera y hacer que cualquier rato juntos se convierta en una fiesta. A Lara por compartir tantas horas de piso juntas y tardes de compras y saber que siempre estás cerquita, para lo que necesite, en la calle de al lado. A mi Ru, mi "vecina rubia", por ser mi amiga desde siempre y para siempre y por contagiarme tu alegría siempre que estás cerca.

Para acabar, me queda agradecer este momento a mi familia. Ellos son mi pilar más fuerte y esta tesis no habría llegado a este punto sin su apoyo. Ante todo y sobre todo, a mis padres, Mari Carmen y Juangi, por vuestro amor incondicional. Por cuidarme y apoyarme siempre en todas las etapas de mi vida. Por vuestra alegría. Por ser para mí un ejemplo a seguir y enseñarme que en esta vida con esfuerzo y trabajo puedo llegar donde me proponga. Porque no hay nada que me haya dado más fuerzas para llegar hasta aquí que saber que estáis tan orgullosos de mí como yo de vosotros.

A mis tíos, primos y sobrinos, Irene, Antonio, Jaby, Iskia, Erik, Rosa Mari y José Luis, por vuestro cariño, por darme tanta felicidad, por estar siempre ahí y por creer en mí. También a Álvaro, Charo, Saida, Mario, Álvaro, Cristina y Martina, por vuestro ánimo y cariño. A mi "primi", Cristina, por ser tan especial y cariñosa y apoyarme en la distancia. A mis abuelitos, Luis y Mariano, porque seguro que se sienten orgullosos allá donde estén. A mi abuelita Beatriz porque, además de orgullosa, está a mi lado para celebrarlo.

Y a mi abuelita Natalia por nuestra adoración mutua, porque ahora mismo será la persona más orgullosa del cielo y porque su forma de ser y carácter me sirvieron de inspiración para ser como soy. A mi Brisa porque estará dando saltos de alegría en el arcoíris. A mi familia gallega, Guilli, Ángeles, Paula, Marta, Javi, Carlota y Saúl, por vuestros ánimos en la distancia, vuestro cariño y por hacer que sea una más de la familia.

Y, por último, a mi chico Diego. Por toda la fuerza que me has dado en forma de *chuchonsitos*, *carisitos* y *carimitos* para que acabe esta tesis. Por tu paciencia infinita aguantando mis horarios nocturnos de tesis, mis viajes a congresos y mis subidas y bajadas de ánimo en momentos complicados. Por estar a mi lado día a día en esta etapa de mi tesis y en el resto de mi vida. Por ese beso al despertarme, el de antes de dormirme y los del resto del día. Por apoyarme en todos mis sueños e ilusiones. Por hacerme reír todos los días. Porque no he podido encontrar mejor compañero de viaje y de vida. Porque nos quedan millones de aventuras, risas, alegrías, sorpresas y experiencias por vivir y pienso disfrutar cada una de ellas contigo. Porque, gracias a ti, soy feliz.

Puede que me haya olvidado de mencionar a alguien que, de alguna u otra manera, haya formado parte de esta etapa. A vosotros también, gracias. Porque gracias a todos vuestros granitos de arena, por fin, este momento ha llegado.

Resumen

El diagnóstico, tratamiento e investigación sobre enfermedades tan complejas como el cáncer de mama es una tarea cada vez más complicada por la gran cantidad y diversidad de datos implicados y por la necesidad de relacionarlos adecuadamente para obtener conclusiones relevantes. La generación de los datos clínicos tiene que estar acompañada de una gestión eficiente de los mismos. Ello hace imprescindible la utilización de tecnologías avanzadas de Sistemas de Información que aseguren un correcto almacenamiento, gestión y explotación de los datos.

Tras un profundo estudio del dominio y de las tecnologías utilizadas para el almacenamiento y gestión de datos clínicos y biológicos sobre la enfermedad, el objetivo principal de esta tesis es ofrecer una base metodológica que permita diseñar y desarrollar sistemas software para la manipulación eficiente y fiable de la información sobre el cáncer de mama. La utilización de técnicas de Modelado Conceptual en un entorno donde su uso no es tan habitual como debiera ser, permitirá disponer de un sistema de información perfectamente adaptado al dominio de aplicación.

Bajo este planteamiento, en esta tesis se ha llevado a cabo el modelado conceptual del dominio del diagnóstico, tratamiento e investigación del cáncer de mama, el diseño de arquetipos bajo el estándar ISO13606 para ofrecer interoperabilidad entre sistemas, la integración de datos de distintos orígenes relacionados con el cáncer de mama en una base de datos unificadora y el diseño de un prototipo de herramienta de gestión y análisis de datos clínicos y de expresión génica. Para validar la idoneidad de esta propuesta, se ha llevado a cabo un proceso de validación en un entorno real como es la Fundación de Investigación INCLIVA de Valencia, donde investigadores clínicos y biólogos han probado y valorado la eficiencia de la solución planteada en esta tesis doctoral.

Resum

El diagnòstic, tractament i investigació sobre malalties tan complexes com ara el càncer de mama és una tasca cada vegada més complexa per la gran quantitat i diversitat de dades implicades i per la necessitat de relacionar-les adequadament per a obtenir conclusions rellevants. La generació de dades clíniques ha d'estar acompanyada d'una gestió eficient de les mateixes. Açò fa imprescindible la utilització de tecnologies avançades de Sistemes d'Informació que asseguren un correcte emmagatzematge, gestió i explotació de les dades.

Després d'un profund estudi del domini i de les tecnologies utilitzades per l'emmagatzematge i gestió de dades clíniques i biològiques sobre la malaltia, el principal objectiu d'aquesta tesi és oferir una base metodològica que permeta dissenyar i desenvolupar sistemes programaris per a la manipulació eficient i fiable de la informació sobre el càncer de mama. La utilització de tècniques de Modelat Conceptual en un entorn on el seu ús no és tan habitual com deuria ser, permetrà disposar d'un sistema d'informació perfectament adaptat al domini d'aplicació.

Baix aquest plantejament, en aquesta tesi s'ha dut a terme el modelat conceptual del domini del diagnòstic, tractament i investigació del càncer de mama, el disseny d'arquetips baix l'estàndard ISO13606 per oferir interoperabilitat entre sistemes, la integració de dades de distints orígens sobre el càncer de mama en una base de dades unificadora i el disseny d'un prototip d'eina de gestió i anàlisi de dades clíniques i d'expressió gènica. Per a validar la idoneïtat d'aquesta proposta, s'ha dut a terme un procés de validació en un entorn real com és la Fundació d'Investigació INCLIVA de València, on investigadors clínics i biòlegs han provat i valorat l'eficiència de la solució plantejada en aquesta tesi doctoral.

Abstract

Diagnosis, treatment and research about such complex diseases as breast cancer is an increasingly complex task due to the big quantity and diversity of involved data and the need of relating them properly to obtain relevant conclusions. Clinical data generation has to be followed by an efficient data management. So, the use of advanced information system technologies is essential to ensure a correct storage, management and exploitation of data.

Following a deep study of domain and technologies used to store and manage clinical and biological data about the disease, the main goal of this thesis is to provide a methodological basis to design and implement software systems to manage breast cancer data in a trustable and efficient way. Using Conceptual Modelling techniques in an environment where their use is not as common as it should be, allows to create information systems perfectly adapted to the studied domain.

Under this approach, in this thesis some tasks have been carried out among which are conceptual modelling of diagnosis, treatment and research of breast cancer's domain; archetypes' designing under ISO13606 standard to allow systems interoperability; breast cancer data integration from different data sources in a unified database; and designing a prototype of tool for managing and analysing clinical and genic expression data. In order to validate the proposal, a validation process in a real environment as Research Foundation INCLIVA in Valencia has been carried out. During this process, medical and biological researchers have use and assess the efficiency of solution proposed in this doctoral thesis.

ÍNDICE

1. Introducción 1

1.1. Motivación 5

1.1.1. El Cáncer de Mama en la sociedad 5

1.1.2. La genómica y el Cáncer de Mama. 10

1.2. Objetivos 18

1.3. Metodología de investigación 20

1.3.1. Aplicación de la metodología de investigación 22

1.4. Estructura y planificación de esta Tesis Doctoral 25

2. Estado del arte 27

2.1. Estudio del dominio del diagnóstico y tratamiento del Cáncer de Mama 27

2.1.1. Clasificación histológica 28

2.1.2. Clasificación según el perfil molecular 30

2.1.3. Técnicas utilizadas actualmente para el diagnóstico del cáncer de mama 33

2.1.4. Guías de tratamiento 36

2.1.5. Procedimiento clínico para el diagnóstico y tratamiento del cáncer de mama 36

2.1.6. Almacenamiento y gestión de los datos en los hospitales y los laboratorios genéticos. 45

2.2. Estudio de las bases de datos sobre el Cáncer de Mama 57

2.2.1. Bases de datos genéticas 58

2.2.2. Bases de datos con información sobre microARNs 83

2.2.3. Bases de datos con información específica del cáncer de mama 90

2.2.4. Conclusiones 96

2.3. Herramientas o tecnologías existentes para la gestión de datos clínicos 101

2.3.1. Historia Clínica Electrónica: Estándar ISO 13606 101

2.3.2. Bases de datos para la investigación traslacional 110

2.3.3. Registro del cáncer 113

2.3.4. Orion Clinic 115

2.3.5. Conclusiones 118

2.4. Justificación de la utilización de técnicas de Modelado Conceptual en el entorno clínico y biológico 121

3. Diseño de la solución 127

3.1. Modelado Conceptual y diseño del Sistema de Información 127

3.1.1. Perspectiva Clínica 128

3.1.2. Perspectiva de Expresión Génica 149

3.1.3. Perspectiva de Secuenciación Masiva 156

3.2. Estandarización de la información relacionada con el Cáncer de Mama utilizando el estándar ISO 13606 164

3.2.1. Diseño de los arquetipos relacionados con la perspectiva Clínica. 166

3.2.2. Diseño de los arquetipos relacionados con la perspectiva de Expresión Génica 176

3.2.3. Diseño de los arquetipos relacionados con la perspectiva de Secuenciación Masiva 180

3.3. Implementación y carga de la base de datos 184

3.3.1. Estudio previo a la selección de un Sistema de Gestión de Bases de datos 184

3.3.2. Inferencia del Esquema de base de datos e implementación de la base de datos en MySQL 203

3.3.3. Integración de la información pública relevante para el Cáncer de Mama en la Base de Datos 205

3.4. Diseño e implementación de una herramienta software para la gestión de datos clínicos y biológicos del Cáncer de Mama 225

3.4.1. Ventana de inicio 226

3.4.2. Lista de pacientes 227

3.4.3. Paciente 227

3.4.4. Episodio 232

3.4.5. Lista de tratamientos, pruebas, muestras, metástasis o síntomas 233

3.4.6. Tratamiento 234

3.4.7. Prueba 235

3.4.8. Muestra 236

3.4.9. Metástasis 238

3.4.10. Síntoma 239

3.4.11. Filtrado de pacientes 240

3.4.12. Carga de datos de microARNs 241

3.4.13. Análisis de expresión de microARNs 244

3.4.14. Consulta de análisis por microARNs 247

3.4.15. Consulta de análisis por genes afectos 248

3.4.16. Consulta de análisis por *pathways* alterados. 249

4. Validación: Implantación de la solución en el Proyecto Cáncer de Mama en Mujeres Jóvenes 253

4.1. Objetivo 254

4.2. Caso de estudio 256

4.2.1. Qué es el INCLIVA....256
4.2.2. Qué es el proyecto de Cáncer de Mama en Mujeres Jóvenes.257
4.2.3. Necesidades del grupo de investigación....258
4.2.4. Sujetos experimentales....259
4.2.5. Preparación del sistema: Carga de la base de datos con los datos clínicos....259
4.3. Experimento....279
4.3.1. Preguntas de investigación y formulación de la hipótesis....279
4.3.2. Factores y tratamientos....280
4.3.3. Variables respuesta y métrica....281
4.3.4. Diseño del experimento....283
4.3.5. Procedimiento del experimento....283
4.3.6. Amenazas a la validez....285
4.3.7. Instrumentos utilizados en el experimento....286
4.4. Resultados....290
4.4.1. Análisis de datos....290
4.4.2. Sesión de focus-group....299
5. Conclusiones....309
5.1. Trabajo futuro....313
5.2. Aportación a la comunidad científica....316
5.2.1. Publicaciones....316
5.2.2. Participación en proyectos de investigación....318
5.2.3. Organización de congresos....319
6. Referencias bibliográficas....321
Anexo I: Cuestionarios diseñados para la validación....333
Anexo II: Tabla de resultados de la validación....337

Índice de figuras

Figura 1. Evolución anual de la incidencia del Cáncer de Mama en mujeres en distintos países. Valores por cada 100.000 habitantes (GLOBOCAN 2012) 6
Figura 2. Evolución anual de la mortalidad del Cáncer de Mama en mujeres en distintos países. Valores por cada 100.000 habitantes (GLOBOCAN 2012) 7
Figura 3. Tasas de incidencia y mortalidad de distintos tipos de cáncer en mujeres en España en 2012 (Sociedad Española de Oncología Médica SEOM. 2014) 8
Figura 4. Distribución del cáncer de mama familiar. 11
Figura 5. Ciclos regulativos anidados que representan la metodología de investigación seguida en esta tesis. 23
Figura 6. Procedimiento de diagnóstico y tratamiento del cáncer de mama. Etapa 1: Sospecha inicial radiológica 37
Figura 7. Procedimiento de diagnóstico y tratamiento del cáncer de mama. Etapa 2: Estudio de la muestra 38
Figura 8. Componentes del Comité de tumores 39
Figura 9. Procedimiento de diagnóstico y tratamiento del cáncer de mama. Etapa 3: Comité de tumores 42
Figura 10. Procedimiento de diagnóstico y tratamiento del cáncer de mama. Etapa 4: Tratamiento 43
Figura 11. Procedimiento de diagnóstico y tratamiento del cáncer de mama. Etapa 5: Seguimiento de la paciente 44
Figura 12. Carpetas que forman el Historial Clínico de un paciente en formato papel. 45
Figura 13. Plantilla utilizada en la Unidad de Hematología y Oncología del Hospital Clínico de Valencia para almacenar los datos de las pacientes de Cáncer de Mama 47

Figura 14. Detalle de un informe de diagnóstico realizado a partir de la plantilla en la Unidad de Hematología y Oncología del Hospital Clínico de Valencia. 48
Figura 15. Ejemplo de la hoja de cálculo de Microsoft Excel utilizada para almacenar datos clínicos para investigación. 50
Figura 16. Ejemplo del archivo de Microsoft Excel donde se almacenan los datos de extracción de ARN 51
Figura 17. Ejemplo del fichero de Microsoft Excel donde se almacenan las expresiones de los microARNs analizados en cada muestra mediante el chip de expresión. 52
Figura 18. Ejemplo del fichero de Microsoft Excel donde se almacenan los resultados del T-Test 53
Figura 19. Hoja del fichero Excel donde guardan los resultados del análisis realizado utilizando la herramienta DIANA mirPath 54
Figura 20. Frangmento de fichero Excel donde se almacenan los datos de los resultados de la PCR-Cuantitativa 55
Figura 21. Ejemplo del fichero de Microsoft Excel donde se almacenan los datos estadísticos finales 55
Figura 22. Logo de dbSNP 58
Figura 23. Logo de Ensembl 64
Figura 24. Logo de HGMD 67
Figura 25. Logo de UniProt 70
Figura 26. Logo de COSMIC 74
Figura 27. Logo de GWAS Catalog 77
Figura 28. Logo de OMIM 79
Figura 29. Logo de KEGG 80
Figura 30. Logo de miRBase 84
Figura 31. Logo DIANA Tools 86
Figura 32. Logo de TargetScan Human 89
Figura 33. Logo de G2SBC 91
Figura 34. Estructura de la herramienta LinkEHR 104
Figura 35. Modelo de Referencia resumido del estándar ISO 13606 106
Figura 36. Representación gráfica de la jerarquía de clases del Modelo de Referencia 107

Figura 37. Arquetipo Historia Clínica Resumida creado por el MSSSI.....109
Figura 38. Captura de la pantalla de entrada de Orion Clinic.....115
Figura 39. Perspectiva Clínica. Parte I. 139
Figura 40. Perspectiva Clínica. Parte II.....143
Figura 41. Perspectiva de Expresión Génica.....155
Figura 42. Perspectiva de Secuenciación Masiva 162
Figura 43. Arquetipo *CEN-EN13606-FOLDER.Cancer_de_Mama.v1.adl*166
Figura 44. Arquetipo CEN-EN13606-COMPOSITION.Anamnesis.v1.adl 167
Figura 45. Detalle del *Archetype Slot* que hace referencia al arquetipo *CEN-EN13606-ENTRY.Vida _Reproductiva.v1.adl*.....168
Figura 46. Arquetipo *CEN-EN13606-ENTRY.Vida_Reproductiva.v1.adl* 168
Figura 47. Detalle del arquetipo *CEN-EN13606-ENTRY.Vida _Reproductiva.v1.adl* donde se muestran los tipos de valores. 169
Figura 48. Arquetipo *CEN-EN13606-COMPOSITION.Episodio.v2.adl*...170
Figura 49. Arquetipo *CEN-EN13606-SECTION.Muestras.v1.adl*.....171
Figura 50. Arquetipo *CEN-EN13606-SECTION.Pruebas.v1.adl*.....172
Figura 51. Arquetipo *CEN-EN13606-SECTION.Sintomas.v1.adl*.....173
Figura 52. Arquetipo *CEN-EN13606-SECTION.Tratamientos.v1.adl*.....174
Figura 53. Arquetipo *CEN-EN13606-CLUSTER.Toxicidad.v1.adl*.....175
Figura 54. Arquetipo CEN-EN13606-COMPOSITION.Informe_de_expresion .v1.adl.....176
Figura 55. Arquetipo CEN-EN13606-CLUSTER.Lista_de_expresiones.v1.adl.....178
Figura 56. Arquetipo *CEN-EN13606-CLUSTER.Gen.v1.adl*.....179
Figura 57. Arquetipo CEN-EN13606-COMPOSITION.Informe_genomico.v1.adl.....181
Figura 58. Arquetipo CEN-EN13606-CLUSTER.Lista_de_variaciones.v1.adl 182
Figura 59. MySQL Workbench 6.3 CE. Hoja de propiedades de la tabla "Paciente".....204
Figura 60. Representación gráfica de la carga de dbSNP.....208
Figura 61. Proceso de extracción, transformación y carga de BIC.....216

Figura 62. Detalle del fichero "hsa.gff3" de miRBase con los microARNs 220
Figura 63. Detalle del fichero "microT-CDS_data" 222
Figura 64. Detalle del fichero "GO_Biological_Process_2015.sdx" 223
Figura 65. Formulario de entrada a la aplicación 226
Figura 66. Ventana "Lista de pacientes" 227
Figura 67. Ventana "Paciente" 228
Figura 68. Detalle de la ventana "Paciente" del apartado "Anticonceptivos" 229
Figura 69. Detalle de la ventana "Paciente" del apartado "Fármacos habituales" 229
Figura 70. Detalle de la ventana "Paciente" del apartado "Estados" 230
Figura 71. Detalle de la ventana "Paciente" del apartado "Performance status" 230
Figura 72. Detalle de la ventana "Paciente" del apartado "Antecedentes médicos" 230
Figura 73. Detalle de la ventana "Paciente" del apartado "Hábitos tóxicos" 231
Figura 74. Detalle de la ventana "Paciente" del apartado "Antecedentes oncológicos familiares" 231
Figura 75. Detalle de la ventana "Paciente" del apartado "Episodios" 232
Figura 76. Ventana "Episodio" de tipo "Diagnóstico" 232
Figura 77. Ventana "Lista de tratamientos" 233
Figura 78. Ventana "Tratamiento" 234
Figura 79. Formulario "Prueba" de tipo "Biológicas" y subtipo "Inmunohistoquímica" 236
Figura 80. Ventana "Muestra" de tipo "Tumor" 237
Figura 81. Ventana "Muestra" de tipo "Ganglios" 237
Figura 82. Detalle de la ventana "Muestra" de tipo "Tumor" 238
Figura 83. Ventana "Metástasis" 238
Figura 84. Ventana "Síntoma" 239
Figura 85. Ventana "Filtrado de pacientes" 240
Figura 86. Ventana "Carga de datos de microRNAs" 241
Figura 87. Ventana de "Análisis de expresión de microRNAs" 245

Figura 88. Ventana "Consulta de análisis por microRNAs" 247
Figura 89. Ventana "Consulta de análisis por genes afectos" 249
Figura 90. Ventana "Consulta de análisis por Pathways" 250
Figura 91. Esquema dibujado en la pizarra durante la sesión de *focus-group* para pegar los post-its con los pros, contras y mejoras. 301
Figura 92. Pizarra con post-its de la valoración de la gestión de datos y análisis de pacientes 302
Figura 93. Pizarra con post-its de la valoración del análisis de microARNs 302

Índice de tablas

Tabla 1. Tabla resumen clasificación molecular 33
Tabla 2. Datos incluidos en el Registro Danés del Cáncer 114
Tabla 3. Correspondencias del fichero de datos de expresión de microARNs, en la pestaña "Extracción RNA" con las tablas de la base de datos 243
Tabla 4. Correspondencias del fichero de datos de expresión de microARNs, en la pestaña "datos_chip_entero_normalizado" con las tablas de la base de datos. 243
Tabla 5. Tabla de correspondencias de la base de datos con el Excel de datos de mujeres jóvenes 270
Tabla 6. Tabla de correspondencias de la base de datos con el Excel de datos de mujeres de edad avanzada 277
Tabla 7. Tabla resumen de las preguntas de investigación, hipótesis, variables respuesta y métrica utilizadas en la validación. 283
Tabla 8. Resumen de las tareas y subtareas a llevar a cabo por los sujetos del experimento 284
Tabla 9. Tabla de categorización de las variables clínicas 288
Tabla 10. Tabla resumen con las medias de los tiempos dedicados y porcentajes de acierto en la realización de los ejercicios de la validación. 292
Tabla 11. Tabla con la facilidad de uso percibida por los usuarios en la validación 296
Tabla 12. Tabla con utilidad percibida por los usuarios en la validación. 297
Tabla 13. Tabla con la intención de uso de ambos métodos planteados en la validación 298

1. Introducción

El estudio del Cáncer de Mama, como el de otras enfermedades complejas, es un proceso en el que se requiere la consulta, manejo y relación entre sí de múltiples datos clínicos y genéticos para llegar a conclusiones correctas que permitan el avance de la investigación en el conocimiento de la enfermedad y, por tanto, de manera concreta, permita una mejora del diagnóstico y seleccionar el tratamiento más adecuado a la misma. Este proceso que tradicionalmente se ha realizado a partir de la historia clínica del paciente se complica cuando los datos a estudiar se incrementan, son más complejos y se encuentran dispersos y almacenados de forma heterogénea en distintas fuentes.

En un intento de unificar todos los datos de publicaciones científicas referentes a variaciones genéticas sobre la enfermedad, varias entidades, tanto públicas como privadas, han desarrollado bases de datos accesibles a través de internet donde se recopila información sobre la enfermedad, principalmente sobre variaciones genéticas con un fenotipo posiblemente relacionado. Con demasiada frecuencia se puede constatar que el formato seguido por cada una de estas bases de datos no sigue estándares predeterminados. Como consecuencia de ello, los datos relevantes aparecen dispersos y almacenados de forma heterogénea, y los profesionales clínicos o genetistas que necesitan consultar estos datos deben conocer perfectamente cada una de las fuentes de datos, su formato, y saber hacer

conversiones entre distintos formatos para poder hacer comparaciones entre datos provenientes de cada una de ellas.

Con respecto a los datos manejados en un entorno hospitalario, el problema de dispersión y heterogeneidad de los datos aparece de manera similar, pero a menor escala (más local), dificultando la gestión de datos clínicos y la obtención de conclusiones relevantes a partir de ellos. En la actualidad se recogen una gran cantidad de datos de pacientes provenientes de las diferentes pruebas clínicas diagnósticas y del tratamiento de las distintas enfermedades. El cáncer de mama es un buen ejemplo de ello. En estos casos, las pruebas se van realizando en distintos centros y unidades hospitalarias, generando cada uno de ellos informes clínicos con los correspondientes resultados y provocando que los datos queden almacenados de forma dispersa. En el mejor de los casos, el centro hospitalario dispone de un sistema de Historia Clínica Electrónica, en el que los informes quedan almacenados a disposición de los clínicos del hospital, en formularios con campos de texto libre donde las distintas unidades incluyen los datos en el formato que cada una de ellas considera más apropiado. Cuando un profesional necesita acceder a los datos para realizar un determinado estudio, se encuentra con el inconveniente de tener que recopilar los datos desde los varios informes médicos generados, paciente por paciente, o acudiendo a la unidad clínica responsable, confiando en que se encuentren almacenados de forma ordenada (por ejemplo en ficheros Excel, muy utilizados en el entorno clínico como bases de datos improvisadas) o en pequeñas bases de datos.

Observando la situación detectada, nos encontramos con un problema de gestión avanzada de datos muy complejo y que, como tal, requiere soluciones muy sofisticadas que aprovechen el conocimiento y experiencia acumulado en el dominio del Diseño de Sistemas de Información. Las soluciones que se han planteado a día de hoy, tanto desde el punto de vista de los datos públicos como desde el punto de vista hospitalario, se encentran muy lejanas a este objetivo. El objetivo esencial de esta Tesis Doctoral va justamente en esa dirección.

Planteando soluciones desde la perspectiva de los Sistemas de Información, se detecta la necesidad de un Sistema de Información holístico diseñado de acuerdo con los principios de los sistemas de información y estrictamente basado en la utilización de modelos conceptuales. Actualmente, se ha demostrado en múltiples entornos con una gran manipulación de datos, como las grandes empresas o la banca, que el uso de las tecnologías adecuadas en el ámbito del diseño de sistemas de información complejos es la vía para resolver problemas complejos de gestión de datos y, por tanto, también debe ser eficaz en un entorno de datos tan complejo como el que estamos comentando en el que están implicados tanto datos clínicos como genómicos relacionados con el Cáncer de Mama.

El sistema de información que se plantea en esta Tesis Doctoral tiene en cuenta, desde el diseño hasta la prueba de concepto, tres características que consideramos esenciales: holístico, independiente de la tecnología y con capacidad de evolución. Integra, utilizando la tecnología del Modelado Conceptual, todos los conceptos en un único esquema de datos con diferentes vistas que permitirá tener una visión holística de toda la información manejada en el dominio. Además, este esquema conceptual permitirá implementar el sistema de información utilizando cualquier tecnología y actualizar los datos y la información a medida que vayan evolucionando con el tiempo. Estos tres factores son esenciales para conseguir el sistema de información de calidad que solucionará los problemas planteados y que se desarrolla en esta Tesis.

La estructura de este capítulo de Introducción desarrolla estas ideas en cuatro puntos fundamentales:

- Una exposición de la importancia del problema tratado, incidiendo en su motivación, en su incidencia social y en la importancia de disponer de entornos de gestión de datos que engloben los datos clínicos y genómicos.
- Como consecuencia del problema investigado desarrollado en el punto anterior, se presentan los objetivos generales y específicos de este trabajo de tesis

- La descripción de la metodología de investigación de *Design Science* que se ha seguido en la realización de esta tesis
- La explicación de la estructura de la Tesis Doctoral seleccionada como la más adecuada para documentar el trabajo realizado y justificar la satisfacción de los objetivos propuestos.

1.1. Motivación

1.1.1. El Cáncer de Mama en la sociedad

Cáncer es un término que define aquellas enfermedades en las cuales las células proliferan sin control y adquieren capacidad para invadir otros tejidos. El cáncer no es una enfermedad, sino muchas enfermedades distintas, llegando a abarcar unos cien tipos diferentes, la mayoría de los cuales se nombran por el nombre del órgano o célula de origen. Los cánceres más prevalentes y los más estudiados son el cáncer de colon, el de pulmón y el de mama, representando el 60% del total de los cánceres.

El cáncer de mama es una de las enfermedades más comunes en la población femenina en todo el mundo. Según datos de la Organización Mundial de la Salud (OMS) en su último Informe Mundial de Cáncer publicado en el año 2014 [1] representa el 16,7% de todos los cánceres femeninos. Según este informe en el 2014 murieron 521.817 mujeres en el mundo y se registra como 69% de las defunciones mundiales por esa causa.

Según los estudios más recientes de GLOBOCAN [2] que analizan los datos del año 2012, el cáncer de mama es el segundo cáncer más común en el mundo y el más frecuente entre las mujeres con 1.670.000 de casos nuevos diagnosticados en 2012 (lo que supone un 25% del total de casos de cáncer). Es el cáncer más frecuente en mujeres tanto en los países más desarrollados como en los menos, habiendo ligeramente más casos en los países menos desarrollados (883.000 casos) que en los más desarrollados (794.000 casos). La incidencia tiene una variación mucho más notable entre las diferentes regiones del mundo, con valores como 27 casos por cada 100.000 habitantes en África central y el este de Asia, hasta valores de 92 casos por cada 100.000 habitantes en América del Norte. Además, este tipo de cáncer aparece como el quinto más mortífero (522.000 muertes). A pesar de que es la causa más

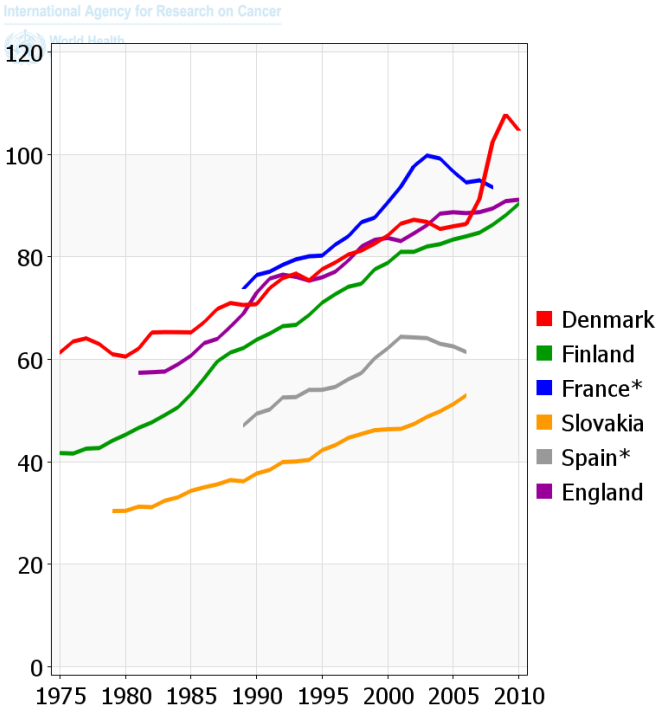

Figura 1. Evolución anual de la incidencia del Cáncer de Mama en mujeres en distintos países. Valores por cada 100.000 habitantes (GLOBOCAN 2012)

frecuente de muerte por cáncer en mujeres en las regiones menos desarrolladas (324.000 muertes, un 14.3% del total), se trata de la segunda causa en las regiones más desarrolladas (198.000 muertes, un 15.4%) después del cáncer de pulmón. El rango en la tasa de mortalidad entre las diferentes regiones del mundo es menor que el de la incidencia, ya que la posibilidad de supervivencia a un cáncer de mama en las regiones más desarrolladas es mayor, con tasas de 6 muertes cada 100.000 habitantes en el este de Asia y de 20 por cada 100.000 en el oeste de África.

Es evidente que estamos ante un problema de alcance internacional. Con los datos disponibles en el informe GLOBOCAN [2], en nuestro contexto geopolítico europeo encontramos que, por ejemplo, en Dinamarca, la proporción es incluso mayor, afectando a una de cada ocho mujeres (como

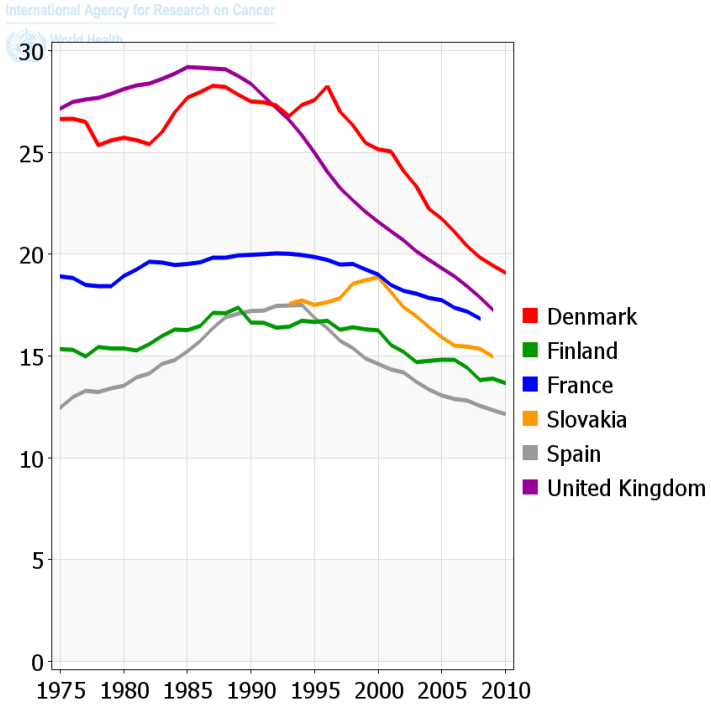

Figura 2. Evolución anual de la mortalidad del Cáncer de Mama en mujeres en distintos países. Valores por cada 100.000 habitantes (GLOBOCAN 2012)

podemos ver en la Figura 1), y siendo además el país con una mortalidad más alta por esta enfermedad (Figura 2). Representa el 20-30% de los cánceres en mujeres y continúa siendo la primera causa de muerte por cáncer en mujeres europeas. Esta incidencia es debida a los estilos de vida sedentarios y los hábitos de vida que caracterizan a las sociedades actuales europeas, y España va de camino a conseguir estos niveles. Además, la edad de máxima incidencia está por encima de los 50 años, pero aproximadamente un 6% se diagnostica en mujeres menores de 35 años. En Europa, el pronóstico es relativamente bueno con una supervivencia a 5 años del 77%.

En España una de cada diez mujeres sufre esta enfermedad al menos una vez en su vida. Según los datos más recientes de GLOBOCAN [2], en 2012 hubo una incidencia de cáncer de mama en mujeres en España de 25.515 casos, lo que supone un ratio de 67,3 por cada 100.000 habitantes. Representando un 29% de los cánceres registrados en mujeres y el 11,7% de los cánceres registrados en población de ambos sexos, el cáncer de mama es el tipo de cáncer con mayor incidencia en las mujeres españolas y el cuarto de mayor incidencia en población general. Sin embargo, la mortalidad fue de 6.075 mujeres, lo que supone un ratio de 11,9 por cada 100.000 habitantes, lo que supone un 15,5% de las muertes por cáncer en mujeres españolas. En términos generales, el cáncer de mama ha sido el tipo de cáncer más frecuente y más mortal en la población femenina de nuestro país en 2012, como se observa en la Figura 3 publicada por la Sociedad Española de Oncología Médica [3].

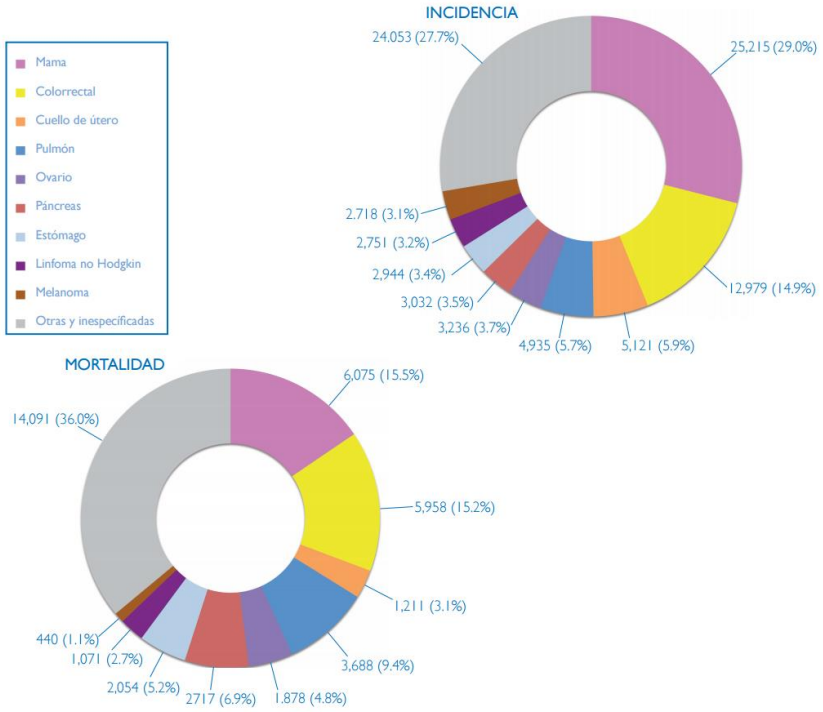

Figura 3. Tasas de incidencia y mortalidad de distintos tipos de cáncer en mujeres en España en 2012 (Sociedad Española de Oncología Médica SEOM. 2014)

La búsqueda de soluciones para reducir su frecuencia y severidad y mejorar la calidad de vida de los pacientes que sufren cáncer de mama es una prioridad del sistema público de salud y uno de los asuntos que más preocupan a la sociedad. Esta preocupación ha llevado a los gobiernos a realizar grandes inversiones en proyectos de investigación médica destinados a estudiar el cáncer de mama y a encontrar nuevos tratamientos, convirtiendo esta enfermedad en una de las más estudiadas. Todos estos esfuerzos han conseguido que el ratio de mortalidad por cáncer de mama se reduzca año tras año.

Esta alta incidencia de la enfermedad en los últimos años y el interés de los sistemas públicos de salud en invertir en el estudio de esta enfermedad hacen que la cantidad de información y datos referentes al cáncer de mama, tanto al tratamiento como a la prevención de la enfermedad, crezca exponencialmente día a día. Nueva información sobre mejoras tanto en los tratamientos como en los procedimientos para incrementar la calidad de vida de los pacientes aparece publicada casi diariamente en publicaciones científicas, quedando almacenada en bases de datos públicas. Además, miles de datos clínicos se generan diariamente en los hospitales y centros médicos durante los procesos de prevención, diagnóstico y tratamiento de pacientes. Estos datos estudiados en conjunto son de gran utilidad para ayudar a mejorar estos procesos en un futuro, reduciendo a su vez la mortalidad, la administración de fármacos inapropiados o mejorando el diagnóstico y la prevención del cáncer de mama.

Con respecto a la etiología del cáncer de mama, todavía no está clara, aunque sí se han identificado ciertos factores de riesgo. La edad es uno de los factores detectados, aumentando el riesgo a desarrollar un cáncer de mama a medida que ésta aumenta. Así pues, la mayoría de los casos se detectan en mujeres mayores de 60 años y se ha visto que éste tipo de cáncer es más frecuente en mujeres de raza caucásica, aunque algunos subtipos más agresivos se presentan en mujeres de raza africana. Debido a la influencia que tienen las hormonas sexuales femeninas en el desarrollo de esta enfermedad, el riesgo a padecerla aumenta con el tiempo que la paciente ha estado expuesta a las mismas, es decir, con una aparición

temprana de la primera regla, con una menopausia tardía, un embarazo tardío o la ausencia de embarazo. Teniendo en cuenta que un 5-7% de los casos de cáncer de mama tienen predisposición genética a la enfermedad, el historial familiar también es un aspecto a valorar, aumentando el riesgo si un familiar de primer grado ha padecido la enfermedad o si la misma paciente ha padecido la enfermedad con anterioridad.

¿Cuál es la conexión de toda esta información con el trabajo de Tesis realizado? En primer lugar, se pretende justificar la decisión de la patología seleccionada en base a su enorme proyección social. En segundo lugar, cabe destacar que el volumen de datos que se generan diariamente no deja de crecer, lo que convierte a la gestión adecuada de toda esa información en la estrategia esencial para encontrar las soluciones buscadas. Sobre esa base, vamos a incidir tanto en la complejidad de la gestión de los datos implicados en el diagnóstico y tratamiento de la enfermedad, como en la importancia de disponer de Sistemas de Información diseñados correctamente y preparados para abordar la gestión de la información implicada de forma eficiente y efectiva.

1.1.2. La genómica y el Cáncer de Mama.

La complejidad a la que nos referimos engloba tanto a la información clínica "convencional" como a los descubrimientos recientes en el campo genómico que están llevando a la llamada Medicina de Precisión a la primera línea de batalla contra la enfermedad [4]. Por ejemplo, BRCA1 y BRCA2 son genes implicados en el crecimiento y división celular. Una mutación en alguno de estos genes interfiere en estas funciones celulares, pudiendo llegar a convertir a la célula en cancerígena. Los individuos con mutación en alguno de estos genes pueden desarrollar cáncer a edades precoces y pueden desarrollar cáncer de mama bilateral.

El tipo de cáncer más frecuente asociado a alteraciones en los genes BRCA1 y BRCA2, es cáncer de mama y ovario en la mujer. Sin embargo, y mucho menos frecuente, una mutación en BRCA2 aumenta la posibilidad de desarrollar cáncer de mama en varón. Es importante recordar que la

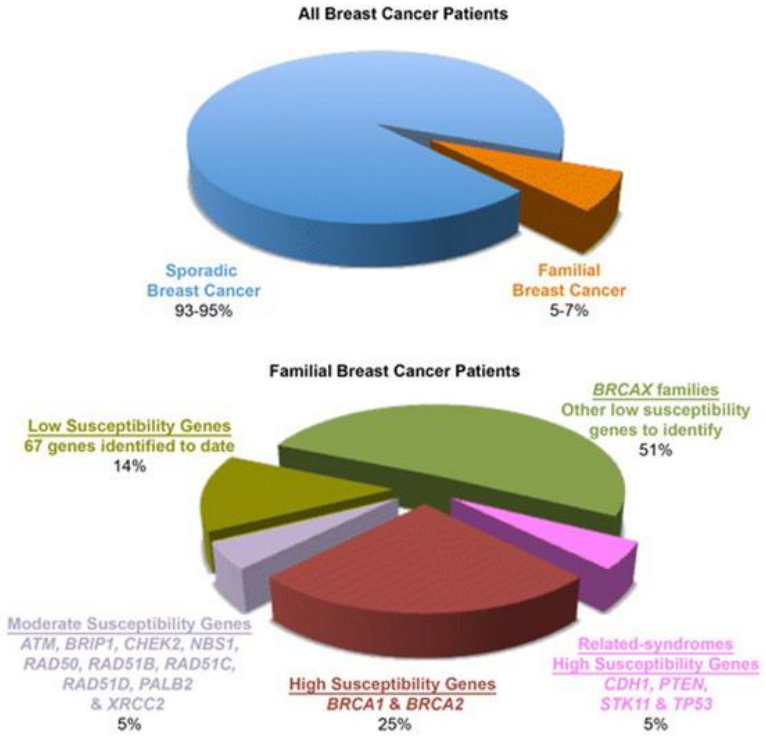

Figura 4. Distribución del cáncer de mama familiar.

mayoría de los cánceres de mama no son hereditarios. En los tumores de mama esporádicos no se conoce el o los genes responsables.

Como podemos ver en la Figura 4, existen otros genes conocidos de alta penetrancia que incrementan el riesgo de cáncer de mama (como TP53, PTEN, SKT11 y CDH1) y genes de penetrancia intermedia (como RAD51C y ATM, BRIP1, CHECK2 y PALB2) donde también se han descrito mutaciones relacionadas con el cáncer de mama [5].

Puesto que sólo entre un 5-7% de los cánceres de mama puede atribuirse a síndromes de predisposición familiar al cáncer que suponen mutaciones de genes únicos con alta penetrancia, probablemente gran parte de la fracción etiológica restante relacionada con una causa genética se deba a polimorfismos de un nucleótido único (SNP) o a una combinación de SNPs. Los polimorfismos más investigados incluyen los genes implicados

en las rutas de síntesis y metabolismo de los estrógenos, la reparación del ácido desoxirribonucleico (ADN) y las principales rutas metabólicas [6, 7].

Prevención

El modelo de predisposición genética que siguen los genes BRCA1 y BRCA2, relacionados con la herencia genética del cáncer de mama, es autosómico dominante de alta penetrancia. Autosómico significa que el gen se encuentra en uno de los 22 pares de cromosomas no sexuales, o autosomas, pudiendo haberlo heredado con igual probabilidad del padre o de la madre y, de la misma manera, pudiendo afectar con igual probabilidad a hijos e hijas. Dominante significa que solamente con heredar una sola copia del gen alterado aumenta la probabilidad de desarrollar la enfermedad. Por lo tanto, la herencia de una única mutación en uno de estos genes conlleva un incremento del riesgo de padecer esta enfermedad (entre el 45 al 85%) a lo largo de su vida. Se han detectado variaciones en estos genes en el 10-15% de las mujeres con un historial familiar de cáncer de mama y en el 60-80% de las mujeres con historial familiar de cáncer de mama y ovario. Todos los individuos tienen dos copias de ambos genes BRCA, una copia procedente de cada progenitor (padre y madre). Si un individuo tiene un antecesor con una mutación en BRCA, él o ella puede heredar el gen mutado (no funcional) o bien heredar la copia no mutada (funcional). Es decir, el hijo o hija de un portador/a de la mutación en el gen BRCA tiene un 50% de posibilidades de heredar gen BRCA mutado e incrementar la probabilidad de desarrollar un cáncer, y un 50% de probabilidad de no heredar la mutación y por tanto tener el mismo riesgo de desarrollar cáncer que la población general.

A pesar de estos datos, un 85-90% de los pacientes con historial familiar de cáncer de mama no tienen mutaciones en estos genes. Un resultado negativo puede significar varias cosas. Primero puede significar que haya una mutación en BRCA que no haya sido detectada debido a una limitación en los test genéticos. También podría significar que la enfermedad puede estar relacionada con mutaciones en otros grupos de genes que actúen en la reparación del ADN en concordancia con patrones de herencia

poligénica. Incluso es posible que el cáncer en esta familia no sea heredado. Debido a que el cáncer de mama es muy común, por casualidad, puede haber varias mujeres de una misma familia afectadas por el cáncer de mama. De todas formas, puede ser que una mutación particular que se está investigando no sea encontrada, pero puede haber mutaciones todavía no conocidas.

Secuenciar el ADN del paciente con cáncer de mama y detectar una alteración, permite analizar el ADN de los familiares para la alteración en concreto. Esto facilitará la predicción del riesgo de cada familiar a padecer la enfermedad, facilitará el consejo genético a toda la familia y facilitará acciones de control y análisis preventivos para detección temprana de la enfermedad. La detección de mutaciones en una edad temprana afecta directamente a la supervivencia de los pacientes afectados y sus familias para establecer políticas preventivas de prevención temprana.

Tratamiento

La secuenciación del ADN, además de ser útil para prevención, lo es también para el diagnóstico y tratamiento. Hasta hace poco tiempo, el tratamiento para el cáncer se basaba principalmente en el lugar del cuerpo donde el tumor comenzaba, por ejemplo, el pulmón o la mama. En la actualidad, este tratamiento depende cada vez más de factores específicos del tumor de un paciente, como las mutaciones genéticas o las proteínas que suelen ser características de las células cancerosas, independientemente de la ubicación original del cáncer. La terapia dirigida es un tratamiento que apunta a los genes o las proteínas específicos de un tumor, o a las condiciones del tejido que contribuyen al crecimiento y la supervivencia del cáncer.

En la actualidad, el 70% de los fármacos que se están desarrollando para el tratamiento del cáncer están relacionados con variaciones genéticas tumorales. Debido a esto, es muy importante identificar el perfil genético del tumor para poder administrarle al paciente el tratamiento más adecuado a las características de su enfermedad. Además, la lista de genes a

analizar en una muestra tumoral es mucho más amplia que la de prevención, lo que complica la realización del estudio genético para la aplicación de un tratamiento. Afortunadamente, hoy en día las tecnologías de secuenciación de alto rendimiento se están popularizando y se está empezando a abaratar costes, lo que conlleva un aumento del número de pacientes secuenciados en paneles de genes candidatos o el exoma completo.

Este tema también es de gran interés para la sanidad pública, ya que la aplicación de la medicina personalizada en el sistema nacional de salud contribuiría a la reducción de los costes totales a medio plazo, debido a que la administración de tratamientos inapropiados utilizando la técnica de prueba y error desaparecería, dando preferencia a la administración del tratamiento correcto en primer lugar con el consecuente ahorro asociado. De todas formas, también es importante tener en cuenta que, aunque los tratamientos personalizados evitan toxicidades innecesarias, su diseño les hace bastante caros, lo que conlleva un aumento de los debates ético-políticos asociados a estas decisiones.

Inicialmente, los tumores se clasificaban por la forma en que las células aparecían en el microscopio y por una clasificación histológica (permitiendo identificar los subtipos intrínsecos (ver capítulo 2.1.2 de esta tesis)). En ocasiones, la detección del cáncer en fase metastásica hace que se desconozca el cáncer primario. Sin embargo, la genómica puede ayudar a mejorar la clasificación de los tumores y determinar el lugar de origen de las metástasis. Por ejemplo, perfiles de expresión - la medición de la actividad de muchos genes - ha ayudado a refinar la clasificación de ciertos tipos de cáncer (como el de mama). Los perfiles de expresión también pueden ayudar a determinar el riesgo de recurrencia en cáncer de mama, de colon y hematológico.

Además, diferencias genómicas en el cáncer desempeñan un papel importante en la explicación de por qué ciertas terapias funcionan para algunos pacientes y no para otros. Por ejemplo, el HER2 (receptor 2 del factor de crecimiento epidérmico humano) es una proteína que participa

en el crecimiento de las células. Está presente en células normales y en la mayoría de los tumores, pero en un 15-20% de los tumores de mama se encuentra en concentraciones elevadas y esto confiere al tumor mayor agresividad. Estos tumores con sobreexpresión de HER-2 son con mucha frecuencia sensibles al tratamiento con el anticuerpo monoclonal trastuzumab. Sin embargo, para las pacientes cuyos tumores no tienen concentraciones elevadas de HER2, estos medicamentos no son beneficiosos. En otras palabras, un tratamiento dirigido no es eficaz si el tumor no tiene la diana. Otro ejemplo es la presencia de proteínas p73 y p63 en pacientes con cáncer de mama triple negativo, lo cual es indicativo de que el medicamento cisplatino será eficaz, ya que descompone la unión de estas proteínas y reactiva la apoptosis (muerte celular). Una prueba genética sería suficiente para saber si las mujeres que padecen el tipo de cáncer de mama con peor pronóstico, el que no responde a los tratamientos antihormonales ni a la última generación de fármacos dirigidos a la proteína HER2, podrían beneficiarse de la terapia con el medicamento cisplatino [8].

El tratamiento del cáncer exige un proceso efectivo y seguro. Sin embargo, los tratamientos actuales no siempre cumplen con las necesidades de los pacientes. En la investigación actual hacia nuevos tratamientos tenemos dos vertientes: una que pretende reducir los efectos secundarios y toxicidad en los tratamientos actuales que funcionan en un 85% de los casos y la otra que se centra en el desarrollo de los nuevos tratamientos para los casos en los que los tratamientos actuales no son efectivos y existen recaídas. La mayoría de los tratamientos convencionales son no específicos. La quimioterapia clásica y la radiación acaban por destruir células normales junto con las células cancerosas, dando lugar a numerosos efectos secundarios. A medida que se acumulan más conocimientos sobre los genomas de los cánceres, los investigadores pueden ser capaces de desarrollar más tratamientos dirigidos a pacientes individuales que minimizan el daño a las células normales. De ello se encargan la farmacogenética y farmacogenómica, las cuales ofrecen mejores resultados clínicos, con una disminución de la morbi-mortalidad de los pacientes y

una importante reducción de la necesidad de administrar distintos y sucesivos medicamentos hasta encontrar el más eficaz y seguro para cada uno, lo que incrementará la calidad de vida de los pacientes. En estos momentos ya hay un grupo de medicamentos - entre los que se encuentran cetuximab, trastuzumab o imatinib - en que la población de pacientes diana que deben ser tratados se selecciona a partir de un test de farmacogenómica predictiva según la determinación genética de una serie de polimorfismos en las biopsias tumorales. De esta manera el tratamiento es más eficaz y seguro en cada paciente, pudiéndose elaborar una quimioterapia combinada e individualizada, lo que podría generar un resultado clínico más satisfactorio y evitar el uso innecesario e inadecuado de medicamentos. Además, con esas medidas se prevendría la aparición de efectos adversos y se ahorraría recursos al reducirse el número de visitas necesarias y el total de fármacos administrados.

Todo lo expuesto en los párrafos anteriores persigue mostrar la gran cantidad de información de procedencia diversa que interviene en el proceso de prevenir, diagnosticar y tratar de forma eficiente y efectiva la enfermedad. No solo de procedencia diversa, sino que también de carácter heterogéneo, lo que complica sobremanera su gestión. Es importante destacar que la integración de la información relevante debe incluir de una forma holística tanto la vertiente clínica que hemos llamado convencional, como la perspectiva genómica. Esta perspectiva está continuamente presente en el desarrollo de este trabajo de investigación.

De hecho, la necesidad evidente de combinar datos genómicos con datos clínicos de los pacientes, tanto en prevención como en tratamiento del cáncer de mama, complica mucho el trabajo de los profesionales de la medicina y la biología intentando obtener conclusiones conjuntas de los estudios realizados de forma independiente sobre los mismos pacientes. Es esencial poder combinar esta información para detectar características biológicas de la enfermedad que nos permitan, por ejemplo, detectar posibles dianas de tratamiento, diseñar fármacos adecuados a un perfil tumoral específico o encontrar subtipos tumorales dentro de un mismo perfil tumoral desconocidos hasta el momento. El manejo de esta compleja

y variada cantidad de información supone un trabajo que requiere mucho tiempo y esfuerzo, y que con un adecuado sistema de información podría llevarse a cabo de una forma mucho más eficiente y rápida.

1.2. Objetivos

El objetivo principal de este trabajo es el diseño de un sistema de gestión de datos clínicos y genómicos para su utilización en el ámbito clínico basado en Modelos Conceptuales.

Para resolver este objetivo se plantean seis subobjetivos:

- Evaluar el campo de trabajo
  - Conocer de forma detallada los procedimientos actuales de diagnóstico y tratamiento del cáncer de mama
  - Estudiar las principales bases de datos clínicas y genómicas consultadas por los profesionales médicos y biólogos que contienen información sobre el Cáncer de Mama
  - Conocer y valorar las técnicas y métodos implementados en la actualidad para almacenar este tipo de información

- Diseñar los Esquemas de Modelado Conceptual que, aplicados al ámbito de la clínica y la bioinformática donde su uso no es tan habitual como debiera ser, permitan representar con precisión el soporte ontológico del dominio analizado.
  - Se incluirán todos los conceptos clínicos, biológicos y genómicos utilizados en el contexto de la prevención y tratamiento del cáncer de mama.

- Estandarizar la información modelada conforme a la norma ISO 13606 de Historia Clínica Electrónica, facilitando la interoperabilidad semántica y el intercambio de información entre sistemas de gestión de datos clínicos.

- Proporcionar un entorno de base de datos, creado a partir del Esquema Conceptual del dominio modelado previamente, donde la información procedente de los entornos clínico, biológico y genómico relacionada con el cáncer de mama pueda almacenarse

de forma holística, integrada y relacionada de forma apropiada.

- Diseñar y desarrollar un prototipo de sistema de información que, tomando como base los Esquemas Conceptuales diseñados permita:
  - Gestionar de forma adecuada los datos clínicos y biológicos
  - Facilitar las tareas de análisis de estos datos
  - Obtener conclusiones relevantes de forma ágil y sencilla.

- Validar en un entorno real la eficacia del sistema de información propuesto en la mejora de la calidad de la gestión de datos en un entorno clínico específico de una enfermedad compleja, como el cáncer de mama.
  - Conocer las ventajas e inconvenientes que supone la utilización de estas herramientas en estos entornos clínicos.

1.3. Metodología de investigación

Para la realización de la investigación de esta tesis se ha seguido la metodología de investigación denominada *Design Science* [9, 10]. Se trata de una metodología que promueve la creación de artefactos que solucionan problemas del entorno y contribuyen a la base del conocimiento actual. Específicamente, la metodología de investigación en la que se ha basado esta tesis es la propuesta de Wieringa definida en [11] en la que plantea la metodología de *Design Science* como un conjunto de ciclos regulativos anidados.

Wieringa propone resolver los problemas de ingeniería e investigación a los que se enfrentan los investigadores descomponiéndolos en subproblemas de ingeniería e investigación, de manera que el problema principal está compuesto por un conjunto de subproblemas anidados. Un problema de ingeniería es la "diferencia entre la forma en la que se percibe el problema por parte de los actores y cómo les gustaría que fuese realmente" y un problema de investigación es "la diferencia entre el conocimiento actual de los actores sobre el mundo y lo que les gustaría conocer".

El planteamiento propuesto para resolver ambos tipos de problemas es seguir un ciclo regulativo basado en cinco fases principales: "Investigación del problema", "Diseño", "Validación", "Implementación" y "Evaluación". Sin embargo, dependiendo del tipo de problema el ciclo regulativo varía ligeramente.

Un ciclo de ingeniería consta de las siguientes fases:

- Investigación del problema
- Diseño de la solución
- Validación de la solución
- Implementación de la solución
- Evaluación de la implementación

En cambio, un ciclo de investigación contiene las fases siguientes:

- Investigación del problema
- Diseño de la investigación
- Validación del diseño
- Ejecución de la investigación
- Análisis de los resultados

Para resolver un problema, la metodología empieza caracterizando el problema general como una investigación o un problema de ingeniería, aplicando el correspondiente ciclo. Si el problema encontrado es un problema de ingeniería, iniciaremos el ciclo por la fase de "Investigación del problema", donde analizaremos detalladamente el problema al que nos vamos a enfrentar.

Una vez caracterizado el problema entraremos en la fase de "Diseño de la solución". Para ello, es necesario investigar previamente el dominio para verificar que no hay una solución existente que resuelva dicho problema. En el caso en el que no se haya encontrado solución posible, podemos proponer una nueva. Como conclusión, se diseña una nueva propuesta de solución con el propósito de resolver el problema planteado.

Cuando el diseño de la propuesta se ha completado es necesario validar la solución antes de su implementación final, entrando en la fase de "Validación de la solución". Por este motivo, las propiedades de la solución son evaluadas de acuerdo a los criterios definidos en la fase de "Investigación del problema", caracterizando el contexto de la aplicación y la cobertura de la solución. Si la solución tiene el efecto deseado para los actores en su contexto, la solución puede ser finalmente implementada, pasando a la fase "Implementación de la solución". La evaluación de la solución implementada se llevará a cabo en la siguiente fase del ciclo, llamada "Evaluación de la solución".

Por otro lado, si el problema planteado es un problema de investigación, entraremos en una primera fase de "Investigación del problema", donde caracterizaremos en detalle el problema a investigar.

Después de haber estudiado el problema, es necesario diseñar cómo se va a llevar a cabo la investigación estableciendo como se van a recoger los datos, el entorno, los instrumentos y los métodos de análisis de los datos, en la fase de "Diseño de la investigación". En la siguiente fase, llamada "Validación del diseño", evaluaremos las amenazas a la validez del experimento. Tras asegurar la validez del diseño de la investigación, esta se puede llevar a cabo entrando en la fase de "Ejecución de la investigación". Finalmente, en la fase "Análisis de los resultados", se analizan los resultados obtenidos de la investigación, extrayéndose las conclusiones pertinentes.

Durante la ejecución de un ciclo, tanto de ingeniería como de investigación, es normal que aparezcan nuevos subproblemas que tengamos que resolver para finalizar el ciclo principal. Para ello, la metodología propone abrir nuevos ciclos y llevar a cabo sus tareas antes de continuar con el ciclo anterior. Finalmente, se completarán las tareas de todos los ciclos y el problema original principal se habrá resuelto.

1.3.1. Aplicación de la metodología de investigación

Teniendo en cuenta los objetivos de esta tesis planteados en el capítulo 1.2 y la metodología de *Design Science* que vamos a seguir para la realización de esta investigación debemos caracterizar el problema principal de esta tesis como un problema de ingeniería, con la consecuente aplicación de un ciclo regulador de ingeniería y la realización de las tareas asociadas al mismo.

Siguiendo la metodología, se ha diseñado el conjunto de ciclos anidados que presentamos en la Figura 5, donde podemos ver un ciclo de ingeniería principal, del cual nace un ciclo anidado de investigación.

Comenzaremos el ciclo principal con la fase de "Investigación del problema", que hace referencia a la motivación de esta tesis, los objetivos

planteados, la descripción de la metodología a seguir y la estructura de la misma, que se incluyen en el capítulo 1 de este documento.

Continuaremos con la fase de "Diseño de la solución", donde se incluye el estudio del estado del arte (capítulo 2) y donde analizamos detalladamente el dominio para conocer en profundidad el problema planteado en la fase anterior y verificar que no haya una solución existente que resuelva dicho problema. Tras realizar la verificación, se plantea el diseño de la solución al problema (capítulo 3) donde incluimos la realización del Esquema Conceptual del Cáncer de Mama, la estandarización de la información utilizando el estándar ISO13606, la implementación y carga de la base de datos y el diseño e implementación del prototipo propuesto.

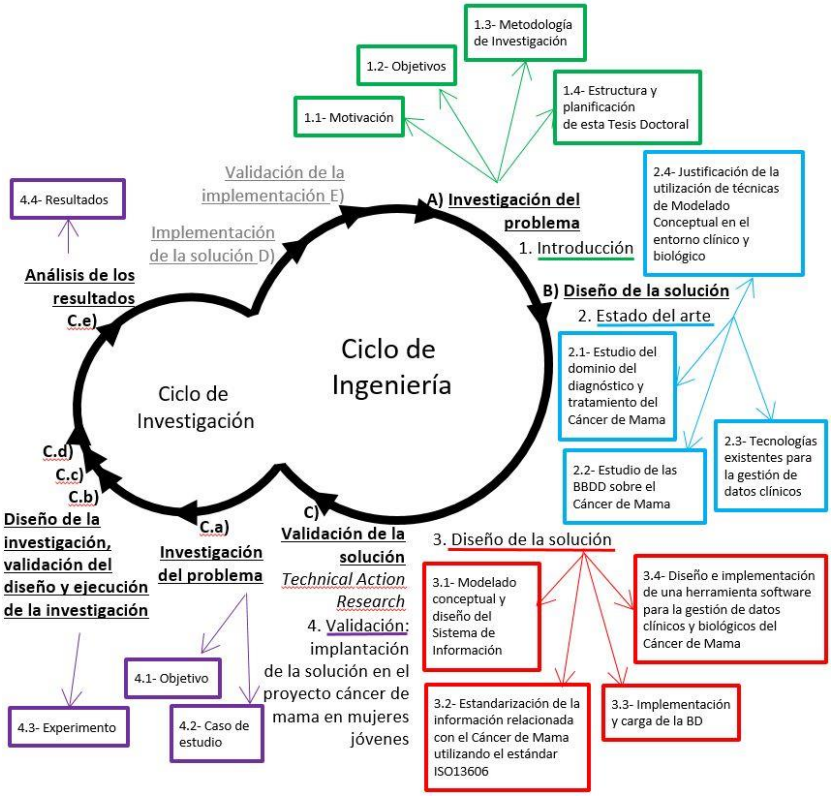

Figura 5. Ciclos regulativos anidados que representan la metodología de investigación seguida en esta tesis.

La siguiente fase a completar es la fase de "Validación de la solución". En este caso, vamos a seguir el método de validación de *Technical Action Research* descrito por Wieringa en [12]. Es en este punto donde se abre un nuevo ciclo de investigación anidado al ciclo de ingeniería principal. Este ciclo corresponde enteramente con el capítulo 4 de esta tesis, donde se describe la validación de la solución diseñada en la fase anterior.

En este nuevo ciclo de investigación, nos encontramos con una primera fase de "Investigación del problema" donde estudiamos quienes son los sujetos van a realizar el experimento, bajo qué entorno, las necesidades o el problema al que se enfrentan y el proceso preparatorio necesario para llevar a cabo el experimento. Esta fase queda descrita en los capítulos 4.1 y 4.2.

Las tres fases siguientes de este ciclo de investigación ("Diseño de la investigación", "Validación del diseño" y "Ejecución de la investigación") las describiremos de forma conjunta en el capítulo 4.3, donde incluiremos detalles sobre las preguntas de investigación e hipótesis planteadas, las variables y métricas, las amenazas a la validez, los instrumentos utilizados y el procedimiento llevado a cabo durante la ejecución del experimento.

Para finalizar este ciclo de investigación, realizaremos la fase de "Análisis de los resultados", donde valoraremos los resultados obtenidos en los ejercicios del experimento, que quedan descritos en el capítulo 4.4.

De vuelta al ciclo principal, y para cerrarlo por completo, faltaría abordar las fases de "Implementación de la solución" y "Validación de la implementación". Estas dos últimas fases del ciclo de ingeniería se corresponden con la transferencia a la industria de la solución propuesta, planteamiento que no se presenta como objetivo en esta tesis.

1.4. Estructura y planificación de esta Tesis Doctoral

Con el fin de cumplir los objetivos planteados en el apartado 1.2, esta tesis se ha estructurado en tres capítulos generales:

- Un estudio profundo del estado del arte centrado en los procesos de diagnóstico y tratamiento del cáncer de mama y las principales bases de datos clínicas y genómicas utilizadas en el estudio de la enfermedad, que se muestra en el capítulo 2. Este capítulo tendrá un doble subobjetivo:
  - Ayudar a conocer la información utilizada en este dominio, las técnicas de almacenamiento y análisis utilizadas en el entorno clínico, así como la información disponible en la red sobre la enfermedad y las tecnologías utilizadas para almacenarla. Esto nos permitirá tener una visión completa de la información manejada en el estudio, prevención, diagnóstico y tratamiento de la enfermedad y la dispersión y heterogeneidad de información presente en este entorno.
  - Como consecuencia de lo anterior, justificar la necesidad de disponer de Sistemas de Información diseñados conforme a las buenas prácticas que marca el Modelado Conceptual, para integrar información tan diversa y compleja desde una perspectiva integral.
- En el capítulo 3, nos centraremos en el diseño de la solución al problema planteado en los capítulos anteriores. Esta solución queda dividida en varios subapartados:
  - El diseño del Esquema Conceptual del Cáncer de Mama, dividido en tres vistas: la perspectiva Clínica, de Expresión Génica y de Secuenciación Masiva, representando los datos clínicos sobre el tratamiento, pruebas y muestras del cáncer de mama, los datos de expresión génica manejados en el estudio enfermedad y los datos obtenidos en los análisis genómicos del

cáncer de mama llevados a cabo utilizando técnicas de NGS (Secuenciación de Nueva Generación, en inglés *Next Generation Sequencing*).

- También se utilizará este Esquema Conceptual como base para la creación de arquetipos que permitan compartir la información modelada según el esquema utilizando el estándar de historia clínica electrónica ISO13606.
- Además, se llevará a cabo la implementación y carga de la base de datos específica para el cáncer de mama, diseñada a partir de los esquemas conceptuales, con datos sobre el cáncer de mama procedente de fuentes de datos públicas, permitiendo integrar en la misma base de datos información diversa relacionada con esta enfermedad.
- El último apartado de este capítulo se centra en el diseño e implementación de un prototipo de herramienta que permita a los clínicos y los biólogos gestionar los datos de sus pacientes de cáncer de mama de forma conjunta relacionados en una misma base de datos, facilitando las tareas de análisis de los mismos y la inferencia de nueva información y conclusiones interesantes para sus estudios.

- El capítulo 4 de esta tesis se centra en la validación del trabajo realizado en un entorno real: el Hospital Clínico de Valencia y el instituto INCLIVA bajo el marco de un proyecto de investigación financiado por el Instituto de Salud Carlos III. Este trabajo nos permitirá probar el sistema de información con datos de pacientes reales, demostrando la validez de la propuesta, la consecución de los objetivos y la satisfacción percibida por los usuarios al utilizar el sistema de información planteado.
- Finalmente, las conclusiones, trabajo futuro y las referencias utilizadas cierran el trabajo de Tesis Doctoral desarrollado.

2. Estado del arte

2.1. Estudio del dominio del diagnóstico y tratamiento del Cáncer de Mama

La detección precoz del cáncer es un tema muy importante para la sanidad pública, ya que con tumores iniciales mejora sensiblemente el tratamiento, y se reducen complicaciones, recaídas y metástasis en las pacientes. Todo esto conduce al ahorro de mucho dinero en hospitalizaciones y en tratamientos agresivos para las pacientes. Para conseguir este propósito, hoy en día se tiende a la detección precoz de la enfermedad mediante cribados poblacionales y a una mejor clasificación molecular tratando de seleccionar el tratamiento más adecuado para las pacientes.

En temas de prevención de cáncer hereditario, cuando se encuentra un caso de antecedentes familiares con cáncer de mama y con una mutación específica detectada se realiza consejo genético, análisis molecular de la mutación familiar concreta y seguimiento a los familiares positivos. Para saber si la paciente ha heredado mutaciones en los genes responsables de la predisposición genética al cáncer, o si todavía no ha desarrollado la enfermedad analizaremos el código genético de una muestra de tejido sano de la paciente. En el caso del cáncer de mama, se buscan mutaciones en los genes BRCA1 y BRCA2 para detectar la predisposición a padecer la enfermedad.

Por otro lado, el diagnóstico de un cáncer de mama puede venir por la detección de un bulto o anomalía en el pecho o por cribado poblacional (mamografías cada dos años a partir de los 45 años). Es en las mamografías donde se detectan precozmente cánceres de mama.

Con una sospecha de cáncer de mama se acude al médico de cabecera o al hospital (en función de la zona o región) donde realizarán otras pruebas y análisis para confirmar el diagnóstico. Además, en el caso en el que se pueda obtener una biopsia de la lesión, ésta será analizada en anatomía patológica.

Existen dos clasificaciones principales para la caracterización del cáncer de mama, según las características histológicas y según el fenotipado molecular, que detallamos en las secciones 2.1.1 y 2.1.2.

Teniendo en cuenta que un objetivo esencial de esta Tesis Doctoral es gestionar de forma eficiente la complejidad de los datos que están asociados al diagnóstico y tratamiento del cáncer de mama, es necesario empezar demostrando que dicha complejidad es real y evidente. Vamos pues a realizar en el inicio de este análisis del "Estado del Arte" una caracterización inicial del entorno analizado, identificando procesos clínicos junto con la información que generan y que debe ser gestionada adecuadamente. Asumiendo el riesgo de entrar en una descripción de carácter clínico altamente especializada, en las próximas secciones se aborda el desafío de entender la complejidad de los datos que han de ser gestionados para que la necesidad de la solución propuesta en esta Tesis emerja de forma lógica y precisa.

2.1.1. Clasificación histológica

La clasificación histológica corresponde a la caracterización morfológica microscópica del tumor. Para ello se consideran elementos de diferenciación celular y grado de diferenciación. Esta clasificación proporciona información como el tejido de procedencia del tumor, su forma de crecimiento y aspecto microscópico, entre otras. La clasificación histológica definida por la OMS la podemos encontrar en el libro [13] , el

cual se encuentra disponible en la página web de la propia organización (http://www.iarc.fr/en/publications /pdfs-online/pat-gen/bb4/). A modo de resumen, de esta clasificación extraemos los tipos histológicos de cáncer de mama más frecuentes en la práctica clínica:

- **Carcinoma ductal invasivo o infiltrante (CDI)**: Representa más del 80% de los casos diagnosticados, las células cancerosas provienen de las células que tapizan el ducto galactóforo. Aunque este carcinoma puede afectar a mujeres de cualquier edad, resulta más común a medida que la mujer envejece. Según la Sociedad Americana del Cáncer, aproximadamente dos tercios de las mujeres que son diagnosticadas con cáncer de mama invasivo tienen 55 años o más.
- **Carcinoma lobulillar invasivo o infiltrante (CLI)**: Es menos frecuente que el carcinoma ductal, pero aun así bastante común, dándose en un 10% de las pacientes de cáncer de mama. Se origina en las células que conforman los lóbulos o acinos glandulares. Aunque, al igual que los CDI, este tipo de carcinoma puede aparecer a cualquier edad, los CLI tienden a aparecer a edades algo más avanzadas que los carcinomas ductales invasivos, alrededor de los 60 años.
- **Carcinoma medular**: Representa cerca del 3 al 5 % de todos los casos de cáncer de mama. Se denomina carcinoma "medular" porque el tumor es una masa suave y pulposa que recuerda al bulbo raquídeo o médula. Es más frecuente en mujeres de 45 a 55 años. El carcinoma medular afecta con más frecuencia a mujeres que tienen una mutación del gen BRCA1. La morfología de sus células es similar a células cancerosas agresivas y muy anómalas, pero no actúan como tales. El carcinoma medular no crece rápidamente y por lo general no se propaga fuera de la mama hacia los ganglios linfáticos.
- **Carcinoma mucinoso**: Representa cerca del 2-3 % de todos los casos de cáncer de mama. En este tipo de cáncer, el tumor se forma a partir de células anómalas embebidas en acumulaciones de

mucina. El carcinoma mucinoso suele afectar a las mujeres postmenopáusicas, con una edad promedio al momento del diagnóstico de 60 años o más. Tiene menos probabilidad de propagarse a los ganglios linfáticos que otros tipos de cáncer de mama y, además, es más fácil de tratar.

- **Carcinoma tubular**: El carcinoma tubular representa menos del 2% de los casos de cáncer de mama. Se trata de un tipo especial de carcinoma de mama con un pronóstico particularmente favorable compuesto por estructuras tubulares bien diferenciadas y con lumina abierta revestida por una sola capa de células epiteliales. Son fácilmente detectables por mamografía, con una frecuencia del 9-19% en los programas de cribado poblacional.
- **Carcinoma de células escamosas**: Con una frecuencia de menos del 1% de todos los carcinomas invasivos mamarios, este carcinoma de mama está compuesto enteramente de células escamosas metaplásicas que pueden ser queratinizantes, no queratinizantes o fusiformes.

2.1.2. Clasificación según el perfil molecular

La clasificación más útil para definir y decidir el tratamiento del cáncer de mama es la clasificación según el perfil molecular. Esta clasificación está realizada utilizando técnicas de inmunohistoquímica y actualmente complementa a la clasificación histológica. Clasifica los tumores en 5 niveles (ver resumen en la Tabla 1) dependiendo del nivel de expresión de 3 genes que codifican para los receptores de estrógeno (RE), progesterona (RP) y HER2, dividiendo el cáncer de mama en lo que se conoce como subtipos intrínsecos:

- **Luminal A**.
  - Es el subtipo más común, representando un 50-60% del total.
  - Se caracteriza por la expresión génica de RE, RP, Bcl-2 (B-cell lymphoma 2), GATA3 y citoqueratina 8/18, entre otros. Presenta una baja expresión de genes relacionados con la

proliferación celular y ausencia de expresión de HER2, así como una baja tasa de proliferación medida por Ki67 y un grado histológico bajo. También son frecuentes las mutaciones de la vía PI3K (49%).

- Las pacientes con este subtipo tumoral tienen un mejor pronóstico; la tasa de recaída es del 27,8% (significativamente menor que la de los otros subtipos) y la supervivencia media tras la recaída es también mayor.
- El patrón de recaída más habitual es en forma de metástasis óseas.
- La respuesta a la quimioterapia es pobre.

- **Luminal B**.
  - Representan entre el 10% y el 20% de todos los cánceres de mama.
  - En comparación con el subtipo Luminal A, poseen un fenotipo más agresivo, mayor grado histológico, mayor índice de proliferación, peor pronóstico y una respuesta intermedia a la quimioterapia.
  - La enfermedad ósea es el lugar de recaída más frecuente (30%).
  - Expresa también RE, pero se evidencia un aumento en la expresión de genes de proliferación, tales como MK167, ciclina D1 y en ocasiones el factor de crecimiento epidérmico (EGFR) y el gen HER2.
  - Las mutaciones de PI3K no son tan frecuentes como en el subgrupo anterior, son de un 32%, y coexisten con pérdidas de pTEN. También es frecuente la hipermetilación del DNA.

- **HER2 positivo**:
  - De un 15 a un 20% de todos los cánceres de mama corresponden a este subtipo molecular.
  - Se caracterizan por una alta expresión del gen HER2, y otros genes asociados con la vía HER2 y/o por la amplificación del gen HER2 situado en el cromosoma 17q12. Presenta también

una sobreexpresión de los genes relacionados con la proliferación celular. Otra característica diferencial de este subtipo es la baja expresión de los genes del subtipo Luminal y del basal.

- Son tumores muy proliferativos, con un alto grado histológico, inestabilidad genómica y más del 40% tienen mutaciones de p53 y un 42% mutaciones de PI3K.
- Tiene una alta quimiosensibilidad
- El pronóstico es malo, aunque en los últimos años el tratamiento anti-HER2 ha mejorado la supervivencia tanto del cáncer de mama metastásico como de la enfermedad localizada.

- **Basal like**:
  - Representa el 10-20% de todos los carcinomas de mama.
  - Se caracteriza por una baja expresión de los genes Luminales, baja expresión del grupo de genes HER2 y alta expresión de los genes de proliferación y del denominado grupo de genes basal. El grupo de genes basal incluye citoqueratinas como la CK5, 6, 14, y la 17; el gen EGFR; c-Kit; Vimentina; P-Cadherina; Fascina y Caveolinas 1 y 2. Poseen también una gran frecuencia de mutaciones de P53 y es frecuente la hipometilación del DNA, la inestabilidad genómica y la aneuploidia,
  - Se caracterizan por su aparición a una edad más temprana, un mayor grado histológico y tamaño tumoral, así como mayor frecuencia de afectación ganglionar.
  - La recaída de estos tumores es, predominantemente, en órganos viscerales.

- **Claudin-low**:
  - Después de la clasificación inicial en los cuatro subtipos anteriores, en el año 2007 se identifica este nuevo subtipo.
  - Se caracteriza genéticamente por la baja expresión de genes relacionados con la adhesión celular como las Claudinas 3, 4 y

7, y la E-Cadherina, y una alta expresión de genes mesenquimales como Vimentina, Snail1, Snail2 y Twist1. Este subtipo tiene una expresión genética muy similar al subtipo Basal-like como una baja expresión de los genes del subtipo Luminal y HER2, pero a diferencia del Basal-like, se sobreexpresan genes relacionados con la respuesta inmune y con la transición epitelio-mesenquimal. Poseen, con mayor frecuencia, mutaciones en los genes BRCA.

- Tienen un peor pronóstico.
- La respuesta a la quimioterapia es mala.

Tipo	RE	RP	HER 2	Otros marcadores	Pronóstico
Luminal A	+	+	-	CK8,CK18,GATA	Bueno
Luminal B	+	+/-	+/-	CK8,CK18,GATA	Intermedio
HER2	-	-	+	HER, TP53	Malo
Basal-like	-	-	-	CK5/14/17, EGFR, c-KIT, CD44, Nestina, Caveolina2, P-Cadherina	Malo
Claudin-low	-	-	-	Claudinas, E-Cadherinas, Ocluidina	Malo

Tabla 1. Tabla resumen clasificación molecular

2.1.3. Técnicas utilizadas actualmente para el diagnóstico del cáncer de mama

Para la realización del diagnóstico inicial del tumor de la paciente se tienen en cuenta un conjunto de datos bastante amplio. Algunos de los datos provienen de análisis genómicos del tumor, como detección de variaciones

o análisis de la sobreexpresión o la inhibición de determinados genes. Esta alteración de la funcionalidad genética se obtiene analizando el ARN (Transcriptómica) o las proteínas (Proteómica). Además, algunas enfermedades se producen por las metilaciones del ADN, las cuales tampoco se detectan analizando las proteínas ni utilizando técnicas de secuenciación. En estos casos se recurre a análisis de Epigenómica, donde se estudia la estructura química que envuelve al ADN y sus reacciones con el medio ambiente y otros estímulos permitiendo o impidiendo la expresión de ciertos genes.

Actualmente se llevan a cabo 3 tipos de pruebas diagnósticas alternativas a la secuenciación:

- **Inmunohistoquímica**: Clásica, la que se utiliza de rutina en los servicios de anatomía patológica de los hospitales. Se usan anticuerpos para identificar ciertos antígenos en una muestra de tejido tumoral, generalmente parafinado. El anticuerpo se une a una sustancia radiactiva o un tinte que hace que los antígenos en el tejido se iluminen al microscopio. Puede observarse tanto sobreexpresión como falta de la misma.
- **Fluorocromo in situ hybridisation** (FISH): Es una técnica molecular que permite localizar un determinado fragmento de ADN y pone de manifiesto la presencia o ausencia de secuencias génicas específicas. Se utiliza para la detección de amplificación de HER2+.
- **Plataformas pronósticas** (Chips de ADN o microarrays): Son herramientas para medir la sobreexpresión de genes de proliferación, resistencia a fármacos o potencial de metastasificación entre otras, y relacionarlo con el pronóstico de supervivencia.

Utilización de plataformas pronósticas (microarrays)

Las plataformas pronósticas analizan la actividad de un grupo de genes para predecir el comportamiento del tumor de la paciente y ayudar a definir qué tipo de tratamiento es el más adecuado. Estas plataformas ajustan mejor el riesgo y el tratamiento de los resultados clásicos. Sin embargo, no están indicadas para todos los tipos de pacientes y tumores, por lo que se deben cumplir unos criterios muy estrictos para poder utilizar esta herramienta de análisis.

Las dos plataformas más utilizadas hoy en día son *Oncotype* [14] y *Mammaprint* [15]. La primera, fabricada por *Genomic Health*, es un test que, analizando 21 genes, permite personalizar la terapia adyuvante para pacientes con cáncer de mama por medio de un *Recurrence Score* que cuantifica la probabilidad de recidiva (o reaparición) del tumor a diez años y la probabilidad de que la paciente se beneficie del tratamiento con quimioterapia. *Mammaprint* es un dispositivo de la compañía *Agendia* que, gracias al análisis de 70 genes, clasifica los pacientes con cáncer de mama en 2 grupos (mal y buen pronóstico) según el riesgo de desarrollar metástasis y estima el beneficio de la quimioterapia en cada paciente.

La utilización de una plataforma u otra puede depender de varios factores. Por ejemplo, la plataforma Oncotype únicamente sirve para detectar el riesgo en cáncer de mama con los receptores hormonales positivos. Ofrece el resultado en tres niveles de riesgo (bajo, medio y alto). Cuando el nivel es bajo o medio no hay necesidad de quimioterapia, ya que como mucho se podría obtener un beneficio del 10%. Cuando el riesgo es alto el beneficio que se puede obtener con la quimioterapia es mayor.

Cuando nos encontramos con cáncer de mama de tipo triple negativo (como los tipos Basal-like y Claudin-low) no es recomendable el uso de las plataformas, ya que el pronóstico que ofrecen no es bueno.

2.1.4. Guías de tratamiento

En ellas se detallan los protocolos de actuación para la detección y el tratamiento asociado a cada tipo de cáncer.

Es necesario crear y actualizar este tipo de guías clínicas a nivel nacional o incluso regional para establecer protocolos de actuación a la hora de usar ciertas medicaciones en un cáncer que haya sido identificado a nivel molecular.

Las más utilizadas son:

- **NICE**: Sistema inglés de aprobación de fármacos de forma local (cada hospital tendrá una guía diferente). En ella se catalogan los medicamentos por eficiencia y ayudan a realizar el estudio coste/eficiencia de un fármaco.
- **NCCN**: Guías de manejo del paciente oncológico. En ella se muestra paso por paso los distintos análisis que deben realizarse y el tratamiento a aplicar según los resultados.

2.1.5. Procedimiento clínico para el diagnóstico y tratamiento del cáncer de mama

El proceso diagnóstico y tratamiento de un cáncer es un proceso colectivo en el que intervienen diversos especialistas. Para explicar este procedimiento clínico más detalladamente vamos a dividirlo en varias etapas:

1ª Etapa: Sospecha Inicial Radiológica

Esta primera etapa está representada gráficamente en la Figura 6. El diagnóstico se inicia cuando la paciente va al Screening (**Radiología**) o a la consulta por que se ha notado un bulto (**Cirugía, Ginecología**, **Medicina**

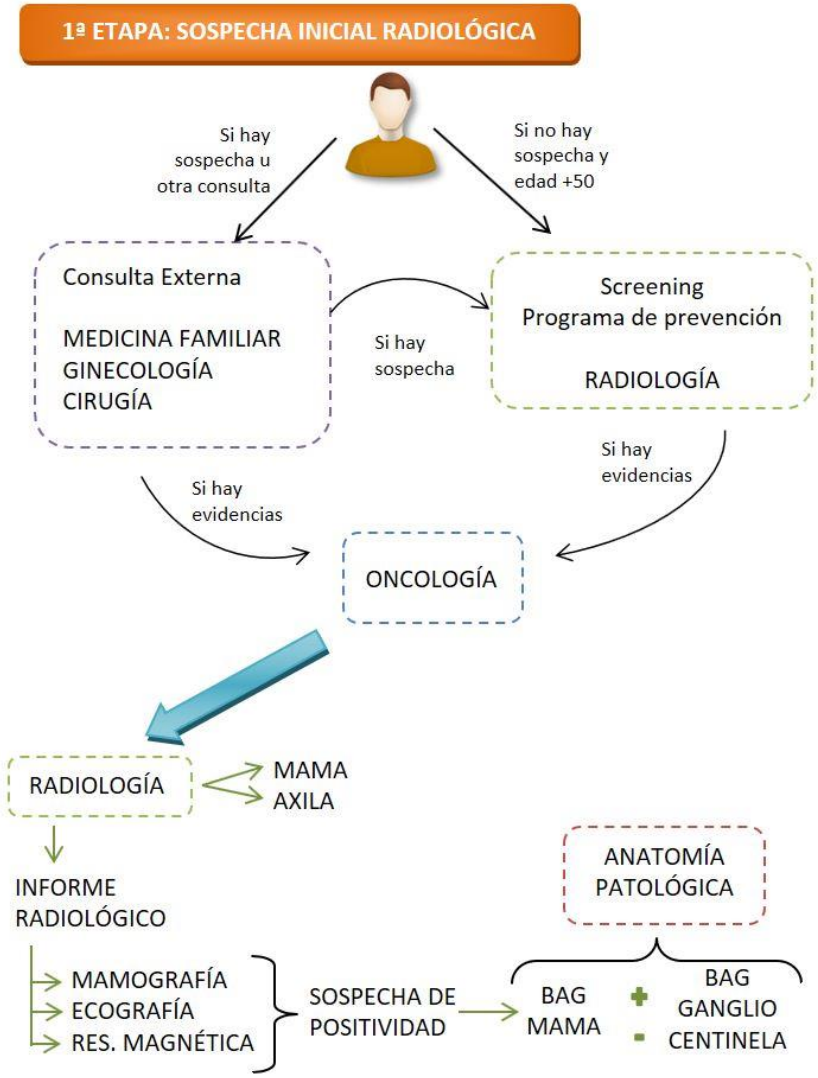

Figura 6. Procedimiento de diagnóstico y tratamiento del cáncer de mama. Etapa 1: Sospecha inicial radiológica

Familiar u Oncología, si hay sospechas desde Medicina Familiar). La lesión detectada se envía a Radiología.

En este momento, la paciente inicia un proceso en **Radiología** donde le hacen una mamografía, ecografía y, si está indicado, resonancia magnética. Si las técnicas de imagen dan una sospecha de positividad desde Radiología se hace una biopsia de tipo BAG (Biopsia Aguja Gruesa) de la mama y del ganglio axilar centinela (GC) si existe alguno que el radiólogo considere sospechoso. Las biopsias se remiten a **Anatomía Patológica**.

2ª Etapa: Estudio de la muestra

En caso de que se confirme malignidad en la muestra y exista indicación para ganglio centinela, la paciente es citada en el servicio de **Medicina Nuclear**, como podemos ver en la Figura 7. Para detectar cuál es el ganglio centinela que hay que biopsiar, en Medicina Nuclear marcan con isótopos radioactivos el ganglio centinela y lo detectan por gammagrafía para que el cirujano sepa cuál es.

A partir de entonces, se realiza el análisis de la muestra y la valoración del caso. Cuando se recibe el tejido biopsiado en **Anatomía Patológica**, comienza el procesado del mismo que dura aproximadamente 12 horas. A continuación, el patólogo elabora un informe patológico que envía al

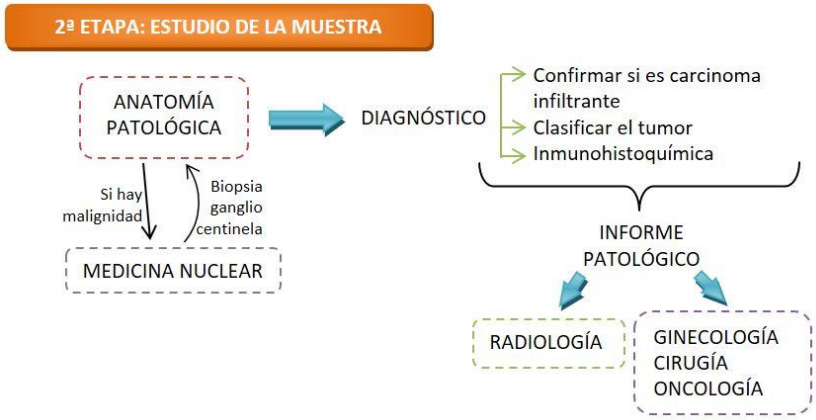

Figura 7. Procedimiento de diagnóstico y tratamiento del cáncer de mama. Etapa 2: Estudio de la muestra

servicio peticionario (**Ginecología, Cirugía** u **Oncología** (si se ha visitado previamente), y se puede mandar también a **Radiología**) donde se indica:

- Confirmación de si es o no un carcinoma infiltrante.
- Clasificación histológica del tumor.
- Inmunohistoquímica (RE, RP, Ki67, Her2)

3ª Etapa: Comité de Tumores

Una vez analizado el caso, si éste es benigno y no requiere otra intervención hospitalaria, vuelve al circuito de cribado. En caso contrario, se inicia el proceso representado en la Figura 9, donde se discute el caso en la Unidad de Mama y se toma la decisión terapéutica más adecuada para la paciente.

La Unidad de Mama consiste en un Comité de Tumores especializado en Mama formado por profesionales de todos los departamentos involucrados (ver Figura 8) que se encargará de tomar las decisiones importantes, supervisar el proceso y asegurarse de que el protocolo de actuación se cumple correctamente. El comité debe valorar el caso y determinar si están

Figura 8. Componentes del Comité de tumores

todas las pruebas, realizar estudios de extensión, decidir el tratamiento sistémico que se le debe aplicar al paciente y tomar decisiones como si se deben realizar nuevas pruebas, acelerar procesos diagnósticos, definir tratamientos, ...

En función de los resultados de la inmunohistoquímica se estudian los factores pronósticos y predictivos para poder tomar mejores decisiones. En algunos casos, si se requiere, se solicitan más pruebas moleculares (MammaPrint u Oncotype).

El Comité de Tumores realiza su actividad en dos escenarios distintos que transcurren de forma secuencial, relacionados con el tratamiento neoadyuvante (el que se le aplica a la paciente de forma previa a la extirpación del tumor) y con la intervención quirúrgica. Pasamos ahora a describir con más detalle su estructura:

Escenario A: Neoadyuvancia

La paciente se dirige a **Oncología** para que se le aplique un tratamiento neoadyuvante (tratamiento quimioterápico previo a la cirugía). En función del subtipo inmunohistoquímico se decidirá el mejor tratamiento quimioterápico y si es necesario asociar un tratamiento biológico. Habitualmente este proceso tiene una duración de 6 meses. Puede aplicarse una de estas opciones o varias combinadas:

- **Quimioterapia** (basada en taxanos y antraciclinas)
- **Hormonoterapia**: (únicamente en pacientes frágiles y ancianas)
- **Fármacos dirigidos a dianas terapéuticas**: Normalmente son proteínas celulares mutadas, como la producida por el gen HER2/une mutado, cuya proteína HER2 se trata con Trastuzumab o Lapatinib.

Una de las finalidades del tratamiento neoadyuvante es conseguir un aumento en el número de cirugías conservadoras (entre un 10 o 30 % de las cirugías que se habían previsto como mastectomías pasan a ser

conservadoras tras la neoadyuvancia). Posteriormente se pasa a Cirugía iniciándose el Escenario B.

Escenario B: Cirugía

Si la BAG realizada previamente ha sido negativa para ganglio centinela o si se ha enviado al paciente directamente a cirugía (sin pasar por el escenario A) se avisa también a **Medicina Nuclear** para que seleccione cual es el ganglio centinela para que el cirujano pueda extirparlo y sea analizado por **Anatomía Patológica**, realizando así un diagnóstico intraoperatorio del mismo.

A partir de este momento, la paciente se pone en manos del cirujano, que puede proceder de las siguientes maneras:

- Si inicialmente el ganglio es positivo para metástasis por carcinoma de mama (bien por BAG o por BSGC (biopsia selectiva de ganglio centinela), se realiza Linfadenectomía axilar conjuntamente con Mastectomía o Cuadrantectomía.
- Si la BAG del ganglio fue negativa o no se hizo biopsia ganglionar, se extrae el ganglio centinela para su estudio intraoperatorio por parte del servicio de Patología. En el mismo acto quirúrgico, se realiza la cirugía sobre la mama (Mastectomía o Cuadrantectomía) Además, si el diagnóstico del ganglio centinela es positivo se hace también una linfadenectomía axilar.

Si el ganglio centinela es positivo siempre se realizará la extracción de la cadena ganglionar para intentar evitar posibles metástasis. Si es negativo indica que el resto de ganglios también están sanos y, por tanto, no se extirpan.

Todas las piezas extraídas por Cirugía se envían al servicio de **Anatomía Patológica** donde se estudian igual que se hizo con la muestra extraída por BAG, pero con el tumor completo donde se estudia mucho más profunda y detenidamente la lesión y el ganglio o ganglios extraídos. Como consecuencia de este estudio se genera un nuevo informe patológico

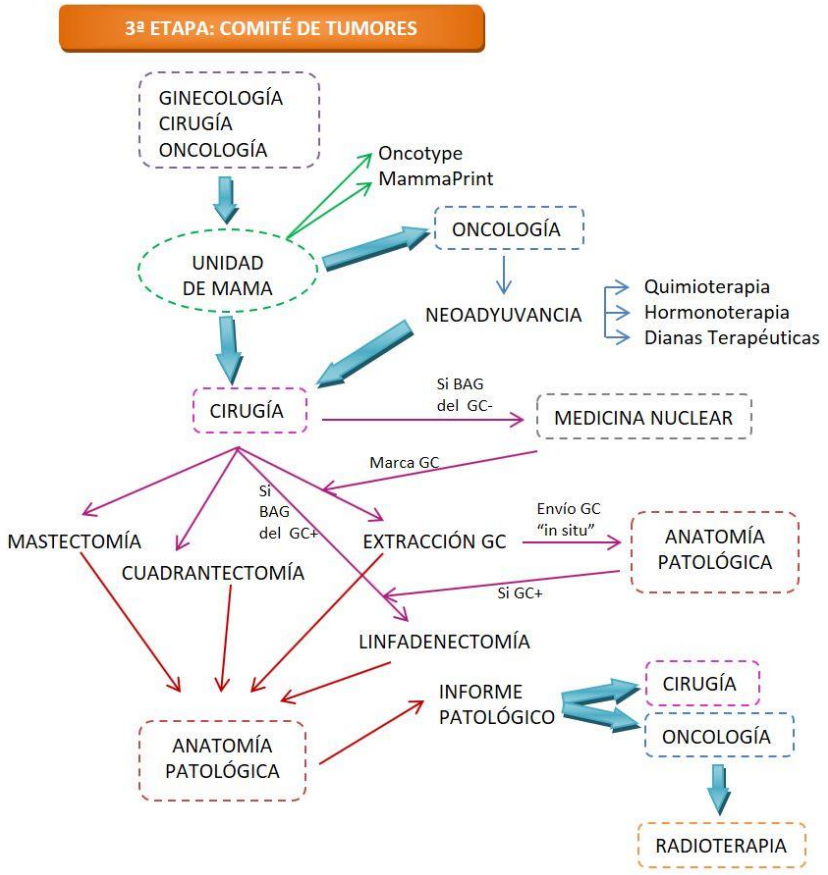

Figura 9. Procedimiento de diagnóstico y tratamiento del cáncer de mama. Etapa 3: Comité de tumores

definitivo, ya que el estudio ha sido más extenso y preciso. En este informe se incluye la inmunohistoquímica. Este informe se envía a **Oncología** y **Cirugía**. El oncólogo con este informe ya puede decidir qué tratamiento debe aplicar al paciente.

A continuación, será el oncólogo el que realice el seguimiento del paciente y el que decida el tratamiento a aplicar -si necesita tratamiento-. Será el oncólogo el que derive al servicio de **Radioterapia** si la cirugía ha sido conservadora o cumple criterios para radioterapia (cirugía conservadora, ganglios axilares afectos, ...).

Si la cirugía fue conservadora y, posteriormente, se diagnostican bordes afectos en la pieza, se discute en la Unidad de Mama para indicación de reintervención o radioterapia. Si se realiza reintervención, las nuevas muestras se envían de nuevo a **Anatomía Patológica** para asegurar la extirpación completa de la lesión.

En los casos en los que, después de todo este proceso, se produzca metástasis o un nuevo tumor, el paciente puede volver a pasar por el mismo proceso. Si hay metástasis pueden llegar muestras de otros órganos que se consideran metástasis del tumor primario de mama.

4ª Etapa: Tratamiento

En el departamento de **Oncología** se encargan de aplicarle el tratamiento farmacológico necesario para superar la enfermedad, como podemos ver en la Figura 10. Dependiendo de las características del tumor se le aplicará uno de los siguientes tratamientos o varios de ellos combinados:

- **Quimioterapia**: Tratamiento con fármacos que se encargan de destruir las células cancerosas. Es muy agresivo para el paciente.
- **Hormonoterapia**: Tratamiento hormonal.
- **Dianas terapéuticas**: Tratamiento dirigido a una determinada proteína.

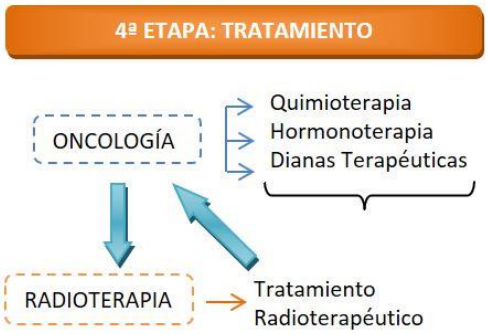

Figura 10. Procedimiento de diagnóstico y tratamiento del cáncer de mama. Etapa 4: Tratamiento

Además, dependiendo de las características del tumor el oncólogo puede considerar necesario enviar al paciente al departamento de **Radioterapia**. En este departamento se le aplica un tratamiento radioterapéutico dirigido a la zona donde ocurrió el tumor. Cuando se ha realizado cirugía conservadora, siempre hay que enviar al paciente a Radioterapia, además de aplicarle el tratamiento farmacológico correspondiente al paciente. Sin embargo, cuando hay extirpación total de la mama, normalmente la **Radioterapia** no es necesaria, aunque puede aplicarse si hay bordes afectos o en la zona ganglionar axilar si está afecta.

5ª Etapa: Seguimiento de la paciente

La paciente tendrá un seguimiento de la enfermedad, que se representa gráficamente en la Figura 11, y que se lleva a cabo desde **Oncología** durante un tiempo mínimo de 10 años. Al principio, la paciente acude a la consulta cada 3 meses durante los 2 primeros años, luego cada 6 meses y a partir de

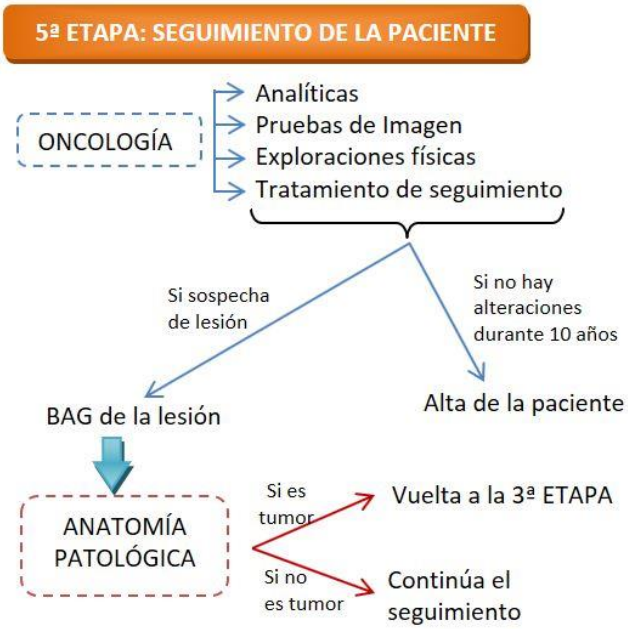

Figura 11. Procedimiento de diagnóstico y tratamiento del cáncer de mama. Etapa 5: Seguimiento de la paciente

los 5 años anualmente. En estas consultas se realizan: analíticas, pruebas de imagen, exploraciones físicas y el seguimiento del tratamiento pautado.

Siempre que haya algún elemento susceptible de ser tumoral se biopsia y se envía a **Anatomía Patológica**. Si se detecta un tumor vuelve de nuevo al circuito. Si en esos 10 años no hay recaídas por la enfermedad, desde Oncología se le da el alta al paciente y se cierra el caso. Sin embargo, puede haber seguimiento si se considera necesario.

2.1.6. Almacenamiento y gestión de los datos en los hospitales y los laboratorios genéticos.

En la mayoría de los hospitales con servicio de Oncología se recogen grandes cantidades de datos de los pacientes que se almacenan de maneras muy diversas. Hace algunos años (muchos menos de los que nos imaginamos), el almacenamiento y gestión de estos datos se limitaba a informes en papel que se incluían en el Historial Clínico del paciente. Cuando hablamos de Historial Clínico nos referimos a un conjunto de carpetas físicas (como la que podemos ver en la Figura 12) donde se incluyen los informes médicos del paciente en formato papel y se gestionan en habitaciones que hacen la función de almacenes de datos.

En la actualidad, y tras la llegada de los avances en informática a nivel usuario, el procedimiento de generación y almacenaje de informes médicos

Figura 12. Carpetas que forman el Historial Clínico de un paciente en formato papel.

se centra, principalmente, en plantillas realizadas por el mismo médico en ficheros de Microsoft Word. Esta herramienta es ampliamente conocida por la sociedad actual y les permite seleccionar por ellos mismos los datos que desean almacenar y organizarlos de la manera que a ellos les parece más útil y estructurada. Estas plantillas contienen un texto genérico para todos los pacientes redactado previamente en el cual se dejan libres ciertos huecos para ir completando con los datos del paciente. De esta manera, el médico puede recoger los datos en la consulta siguiendo la plantilla, asegurándose de que no se olvida de incluir ningún dato y ahorrándose el tiempo de escribir el texto común para todos los pacientes. Además, puede almacenar esta información en sus propios dispositivos, asegurándose que sea accesible para sus investigaciones.

Estas plantillas se utilizan para almacenar información del paciente en el momento del diagnóstico, completándose después con los procedimientos o resultados pertinentes a medida que el paciente evoluciona. La información que se incluye en estos informes se puede resumir en los siguientes puntos:

- Información propia del paciente como el nombre, NHC (número de historia clínica), número de SIP (Sistema de Información Poblacional, corresponde al número de la tarjeta sanitaria), fecha de nacimiento y de la primera visita y un resumen del primer diagnóstico.
- Información sobre los antecedentes médicos del paciente.
- Historia oncológica familiar.
- Historia oncológica propia.
- Pruebas solicitadas y los resultados obtenidos de cada una de ellas.
- Evolución y pruebas o tratamientos realizados al paciente organizados por fechas.
- Plan de actuación previsto.

En la Figura 13 se incluye un modelo de plantilla de informe, así como una muestra de una plantilla con información completa sobre una paciente en

la Figura 14, cedidas como ejemplo en la Unidad de Hematología y Oncología del Hospital Clínico de Valencia.

INFORME DE
NHC
NSIP
Fecha de nacimiento
Fecha de primera visita
TELÉFONO:

Paciente de años remitida desde ante el diagnóstico de carcinoma de mama para valorar tratamiento

Antecedentes médicos: No RAMs conocidas. No hábitos tóxicos. No HTA, no DM, no DL. Menarquia años. FUR . GAP .Lactancia . EPE:. Toma de AO durante meses. THS durante…
Medicación habitual.
Historia Oncológica familiar:

Historia Oncológica: La paciente consulta por …..y en las exploraciones realizadas se halla:
- Mamografía ( ):
-ECOGRAFIA mamaria ( ):
-BAG ( ) :
Inmunohistoquimia: RE % (HScore /300, Allred /8), RPg % (HScore /300, Allred /8), HER 2 negativo (-), Ki 67 %.

El día 10/07/13 se realiza tumorectomía de ……con técnica del ganglio centinela. Siendo el informe AP de …….
Inmunohistoquimia: RE positivos > 50%, RPg positivo 10-50, cerb2 negativo (+/+++), Ki 67 30-40%.
A la exploración física:.

JDX: CDI pT pN M
PESO
TALLA cm

PLAN: PLAN: Se solicita estudio de extensión completo(TAC t-a-p, Rastreo óseo, FEVI, ECG y analítica completa). En función del resultado se valorará su mejor opción terapeútica.

Dr
Servicio de Hematología y Oncología Médica
Hospital Clínico Universitario de Valencia.

Figura 13. Plantilla utilizada en la Unidad de Hematología y Oncología del Hospital Clínico de Valencia para almacenar los datos de las pacientes de Cáncer de Mama

INFORME DE **********
NHC ******** **NSIP **********
Fecha de nacimiento ******** **Fecha de primera visita **********
TELÉFONO: ********

Paciente de 49 años con el diagnóstico de cáncer de mama metastásico que acude para segunda opinión.

Antecedentes médicos: No RAMs conocidas. No hábitos tóxicos. No HTA, no DM, no DL. Menarquia 15 años. FUR 45 años . G3A0P3 .Lactancia si(17 meses acumulada). EPE:. Toma de AO durante meses. THS durante… Colocación de prótesis mamarias en 2008, con recambio de la izquierda en 2010 por encapsulamiento.

Medicación habitual.

Historia Oncológica familiar:

Historia Oncológica: La paciente acude a su revisión habitual y en la RMN mamaria se objetiva en MD, a nivel de la región areolar, una zona nodular de 7 x 6 mm. Ante dicho hallazgo, la paciente consulta el 15 de febrero de 2013 en el Centro de Patología de la Mama de Madrid, y se realizan las siguientes exploraciones complementarias:

-ECOGRAFIA mamaria dopper (15/03/13): Hipervascularización en MD compatible con neovascularización.

-ECOGRAFIA MAMARIA(15/03/13): En MD a nivel de la región areolar en su parte interna, zona de mayor densidad con especulaciones y microcalcificaciones en su interior de 7 x 6 mm.

-RMN mamaria(15/03/13): Imagen nodular en MD de 8 mm en la región retroareolar interna. No adenopatías axilares significativas. No otros hallazgos en MD ni MI.

-BAG (18/03/13) :CDI , GH II. Inmunohistoquimia: RE 90% , RPg 40%, c-erb 2 ++ , Ki 67 30%.

El 28 de febrero de 2013 se realiza tumorectomía con estudio de ganglio centinela, siendo éste positivo, por lo que se realiza vaciamiento axilar. El informa AP confirma la existencia de un CDI, GH II, de 9 x 8 mm, con 2 adenopatías afectas de 22 extirpadas. Inmunohistoquimia: RE 90% , RPg 30%, c-erb 2 ++(FISH AMPLIFICADO) , Ki 67 30%.

Se solicita estudio de extensión: -Rastreo óseo (28/02/13) : Negativo

Figura 14. Detalle de un informe de diagnóstico realizado a partir de la plantilla en la Unidad de Hematología y Oncología del Hospital Clínico de Valencia.

Por otra parte, desde hace algunos años se han ido implantando en los hospitales sistemas de información clínicos para la generación, gestión y almacenamiento de informes clínicos de todas las especialidades y servicios. Estas herramientas son genéricas para todo el hospital, sin distinción de especialidades, e incluyen formularios de informes que los médicos deben rellenar en cada una de sus consultas. Estos formularios genéricos están compuestos principalmente por varios campos de texto libre donde el médico puede escribir las mismas anotaciones que haría en un formato en papel, pero en este caso en formato electrónico. Estos campos vienen coronados por una cabecera donde se indica la información que se tiene que incluir en la celda que aparece debajo de la misma. Los contenidos de estas celdas han de escribirse a mano paciente por paciente y, teniendo en cuenta que el formulario es el mismo para todas las especialidades y servicios del hospital, los campos son bastante inespecíficos. Si tomamos como ejemplo la herramienta de Orión Clinic [16, 17], encontraremos la siguiente información:

- Datos personales del paciente, entre los que se incluyen el nombre del paciente, NHC, número SIP, episodio en el que se encuentra el paciente y la fecha de la cita donde se genera el informe.
- Descripción del seguimiento actual.
- Descripción de las exploraciones complementarias realizadas.
- Juicio diagnóstico.
- Procedimientos a aplicar.
- Tratamiento a aplicar.
- Plan de actuación previsto.
- Seguimientos previos del paciente.

Teniendo en cuenta el uso de estos métodos descriptivos para el almacenamiento de los datos clínicos, realizar estudios y/o análisis sobre ellos conlleva la extracción de los datos relevantes de forma manual de estos soportes. Para almacenar los datos de una forma más estructurada,

	A	B	C	D	E	F	G	H	I
1	ID	Edad al Dx	DCO	Tipo	Grado	Tamaño (cm	pT	N	pN
2	C9	65	CDI	LUMINAL B	III	2.1	c T2(2,1 cm)	NEG	NEG: c N0
3	C10	65	CDI patron m	HER2+/LUMI	III	3	c T2 (3cm)	POS	POS: c N+
4	C11	59	CDI	LUMINAL A	I	1,2	p T1c	NEG	NEG: p N0/1
5	C14	64	CDI	LUMINAL B	II	BAG		NEG	NEG
6	C16/10COP	58	CDI	LUMINAL A	I	0,9	p T1b	NEG	NEG: p N0/1
7	C17	64	CDI	LUMINAL B	III	2,1	p T2	POS	POS: p N5/21
8	C18	63	CDI(de tipo t	LUMINAL A	I	0,9	p T1b	NEG	NEG: p N0/1
9	C21	54	CDI	LUMINAL A	I	2	c T2(2 cm)	NEG	NEG: c N0
10	C22	61	CDI	LUMINAL A	I	1.1	p T1c	NEG	NEG: p N0/2
11	C23	53	CTI	LUMINAL A	I	1.2	p T1c	NEG	NEG: p N0/1
12	C24	64	CDI	TRIPLE NEGA	III	0.9	p T1b	NEG	NEG: p N0/1
13	C26	64	C micropapil	HER2	II	8	c T3(8cm)	NEG	NEG: c N+
14	1DP		CDI	LUMINAL A	III	2.3	p T2	NEG	NEG (0/16)
15	10DP		CDI	TRIPLE NEGA	II	1.4	T1c	NEG	NEG: p N0/1
16	3COP		CDI	LUMINAL B	II	1.5	pT1	POS	POS (3+/25)
17	5COP		CLI	LUMINAL A	II	1.3	p T1	NEG	NEG (0/4)
18	9COP		CDI	LUMINAL A	I	1.5	Pt1	NEG	NEG(0/1)
19	14COP			TN					
20	EOBC-C28		MIXTO:LOBU	LUMINAL B	II	1.2	p Tm	POS	POS (1mic/6
21	C29	66	CDI	LUMINAL B	II	1,3	p T1	NEG	NEG: N0/16
22	C30	69	CDI mutifoca	LUMINAL B	II	1.7	p Tx p N1/2	POS	N1/2
23	C31	76	CDI	HER2 +	II	1,1	p T1b	NEG	N0/19
24	C32	67	CDI	LUMINAL B	II	1,5	nsayo Fase 0	NEG	N0
25	C33	66	CDI	LUMINAL B	I	0,9	pT1c	NEG	N0/1
26	C34	74	2 tipos:CLI, C	TN	III	1.3	mp T1c	NEG	N0/2
27	C36	79	C.Mucinoso	LUMINAL B	II	3	Pt2	NEG	N 0/12
28	C37	78	CI sin tipo es	LUMINAL A	I	1,2	pT1c		No vaciamient
29	C38	69	CDI con difer	LUMINAL A	II	1,5	pT1c	POS	N 1micr/13
30	C39	69	CDI	LUMINAL B	II	0.9	p T1b	NEG	NO/2
31	C40	68	CI sin tipo es	LUMINAL A	II	1,6	pT1c	NEG	N0/22
32	C41/C4	69	CDI	LUMINAL B	III	1,4	T1c	NEG	N0(sn)
33	C42	68	CDI	LUMINAL A					

Figura 15. Ejemplo de la hoja de cálculo de Microsoft Excel utilizada para almacenar datos clínicos para investigación.

actualmente se utilizan hojas de cálculo de Microsoft Excel donde los clínicos almacenan los datos del paciente, permitiéndoles de esta manera visualizar todos los pacientes juntos en una misma hoja y tenerlos mucho más accesibles para su posterior utilización. Un ejemplo de la hoja de cálculo donde se almacenan los datos clínicos para su utilización en estudios de investigación es la que aparece en la Figura 15.

Por otro lado, si echamos un vistazo a los laboratorios de biología molecular donde se almacenan datos genéticos veremos una gran cantidad de datos almacenados utilizando herramientas que no se diseñaron para tal fin, pero que gracias a su facilidad de uso y a su utilización previa para otras tareas, resultan de gran utilidad para los biólogos en su trabajo diario. En la

	A	B	C	D	E	F	G	H
1	id	ng/ul	260/280	V	RIN	fecha	edad	muestra
2	EOBC008	11,9	-	70	1	2008	26	14/12/2011
3	EOBC005	113,9	2,01	30	N/A	2009	30	25/04/2012
4	EOBC012	38,5	1,97	30	N/A	2010	34	18/04/2012
5	EOBC014	70,9	2	30	1	2008	23	18/04/2012
6	EOBC016	12,3	1,86	30		2010	33	18/04/2012
7	EOBC075	410,4	1,95	30			33	18/04/2012
8	EOBC077	23,1	1,94	30			28	18/04/2012
9	EOBC081	265,1	2,12	30			35	18/04/2012
10	EOBC141	72,5	2,13	30	2,4		28	09/05/2012
11	EOBC145	666	1,97	30	2,1	2005	33	15/05/2012
12	EOBC180	178,6	1,95	30	2,3	2011	36	15/05/2012
13	EOBC-048	321,6	2,03	33	2,3	2003	33	25/05/2012
14	EOBC-056	102,3	1,96	33	2,4	2003	35	25/05/2012
15	EOBC-070	263,8	1,96	33		2008	36	25/05/2012
16	EOBC-082	64,2	1,89	33	2,4	2001	35	25/05/2012
17	EOBC-084	94,5	1,94	33	2,4	2010	29	25/05/2012
18	EOBC-096	670,5	2,07	33	2,2		??	25/05/2012

Figura 16. Ejemplo del archivo de Microsoft Excel donde se almacenan los datos de extracción de RNA

actualidad, cualquier análisis o experimento realizado en un laboratorio supone una generación de datos que se almacenan para su posterior utilización. Estos datos se almacenan principalmente mediante la utilización de herramientas de hojas de cálculo, como Microsoft Excel.

Como ejemplo, detallamos el procedimiento llevado a cabo por el Laboratorio de Biología Molecular del Hospital Clínico de Valencia para analizar la expresión de los microARNs en las muestras tumorales de Cáncer de Mama.

El procedimiento comienza cuando llegan muestras de los tumores mamarios en parafina desde la Unidad de Anatomía Patológica. De la muestra recibida extraen el ARN total con unos kits comerciales, obteniendo el ARN diluido. De esa disolución miden la concentración a la que se encuentra el ARN, la pureza (que se mide por la absorción de rayos UV entre 260/280 y se expresa con un valor de 0 a 3), el RIN (RNA Integrity number), el cual les permite saber si el ARN está muy degradado o poco (está representado por un valor de 0 a 10 con decimales) y su intensidad de

	A	B	C	D	E	F	G	H	I	J	K
1	name_miR	EOBC_084	EOBC_101	EOBC_103	EOBC_105	EOBC_134	EOBC_141	EOBC_143	EOBC_145	EOBC_154	EOBC
2	hp_hsa-mir-1224_st	3,71	4,83	4,24	4,36	5,62	4,14	4,78	4,39	4,8	
3	hp_hsa-mir-1248_s_st	2,62	2,36	3,28	2,62	2,56	2,91	2,6	2,93	3,25	
4	hp_hsa-mir-200b_st	2,96	2,77	3,14	3,26	2,91	3,19	3,74	3,41	3,11	
5	hp_hsa-mir-3180-1_s_st	3,4	4,17	3,56	3,83	4,72	3,22	3,71	3,63	4,02	
6	hp_hsa-mir-3180-3_s_st	3,43	3,35	3,19	3,14	3,74	2,35	2,1	3,29	3,16	
7	hsa-let-7a_st	11,78	11,3	11,7	11,38	10,76	11,6	10,94	11,61	11,63	1
8	hsa-let-7b_st	12,37	12,3	12,6	12,14	11,58	12,19	11,21	11,55	12,11	1
9	hsa-let-7c_st	11,42	11,58	11,88	11,55	10,91	11,47	10,05	11,07	11,56	1
10	hsa-let-7d_st	10,28	9,83	10,23	9,79	9,87	10,23	9,18	10,07	10,22	
11	hsa-let-7e_st	10,62	10,26	10,39	10,46	9,29	10,16	9,41	9,15	9,83	
12	hsa-let-7f_st	8,71	6,6	8,79	8,44	7,03	8,52	6,45	9,12	8,57	
13	hsa-let-7g_st	7,76	6,39	7,57	8,18	6,87	8,16	6,11	8,66	8,06	
14	hsa-let-7i_st	9,33	8,66	9,85	9,23	8,76	9,53	9,13	10,01	9,11	
15	hsa-miR-100_st	7,49	7,96	7,08	7,2	7,51	6,91	4,53	6,45	7,03	
16	hsa-miR-103_st	9,86	9,56	10,2	9,53	9,61	10,02	9,98	9,94	9,53	
17	hsa-miR-106a_st	7,89	7,16	7,77	7,34	8,16	7,99	7,76	8,04	6,73	
18	hsa-miR-106b-star_st	4,23	4,13	3,68	3,54	3,32	4,42	3,77	6,21	4,34	
19	hsa-miR-106b_st	6,97	5,51	7,58	7	5,72	7,22	8,01	9,66	5,76	
20	hsa-miR-107_st	9,53	9,15	9,46	9,15	8,72	9,67	9,62	9,42	9,05	

Figura 17. Ejemplo del fichero de Microsoft Excel donde se almacenan las expresiones de los miRNAs analizados en cada muestra mediante el chip de expresión.

fluorescencia obtenida mediante su hibridación en sondas. Estos datos los almacenan en una hoja de cálculo como la de la Figura 16.

De los análisis de las muestras se obtienen dos archivos por cada muestra: un .cel, que contiene los valores intensidad de los microARNs convertidos a un índice numérico, y un .dat, que contiene la imagen escaneada de la fluorescencia de la hibridación (lo que representa la densidad del microARN en la muestra y sirve para llevar un control de calidad). El siguiente paso consiste en procesar el fichero .cel de forma manual con un software específico para normalizarlo, filtrarlo y formatearlo y representar los valores obtenidos en una matriz donde cada fila corresponde a un microARN y cada columna a una muestra analizada. Una vista parcial de esta tabla la tenemos en la Figura 17. Hay que tener en cuenta que el chip está diseñado para medir la expresión tanto de microARNs humanos como de otras especies, por lo que el filtrado llevado a cabo se encarga de seleccionar los humanos, que son los que interesan. Además, desechan también los que tienen una expresión muy baja en todas las muestras y poca variación entre los valores porque les interesa analizar los que varían entre muestras, ya que indican diferencias.

8		Gene Name	Row number	unadj,p	FDR_indep	Obs_stat	abs(Obs_stat)
9	241	hsa-miR-3196	153	5,00E-06	0,0001141	6,744604	6,744604
10	242	hsa-miR-762	230	5,00E-06	0,0001141	6,645506	6,645506
11	243	hsa-miR-939	240	5,00E-06	0,0001141	6,586022	6,586022
12	244	hsa-miR-1909	85	5,00E-06	0,0001141	6,426879	6,426879
13	245	hsa-miR-149	60	5,00E-06	0,0001141	6,422556	6,422556
14	246	hsa-miR-1228	27	5,00E-06	0,0001141	6,015576	6,015576
15	247	hsa-miR-1275	40	5,00E-06	0,0001141	5,887447	5,887447
16	248	hsa-miR-3197	154	5,00E-06	0,0001141	5,767924	5,767924
17	249	hsa-miR-132	49	5,00E-06	0,0001141	-5,572002	5,572002
18	250	hsa-miR-1908	84	5,00E-06	0,0001141	5,411434	5,411434
19	251	hsa-miR-28	128	5,00E-06	0,0001141	-4,424624	4,424624
20	238	v11_hsa-miR-923	248	1,00E-05	0,0001793	5,672394	5,672394
21	239	hsa-miR-3141	144	1,00E-05	0,0001793	5,295238	5,295238
22	240	hsa-miR-92b	237	1,00E-05	0,0001793	4,980808	4,980808
23	237	hsa-miR-4299	191	1,50E-05	0,000251	5,776172	5,776172
24	236	hsa-miR-3175	147	2,00E-05	0,0003137	4,867366	4,867366
25	232	hsa-miR-1202	23	3,50E-05	0,0004392	5,225671	5,225671
26	233	hsa-miR-1224	25	3,00E-05	0,0004392	4,999676	4,999676
27	234	hsa-miR-1225	26	3,50E-05	0,0004392	4,93298	4,93298

Figura 18. Ejemplo del fichero de Microsoft Excel donde se almacenan los resultados del T-Test

Para continuar, realizan un T-Test para ver si hay variación entre unas muestras y otras, entre las muestras de los casos y de los controles. Una muestra del fichero Excel que contiene estos datos se muestra en la Figura 18. Del T-Test almacenan el p-valor (que en la ilustración aparece como *unadj,p* y representa la medida de la significatividad, que indica la diferencia entre los dos grupos), el FDR (que es el p-valor ajustado por múltiples comparaciones) y el Fold-Change (o *Obs_stat,* que representa la media de las diferencias de expresión entre los dos grupos).

El siguiente paso consiste en buscar la asociación de los microARNs significativos con los genes con los que pueden estar actuando. Esta búsqueda se realiza en bases de datos como Diana mirPATH (algoritmo de búsqueda de bases de datos) y TargetScan (BD publica) obteniendo una tabla de resultados como la que aparece en la Figura 19 donde aparecen ordenadas las rutas metabólicas más desreguladas. Una vez localizan los microARNs seleccionan los genes que están en rutas que estén relacionadas con la enfermedad.

Para continuar, vuelven a analizar la expresión de los microARNs seleccionados en el paso anterior utilizando la técnica de PCR-cuantitativa

	A	B	C	D	E	F
1		#	KEGG pathway	p-value		#genes
2						
3		1.	ECM-receptor interaction (hsa04512)	<1e-16		32 see gene
4			hsa-miR-132-5p|microT-CDS	1,08E+01		5see genes
5			hsa-miR-379-3p|microT-CDS	2,87E+00		4see genes
6			hsa-miR-379-5p|microT-CDS	0,002258525		1see genes
7			hsa-miR-134|microT-CDS	3,48E-02		3see genes
8			hsa-miR-181a-2-3p|microT-CDS	0,01074672		1see genes
9			hsa-miR-1908|microT-CDS	0,001040602		1see genes
10			hsa-miR-149-3p|microT-CDS	5,52E-12		11see genes
11			hsa-miR-1207-5p|microT-CDS	0,002034814		7see genes
12			hsa-miR-92b-5p|microT-CDS	0,000236932		5see genes
13						
14		2.	Glycosaminoglycan biosynthesis - chondroitin sulfate (hsa00532)	4,38E-04		1 see genes
15			hsa-miR-1275|microT-CDS	1,59E-12		1see genes
16			hsa-miR-1207-5p|microT-CDS	8,53E-05		1see genes
17						
18		3.	TGF-beta signaling pathway (hsa04350)	6,46E-04		18 see genes
19			hsa-miR-132-5p|microT-CDS	3,07E-03		10see genes
20			hsa-miR-379-3p|microT-CDS	0,003165472		7see genes
21			hsa-miR-409-3p|microT-CDS	5,47E-01		5see genes
22			hsa-miR-433|microT-CDS	0,000510678		1see genes
23			hsa-miR-92b-3p|microT-CDS	0,000116458		1see genes

Figura 19. Hoja del fichero Excel donde guardan los resultados del análisis realizado utilizando la herramienta DIANA mirPath

en otro conjunto de mujeres con la intención de validar los resultados. Como resultado obtienen una tabla como la que aparece en la Figura 20, donde les interesan principalmente los siguientes campos: el número CT (Threshold cycle), que es lo que se obtiene directamente de la PCR (la prueba se hace tres veces y en lugar de almacenar los 3 CT almacenan la media y la desviación), y la expresión relativa de cada muestra (valor RQ - Relative Quantification).

Para finalizar, calculan las diferencias entre los grupos utilizando un test estadístico ANNOVA y almacenan el valor de expresión normalizado, el p-valor y el p-valor ajustado por múltiples comparaciones y la media de la expresión de cada uno de los grupos. Una muestra de ello aparece en la Figura 21.

Como se ha comentado anteriormente, todos los datos que intervienen en estos procesos, se almacenan en varios archivos de texto sin formato o en ficheros de Microsoft Excel. Esta herramienta, a pesar de ser una hoja de cálculo, es ampliamente utilizada en el dominio clínico para almacenar los datos clínicos y biológicos debido a la facilidad de uso, la accesibilidad, la

	C	D	E	F	G	H	I	J	K	L	M	N
1	Sample	Detector	Task	Ct	delta Rn	delta Ct	Ct Avg	Ct SD	Avg Delta Ct	delta Ct SD	Endo Ct Avg	Endo Ct SD
2	EOBC-C09	hsa-miR-1228	Target	29,9601	1,3829947		29,75728	0,1022299	2,80740		26,949878	0,02121326
3	EOBC-C09	hsa-miR-1228	Target	29,67826	1,3466657		29,75728	0,1022299	2,80740		26,949878	0,02121326
4	EOBC-C09	hsa-miR-1228	Target	29,63349	1,5272666		29,75728	0,1022299	2,80740		26,949878	0,02121326
5	EOBC-C33	hsa-miR-1228	Target	29,73226	1,5903302		29,40274	0,165734	2,83048		26,57225733	0,21358648
6	EOBC-C33	hsa-miR-1228	Target	29,26903	1,2580061		29,40274	0,165734	2,83048		26,57225733	0,21358648
7	EOBC-C33	hsa-miR-1228	Target	29,20692	1,5345563		29,40274	0,165734	2,83048		26,57225733	0,21358648
8	EOBC-C43	hsa-miR-1228	Target	30,56097	1,2849537		30,29236	0,2458253	2,95868		27,33368433	0,03302795
9	EOBC-C43	hsa-miR-1228	Target	30,51468	1,3131013		30,29236	0,2458253	2,95868		27,33368433	0,03302795
10	EOBC-C43	hsa-miR-1228	Target	29,80144	1,7001345		30,29236	0,2458253	2,95868		27,33368433	0,03302795
11	EOBC-C18	hsa-miR-1228	Target	30,50108	1,3180609		30,3955	0,0620712	3,73142		26,66407767	0,05791781
12	EOBC-C18	hsa-miR-1228	Target	30,28615	1,4545288		30,3955	0,0620712	3,73142		26,66407767	0,05791781
13	EOBC-C18	hsa-miR-1228	Target	30,39926	1,2044611		30,3955	0,0620712	3,73142		26,66407767	0,05791781
14	1COP	hsa-miR-1228	Target	30,67189	1,8634493		30,60913	0,0464317	2,87728		27,73184733	0,03275591
15	1COP	hsa-miR-1228	Target	30,63702	1,7596313		30,60913	0,0464317	2,87728		27,73184733	0,03275591
16	1COP	hsa-miR-1228	Target	30,51847	1,9069884		30,60913	0,0464317	2,87728		27,73184733	0,03275591
17	EOBCVAL2	hsa-miR-1228	Target	28,93171	1,2282392		28,86324	0,1154173	1,36047		27,502763	0,04231688
18	EOBCVAL2	hsa-miR-1228	Target	29,01992	1,1394546		28,86324	0,1154173	1,36047		27,502763	0,04231688

Figura 20. Frangmento de fichero Excel donde se almacenan los datos de los resultados de la PCR-Cuantitativa

miR	B-F VAR	anova/K-W	M-m	M-J	m-J	media grupo 1(M)	media grupo 2(m)	media grupo 3(J)	Homog. var	t-te
miR-125	**0,0196**	0,2744	> 0,9999	> 0,9999	0,3965	1,4	2	1,2	**0,0022**	-
miR-132	0,6628	0,8581	0,8694	0,9957	0,8726	0,5	1,2	1,3	0,6191	0,81
miR-195	0,2983	0,5934	0,7525	0,9764	0,5656	1,1	1	1,2	0,0671	0,46
miR-23a	**0,0252**	0,2816	> 0,9999	0,3793	> 0,9999	0,5	0,6	0,9	0,1269	**0,07**
miR-27b	**0,0055**	**0,0674**	> 0,9999	**0,0844**	0,5493	1,1	0,9	0,7	0,463	**0,09**
miR28-5p	**0,0155**	0,5832	0,9892	> 0,9999	> 0,9999	1,3	1,2	1,3	0,2152	0,51
miR30c	**0,0327**	**0,0003**	> 0,9999	**0,0024**	**0,0053**	1,1	1,2	1,3	**0,004**	-
miR30e-3p	0,7729	0,8302	0,8153	0,9318	0,9235	1,1	1,2	1,3	0,8695	0,97

Figura 21. Ejemplo del fichero de Microsoft Excel donde se almacenan los datos estadísticos finales

popularidad y la dedicación que supone aprender a utilizar una nueva herramienta compleja.

La complejidad y la diversidad de los datos implicados en todo el proceso clínico descrito requiere que su gestión y explotación se haga utilizando las herramientas de gestión de datos más avanzadas. Esa es la única forma de garantizar un proceso efectivo y eficiente de toda la información disponible. La falta de un soporte adecuado para el almacenamiento de este tipo de información -como sería un sistema de gestión de bases de datos- hace que los datos estén almacenados de manera poco eficiente y que el acceso y gestión de los mismos se convierta en un complejo problema que hace que se pierdan muchas horas del tiempo de los profesionales, además de afectar a la eficiencia del proceso, algo especialmente crítico considerando cuál es el dominio de trabajo analizado.

Es por ello que uno de los objetivos básicos de esta Tesis Doctoral es proporcionar como solución un entorno de gestión avanzada de datos

clínicos profesional, completo, robusto y diseñado conforme a los principios de ingeniería de datos que la complejidad del problema exige.

En la dirección de conseguir ese objetivo, vamos a analizar como siguiente bloque del Estado del Arte cuáles son las fuentes relevantes de datos que hay que tomar en consideración para diseñar el Sistema de Información deseado.

2.2. Estudio de las bases de datos sobre el Cáncer de Mama

Desde que se empezó a estudiar la genética humana se han estado generando datos genéticos que se han ido almacenando en múltiples bases de datos. En los últimos años, y especialmente a partir de la secuenciación completa del genoma humano y del abaratamiento de las técnicas de secuenciación, la cantidad de datos genéticos disponibles ha ido creciendo exponencialmente, y con ello el nacimiento de nuevas bases de datos donde almacenarlos. Ante la falta de un acuerdo global en la utilización de estándares comunes de almacenamiento, estas nuevas bases de datos se han ido creando de forma heterogénea con distintas estructuras y formatos.

El proceso de análisis de una secuencia de ADN de un paciente es un proceso bastante largo y tedioso. Uno de los principales factores de que este proceso sea tan costoso es la gran heterogeneidad de bases de datos donde los genetistas tienen que buscar la información para poder establecer un diagnóstico genético preciso.

Actualmente nos encontramos con múltiples tipos de bases de datos. Existen bases de datos de información genética donde podemos encontrar información de todo tipo: información sobre genes, cromosomas, variaciones, SNPs, fenotipos, enfermedades, secuencias, fármacos, especies, ... Por otra parte, podemos encontrar bases de datos genéticas específicas de enfermedades, donde encontraremos también información de todo tipo, pero con un factor común: todos los datos están relacionados con la misma enfermedad. Además, también existen ontologías específicas que se utilizan para nombrar determinados conceptos relacionados con la genómica, incluso especificando una rama de la misma. Por último, existen además bases de datos genómicas específicas del tipo de moléculas que analizamos. En este caso, estudiaremos también algunas de ellas específicas de las variaciones en los microARNs, con las que trataremos más adelante en esta tesis.

Centrándonos en el cáncer de mama, en este capítulo se incluyen estudios de bases de datos genéticas genéricas, específicas de la enfermedad, específicas de microARNs y algunas ontologías relacionadas con la misma. Además, se incluyen valoraciones de las mismas y de la información que contienen, determinando cuales de ellas son las más apropiadas para ser incorporadas en la base de datos que utilizaremos para la validación de esta tesis.

2.2.1. Bases de datos genéticas

Para empezar el análisis, estudiaremos en primer lugar un conjunto de bases de datos que contienen información genética sobre todo el genoma, sin tener en cuenta unos genes o una enfermedad en particular. La información contenida en ellas es de interés debido a que también disponen de información relacionada con el cáncer de mama.

La selección de estas bases de datos ha estado basada en el criterio de los biólogos expertos, habiendo seleccionado las más conocidas y las más consultadas para la realización de análisis genéticos. Las bases de datos seleccionadas para este estudio son: dbSNP, GenBank, Ensembl, HGMD, UniProt, Open Access GWAS Database, COSMIC, GWAS Catalog, OMIM y KEGG.

dbSNP

En septiembre de 1998, el National Center for Biotechnology Information (NCBI) creó una base de datos a la que llamó dbSNP para cubrir la necesidad de disponer de un catálogo general de las variaciones del genoma donde poder almacenar y consultar la gran cantidad de información genómica que se extrae de los estudios de asociación, de cartografía

Figura 22. Logo de dbSNP

genética y de biología evolutiva [18, 19]. Se encuentra accesible desde la url http://www.ncbi.nlm.nih.org/SNP, donde se encuentra el logo representativo de la base de datos que se incluye en la Figura 22. Desde su creación, la base de datos dbSNP ha servido como un repositorio público centralizado de variaciones genéticas. Desde el momento en el que las variaciones son identificadas y catalogadas en la base de datos, los laboratorios de todo el mundo pueden utilizar la información sobre dichas variaciones y sobre las condiciones experimentales específicas para futuras aplicaciones en investigación. Al igual que todos los recursos de NCBI, los datos almacenados en dbSNP están disponibles gratuitamente y en múltiples formatos.

No hay ningún requisito de entrada en esta base de datos sobre la frecuencia mínima de los alelos o la neutralidad funcional de los polimorfismos. Por lo tanto, dbSNP se incluyen tanto mutaciones que pueden causar alguna patología, como polimorfismos con un fenotipo neutro. Las entradas de dbSNP incluyen información sobre la secuencia que rodea al polimorfismo, las condiciones experimentales necesarias para llevar a cabo la prueba, la descripción de la población que contiene la variación, y la frecuencia de la variación por población o genotipo individual.

dbSNP enlaza sus variaciones (polimorfismos y mutaciones) con otros recursos disponibles en NCBI gracias al BLAST y los análisis E-PCR de la secuencia de nucleótidos que rodea la variación. Los enlaces a bases de datos bibliográficas se establecen a partir de la información citada en el momento en que se incluyó la variación en la base de datos. La integración de dbSNP con otros recursos de información genómica se extiende más allá de NCBI gracias al uso de 'LinkOut URLs' que hacen referencia a bases de datos externas con más información sobre la variación. Esta integración es importante si se pretende llevar a cabo la gran labor de almacenar toda la información disponible sobre la variación y sus consecuencias para el organismo. La red interna de NCBI ya está estructurada de esta manera en la que las variaciones son catalogadas en dbSNP mientras que las

descripciones funcionales de la región de la secuencia local se describen en otras bases de datos como GenBank, dbSTS, RefSeq, LocusLink o UniGene.

Puesto que los genes y sus nucleótidos están potencialmente implicados en múltiples *pathways* y, por tanto, en múltiples fenotipos intermedios, NCBI no almacena los detalles bioquímicos o las consecuencias fenotípicas de la variación directamente en la misma entrada. Esta información se almacena gracias a los vínculos que dbSNP tiene con bases de datos externas que caracterizan cada uno de los ejes particulares de la variación fenotípica.

dbSNP tiene una frecuencia de actualización variable que suele rondar sobre los 3 meses.

Mecanismos de acceso a la base de datos

Las búsquedas en dbSNP pueden realizarse directamente desde el buscador de la misma base de datos o a través de otros recursos de NCBI [20]. Como cualquier otro miembro de la familia NCBI, los contenidos de dbSNP también están accesibles a través de enlaces disponibles en otros recursos de NCBI. Además, se puede acceder a dbSNP a través de la herramienta de comparación de cadenas BLAST, que es capaz de comparar la secuencia introducida por el usuario contra todas las secuencias que contienen las variaciones almacenadas en dbSNP, devolviendo al usuario las entradas de dbSNP que coincidan con la muestra introducida.

dbSNP se ha diseñado para facilitar las búsquedas a lo largo de cuatro ejes principales de información: ubicación en la secuencia, función, homología entre especies, y el grado de heterocigosidad (grado de variación en la población). Al establecer en las búsquedas los umbrales de inclusión en uno o varios de estos ejes, los usuarios pueden extraer el subconjunto de registros que mejor se adapten a sus necesidades de investigación.

Además, esta base de datos dispone de un servidor FTP de uso público donde podemos descargar toda la información en distintos formatos dependiendo de las necesidades o preferencias de los usuarios. Este FPT se encuentra accesible en ftp://ftp.ncbi.nlm.nih.gov/snp/.

Análisis razonado del interés de la base de datos

La interfaz del usuario que muestra esta herramienta es algo compleja y difícil de entender la primera vez que te encuentras con ella. Además, la documentación que ofrece no es demasiado clara ni precisa como para solucionar las dudas concretas que puedan surgir sobre lo que representa cada uno de los elementos que aparecen en la página de cada variación. El uso continuado de esta base de datos puede mejorar mucho su uso y entendimiento.

A pesar de todos los inconvenientes del uso de los servicios web que ofrecen los gestores de esta base de datos, la información que contiene es muy valiosa y ampliamente aceptada por la comunidad científica. De hecho, la mayoría de las bases de datos que se encuentran en la web nutren sus repositorios de la información encontrada en esta base de datos. Y es que en realidad estamos hablando de una base de datos de información muy fiable y que forma parte de la tan conocida colección de NCBI.

Como indicamos anteriormente, cuando hablamos de NCBI, nos referimos al Centro Nacional de Información Biotecnológica (*National Center of Biotechnology Information*), el cual forma parte de la Biblioteca Nacional de Medicina (*National Library of Medicine*) de los Institutos Nacionales de Salud (*National Institutes of Health*) de Estados Unidos. La información que contienen sus bases de datos se encuentra actualizada y controlada por el personal de este centro, y es utilizada y tenida en cuenta como referencia por la gran mayoría de especialistas e investigadores de todo el mundo.

Otro valor a su favor, es el servidor FTP de donde se puede descargar toda la información disponible en múltiples formatos. Esta información puede descargarse de forma bastante selectiva. De esta manera se puede hacer una selección de aquellos datos que interesen realmente creando un repositorio de datos propio con la precisión deseada por el usuario.

Todo esto convierte a dbSNP en una base de datos muy recomendada para ser utilizada.

Teniendo en cuenta la cantidad de información que recoge esta base de datos, debemos seleccionar aquellos datos que realmente se adapten a nuestras necesidades. La información sobre variaciones que contiene esta base de datos es muy amplia y solicitada. Debemos centrar nuestro análisis en obtener únicamente las variaciones de aquellos genes de la especie humana en los que vamos a centrar las búsquedas. En esta base de datos en particular podemos encontrar tanto SNP's como mutaciones de cualquier especie.

GenBank

Genbank es la base de datos de secuencias genéticas del NIH (*National Institute of Health* de EEUU) gestionada por el NCBI (*National Center for Biotechnology Information*). Esta base de datos pretende hacer una distribución pública de todas las secuencias de ADN anotadas disponibles [21].

El NIH tiene sus orígenes en el año 1887, cuando dentro del Hospital de Servicio de la Marina se creó un laboratorio que fue predecesor del *U.S Public Health Service* (PHS). GenBank se creó en 1982 y hacia finales de 1983 ya tenía almacenadas más de 2000 secuencias.

La web de la base de datos está disponible en https://www.ncbi.nlm.nih.gov/genbank/

GenBank está diseñada para proporcionar acceso dentro de la comunidad científica a la información sobre la secuencia de ADN más actualizada. Además, NCBI no pone ninguna restricción a la difusión y uso de esta información.

NCBI está continuamente desarrollando nuevas herramientas y mejorando las ya existentes para mejorar tanto el acceso como el envío de datos a GenBank. Es posible subscribirse a una lista de noticias para mantenerse informado de los nuevos cambios y actualizaciones. Además, los cambios son publicados en la sección "What's new" de la web.

La frecuencia de actualización de los datos y herramientas ronda los 2 meses. Existe información disponible a través de FTP de la versión actual de GenBank.

Mecanismos de acceso a la base de datos

Además de su interfaz web de búsqueda, GenBank ofrece un sitio FTP para compartir datos.

Toda la información está disponible también a través de su sitio FTP ftp://ftp.ncbi.nih.gov/genbank/

Es posible descargar la información que aparece en el FTP mediante archivos comprimidos TGZ que contienen a su vez archivos de secuencias "seq".

Además, ofrece un software cliente para el acceso a la base de datos (disponible en el sitio FTP en ftp.ncbi.nih.gov/entrez/network) y también un cliente de BLAST (un programa de comparación de secuencias.). Este software lo que hace es emular la información que se puede obtener vía web tanto de Entrez como de BLAST, de hecho, en la web se cita este software dirigiéndonos a la parte web que tiene su misma funcionalidad.

Otra forma de acceder a la base de datos es con peticiones vía URL, usando una herramienta "cgi" llamada e-utilities. También es posible descargar información vía petición URL, por ejemplo, descargar la información de algunos genes en ficheros FASTA.

Finalmente, cabe resaltar que no ofrece una capa de servicios web para poder tratar la información mediante peticiones SOAP o REST.

Análisis razonado del interés de la base de datos

GenBank es una base de datos de referencia dentro de la comunidad científica. Se actualiza diariamente en la web y bimensualmente en el servidor FTP. Cada nuevo hallazgo queda registrado y disponible.

Es importante destacar que la información que se obtiene es fácilmente accesible a través de su directorio FTP. Además, el software proporcionado también puede facilitar la obtención de resultados en combinación con la herramienta web. La obtención de información vía URL puede ser más tediosa y complicada debido a las diversas opciones que se contemplan en la petición.

La web está bien estructurada en secciones y hay un ejemplo de salida de datos al realizar una búsqueda.

Se echa de menos una sección de ayuda propia de la web explicando el tipo de búsqueda que se puede realizar, bien por identificador de gen, descripción, función, etc. Puede ser un poco compleja para neófitos o profesionales recién iniciados en este campo, pero para gente experimentada sin duda supone una herramienta básica en su trabajo diario.

Ensembl

La base de datos de Ensembl proporciona una herramienta bioinformática para organizar los datos biológicos en torno a las secuencias de los grandes genomas. El objetivo del proyecto es la creación de un software libre que facilite las tareas relacionadas con la ciencia genómica proporcionando información de alta calidad sobre los genomas eucariotas más estudiados a través de una infraestructura consistente y accesible [22-24]. La base de datos se encuentra disponible en www.ensembl.org.

Ensembl, cuyo logo representativo podemos ver en la Figura 23, es un proyecto conjunto del *European Bioinformatics Institute* (EBI) y el Centro Sanger, ambos situados en el *Wellcome Trust Genome Campus* de Cambridge (Reino Unido). El 27 de enero de 2000, anunciaron la

Figura 23. Logo de Ensembl

finalización del "Ensembl Milestone 1", la primera versión completa de los datos del proyecto y su interfaz web. Esta versión incluía la información de genes y variaciones disponibles en ese momento y proporcionaba apoyo a las predicciones genéticas basadas en la evidencia científica, incluyendo las comparaciones entre proteínas homólogas. Todos los datos eran accesibles desde la propia web y podían ser descargados vía FTP. Diez años más tarde, Ensembl proporciona un conjunto de información genómica mucho más grande y completo incluyendo conjuntos de genes, alineamientos entre especies, información sobre genes ortólogos y parálogos (genes que tienen similitudes en distintas especies o en la misma especie, respectivamente) y una extensa colección de variaciones e información sobre la regulación. Además, proporciona acceso a los recursos más avanzados para las especies más estudiadas, incluyendo los humanos, el ratón, la rata y el pez cebra, reflejando de esta manera la popularidad y la importancia de estas especies en la investigación biomédica. La base de datos se actualiza varias veces al año, aproximadamente cada 3 meses, con nuevas especies y nuevas versiones de los genomas. Además, para los nuevos conjuntos de datos, la actualización incluye mejoras en el software y la visualización, lo que mejora el código base y la integración de los nuevos datos.

Ensembl integra información procedente de fuentes de datos de especies o dominios específicos, como ZFIN, HGNC, dbSNP, UniProt y ENCODE. Los datos están disponibles a partir de un amplio surtido de interfaces entre las que se incluyen Ensembl Genome Browser, Perl API y BioMart. Además, existe la posibilidad de hacer una copia de todos los datos y código para ser usada libremente por la comunidad científica.

Mecanismos de acceso a la base de datos

Los datos de Ensembl se encuentran disponibles a través una serie de servicios. La elección de cada uno de ellos debe considerarse dependiendo de la cantidad y el tipo de datos que queramos buscar.

Si lo que buscamos consta de pequeñas cantidades de datos, como un único gen, tenemos la opción *Export Data* desde la que se puede descargar la

información disponible en la web. En la ventana de *Export Data* existe una serie de opciones para elegir qué datos deseas descargar y el formato en el que los quieres.

En cambio, si estamos buscando grandes conjuntos de datos como, por ejemplo, todos los genes en un cromosoma, o análisis complejos, podemos utilizar el servidor MySQL de acceso público.

Además, está disponible una API de Perl para descargar información de esta base de datos sin necesidad conocer el esquema de la misma.

Como alternativa, podemos acceder a un subconjunto de datos de Ensembl utilizando el protocolo *DAS (Distributed Annotation System)*, a través de una simple URL.

Para realizar consultas complejas a la base de datos existe la herramienta de minería de datos BioMart [25], que es una herramienta de minería de datos altamente personalizable, lo que facilita la tarea de la extracción de datos de esta base de datos.

Finalmente, si lo que queremos es descargar la base de datos completa podemos hacerlo desde el servidor FTP en varios formatos.

Ensembl ofrece un servicio de comparación de cadenas que usa *Blast* y *Blat*. Este servicio está disponible online y utiliza por defecto el algoritmo de *Blat* por ser, según los creadores de este servicio, más rápido que *Blast (http://www.ensembl.org/Help/Faq?id=429)*.

Análisis razonado del interés de la base de datos

La información en esta base de datos es muy amplia y está muy bien estructurada. Un detalle a tener en cuenta es el posicionamiento de las variaciones respecto a los tránscritos en lugar de hacerlo respecto al ADN. En este caso, si queremos posicionar las variaciones respecto al gen o al cromosoma se debería estudiar la posibilidad de realizar un cambio de coordenadas.

Debemos tener en cuenta que esta base de datos se nutre de la información recogida en otras bases de datos, principalmente en dbSNP, por lo que, si tuviésemos la información de dbSNP cargada en una base de datos, tendríamos que hacer una selección de datos para no volver a introducir los mismos.

Por otra parte, Ensembl posee un servicio de descarga de sus datos con una gran variedad de posibilidades. Cada una de ellas tiene unas características distintas pudiendo elegir los detalles de la descarga, de forma que los datos se ajusten lo mejor posible a nuestras necesidades, ya que parte de la información que contiene la base de datos no va a ser necesaria para esta tesis.

Además, en esta base de datos tenemos mucha información sobre genes, transcritos y variaciones que pueden ser de utilidad. La información está perfectamente relacionada entre sí, y al poder seleccionar los datos que interesa descargar, es mucho más fácil seleccionar únicamente aquellos detalles que nos interesan.

Todo lo comentado hace de esta una base de datos muy útil, siempre teniendo en cuenta el cambio de coordenadas.

HGMD

Human Gene Mutation Database (HGMD) constituye una colección básica de datos sobre mutaciones en los genes, asociadas a alguna enfermedad hereditaria humana [26-28]. Esta base de datos, representada por el logo de la Figura 24, utiliza la extracción manual para obtener de la literatura una lista de las mutaciones relacionadas con trastornos

Figura 24. Logo de HGMD

mendelianos, que luego organiza con un formato estructurado que facilita las tareas de búsqueda. Los datos catalogados incluyen: sustituciones de un único par de bases en áreas codificantes, reguladoras y de "*splicing*"; micro-borrados y micro-inserciones; *indels* (inserciones y borrados en la misma posición); repeticiones de tripletes, así como grandes borrados, inserciones, duplicaciones, y translocaciones complejas. Cada mutación se introduce en HGMD una única vez con el fin de evitar confusión entre las recurrentes y las que tienen resultados idénticos. Esta base de datos se encuentra accesible desde http://www.hgmd.cf.ac.uk/ac/index.php.

HGMD se hizo público por primera vez en abril de 1996. Aunque originalmente se creó para el estudio científico de los mecanismos de mutación en los genes humanos, HGMD ha adquirido una utilidad mucho más amplia para investigadores, médicos y consejeros genéticos, así como para empresas especializadas en productos biofarmacéuticos, bioinformáticos y genómica personalizada. Más tarde, en 2006, se firmó un acuerdo de cooperación entre HGMD y BIOBASE GmbH que abarca la comercialización mundial de la información genómica más actualizada en una versión de HGMD - HGMD Professional-, para su uso académico, clínico y comercial, disponible mediante la compra de una licencia. HGMD se suele actualizar cada 3 meses.

Mecanismos de acceso a la base de datos

HGMD es una base de datos privada con un perfil público limitado. Se necesita una suscripción gratuita para poder acceder al perfil público de la misma, pero hay ciertas limitaciones de acceso a los datos. Los datos disponibles en la versión gratuita tienen una antigüedad de más de dos años y medio y la forma de acceso es únicamente vía web. La versión profesional (HGMD Professional) se actualiza cada 3 meses y provee además herramientas avanzadas de búsqueda e información adicional específica de genes y mutaciones ausente en la versión pública. Se pueden guardar los resultados de las búsquedas en archivos de texto plano o como archivos del navegador del genoma, que pueden visualizarse utilizando navegadores tanto públicos como privados, pudiendo elegir entre el formato de UCSC

(.bed) o el de CLC Genomics Workbench (.gff). Además, existe la posibilidad de descargarse las tablas SQL que componen la base de datos en su totalidad.

Análisis razonado del interés de la base de datos

Esta base de datos contiene un recopilatorio de aquellas mutaciones que han sido publicadas y pueden ser referenciadas por una o varias publicaciones. Los responsables se encargan de revisar la bibliografía para ir introduciendo manualmente las mutaciones en la base de datos. Además, es una base de datos muy frecuentada por múltiples genetistas.

La información que contiene es de alta calidad, gracias al trabajo de los responsables de la misma, y se encuentra perfectamente estructurada, de manera que facilita considerablemente su consulta.

El mayor inconveniente de esta base de datos es la falta de actualización de la versión gratuita y la limitación de que los datos sean de solo consulta, sin posibilidad de descarga. Esto hace casi imprescindible la compra de la licencia para poder disponer de la versión profesional.

En cualquier caso, toda la información que deberá ser incluida en el Sistema de Información que nos proponemos diseñar debe tener en cuenta los datos que proporciona esta base de datos, convenientemente contextualizada para los datos relacionados con las variaciones de tipo patológico del cáncer de mama, la enfermedad que analizamos en esta Tesis.

UniProt

La base de datos se creó el 15 de diciembre de 2003 para dar acceso a la anotación y secuenciación de proteínas [29-31]. UniProt fue creada y es mantenida por el Consorcio UniProt, el cual es fruto de la colaboración entre el Instituto Europeo de Bioinformática (EBI), el Instituto Suizo de Bioinformática (SIB) y el Repositorio de Información de Proteínas (PIR).

Figura 25. Logo de UniProt

Juntos deciden poner en común sus recursos y experiencia formando así el consorcio UniProt, cuyo logo aparece en la Figura 25, y la base de datos ligada a este. Podemos acceder a la misma desde http://www.uniprot.org/.

UniProt incluye información sobre registros anotados de proteínas. La información es extraída de la literatura científica relacionada. La base de datos proporciona información que va desde los productos proteicos derivados hasta información sobre ensayos llevados a cabo o tratamiento informático de estos.

Las principales funciones que proporciona son: búsqueda por texto, búsqueda por similaridad de secuencias (*BLAST*), alineamiento de secuencias, recuperación de lotes y mapeo de los identificadores de bases de datos.

El sitio web de la página contiene links a otros recursos relacionados que pertenecen al consorcio creador de UniProt, como *EBI Services* [32], *ExPASy*: Bioinformatics Resource Portal of SIB [33] o PIR Links [34], donde se ofrecen enlaces a herramientas bioinformáticas o bases de datos con información biomédica.

UniProt se actualiza mensualmente y permite recomendar cambios para siguientes versiones.

Mecanismos de acceso a la base de datos

El acceso a la base de datos puede hacerse online desde la propia web o descargando los *sets* de datos en alguno de los múltiples formatos disponibles en su servidor (http://www.UniProt.org/help/mapping). La información del servidor FTP puede ser descargada en formato: fasta, text o XML.

Análisis razonado del interés de la base de datos

UniProt es un repositorio universal de información acerca de secuencias de proteínas e información acerca de las funciones biológicas de las proteínas. Los evaluadores han extraído la información de la literatura científica relacionada y llevado a cabo análisis de las secuencias ligados a la revisión clínica asociada a datos tanto experimentales como predictivos. A su vez contiene registros analizados computacionalmente y enriquecidos con anotación automática.

La base de datos contiene una sección dedicada a almacenar la captura de la información pública disponible referente a secuencias de proteínas.

Uniprot proporciona un conjunto de servicios que permiten al usuario explorar y analizar la información analizada. Permite llevar a cabo consultas basadas en texto simples o complejas, ejecutar búsquedas basadas en secuencias, realizar alineamientos múltiples, recuperar múltiples entradas y mapear identificadores a partir de una base de datos externa a UniProt o viceversa.

El interés principal de esta base de datos estriba en la combinación de información fiable sobre secuencias de proteínas con un alto nivel de anotación, mínima redundancia, una buena integración con otras bases de datos y un acceso universal y gratuito.

Por el contrario, no se proporciona información clara sobre la estructura interna de la base de datos. El usuario tiene acceso al XML-Schema pero no existe un documento donde se resuman las entidades de mayor relevancia. Además, al estar hablando de variaciones sobre proteínas en lugar de sobre genes, la traducción necesaria para poder utilizar esta información para los análisis de secuencias de ADN es muy compleja y costosa.

Es importante destacar como la complejidad asociada a la gestión de información genómica con proyección clínica, se va poniendo de manifiesto. Solo con el análisis inicial de bases de datos genéticas que hemos introducido hasta ahora, podemos concluir que existen distintos

repositorios de datos que incluyen distintas perspectivas de la información relevantes para el diagnóstico y tratamiento de una enfermedad (genes, cromosomas, variaciones, mutaciones, proteínas...). Y la lista es inevitablemente extensa. De ahí la necesidad de realizar un trabajo de investigación como el desarrollado con esta Tesis, que seguimos presentando.

Open Access GWAS Database

OpenGWAS es una base de datos de acceso libre sobre estudios de asociación del genoma completo (openGWAS proviene de *Open Access Database of Genome-Wide Association Studies*) creada por Andrew Johnson y Christopher O'Donnell del *National Heart, Lung and Blood Institute* de Bethesda, EEUU [35]. El objetivo de esta base de datos es el de solventar el problema de la falta de un repositorio centralizado donde poder almacenar los resultados de los múltiples resultados de los diferentes estudios GWAS que existen. La carga inicial se hizo con los resultados de 118 artículos sobre GWAS, cada uno de los cuales fue seleccionado haciendo una evaluación ciega por parte de 3 o 4 revisores de modo que se asegurara la calidad de los mismos. De estos artículos se extrajeron 56.411 asociaciones fenotípicas SNP significativas. El artículo donde apareció publicada la base de datos data de enero de 2009 [35]. Se puede acceder a información sobre ella desde el artículo publicado en la siguiente URL http://bmcmedgenet.biomedcentral.com/articles/10.1186/1471-2350-10-6. Al no disponer de portal web, tampoco tiene imagen corporativa.

Contiene información que la relaciona con otras bases de datos como RefSeq o dbSNP. Tanto las bases de datos descritas como el propio artículo que la explica son de acceso libre.

La base de datos openGWAS está centrada en los siguientes estudios:

- Desordenes adictivos
- Alzheimer
- Esclerosis lateral amiotrófica (ELA)

- Presión sanguínea, hipertensión
- Cáncer
- Enfermedades cardiovasculares
- Enfermedad de Crohn
- Rasgos relacionados con los lípidos
- Parkinson
- Artritis reumatoide
- Diabetes tipo 2
- Rasgos relacionados con el peso/índice de masa corporal

Mecanismos de acceso a la base de datos

Esta base de datos no puede ser consultada de forma online, solo puede ser descargada en formato Microsoft Access 2007 en la siguiente dirección web:

(http://www.biomedcentral.com/content/supplementary/1471-2350-10-6-s4.zip).

Esta misma información se encuentra disponible como tsv (valores separados por tabuladores) en un fichero de texto

(http://www.biomedcentral.com/content/supplementary/1471-2350-10-6-s2.zip).

Los autores también han puesto disponibles más de 400 análisis GWAS en formato Genome Graphs que pueden ser usados para visualizarlos con la herramienta UCSC Genome Graphs. Se incluye en el mismo fichero un fichero con nombre "JohnsonODonnell_ALLgwas _graph.txt", que contiene un único grafo con todas las asociaciones.

Análisis razonado del interés de la base de datos

Revisando la información sobre cáncer de la base de datos, contiene información sobre cánceres de próstata, colorectal y de mama. Es por ello que la información contenida relacionada con el cáncer de mama en esta base de datos nos puede resultar muy interesante.

Por otra parte, hay que tener en cuenta que no está publicada la frecuencia de actualización de la misma, por lo que podría estar desactualizada y obsoleta. Los datos de esta base de datos son interesantes, pero complementarios a otra base de datos actualizada recientemente.

COSMIC

COSMIC es un catálogo de mutaciones somáticas de cáncer creado por el *Wellcome Trust Sanger Institute* el 3 de febrero de 2004. El motivo de su creación fue el poner disponibles públicamente los datos sobre mutaciones somáticas que provocan alteraciones en las células en las que ocurren [36, 37]. Podemos acceder a su página web en http://cancer.sanger.ac.uk/cosmic, donde se encuentra el logo que vemos en la Figura 26.

Entre sus principales características se encuentra el poder consultar de forma sencilla y online información relativa a los genes relacionados con alteraciones en los tejidos (y subtejidos), pudiendo consultar de los mismos qué genes son los más significativos y cuantas muestras se tienen de ellos en la base de datos. Esta información se visualiza en un histograma para mayor rapidez de consulta. Además, permite acceder de un modo directo a la lista completa de mutaciones y genes con o sin mutaciones para este tejido.

El ritmo de actualización de esta fuente de datos es cada 2 meses.

A diferencia de otras bases de datos, COSMIC pone a disposición de los usuarios información sobre los datos contenidos en la misma. En la versión actual, contiene información sobre 19.737 genes, 177.322 mutaciones, 6.365 fusiones de genes y 619.320 muestras de tumores.

Figura 26. Logo de COSMIC

También da acceso a realizar extracciones completas o parciales (filtrado y selección de atributos) de la información contenida en ficheros o ficheros comprimidos mediante la herramienta online COSMICmart [38]. El formato del mismo está basada en BioMart [25] (la herramienta de minería de datos de Ensembl), dado que la propia base de datos de Cosmic está basada en esa misma tecnología.

Mecanismos de acceso a la base de datos

El acceso a la base de datos puede hacerse online desde la propia web o descargándose la base de datos en alguno de los múltiples formatos disponibles en el servidor FTP (ftp://ftp.sanger.ac.uk/pub/CGP/cosmic). La información del servidor FTP puede ser descargada en formato Excel (xls), valores separados por comas (csv), valores separados por tabuladores (tsv), fasta y base de datos Oracle (sql).

Además, contiene las siguientes exportaciones ya realizadas listas para su uso en todo tipo de experimentos:

- Exportación completa con y sin las mutaciones codificadas.
- Exportación completa con y sin las mutaciones codificadas, incluyendo además las mutaciones de fusión.
- Exportación de las mutaciones de fusión.
- Exportación de todas las mutaciones de inserción, especificando la secuencia insertada.
- Exportación de todas las mutaciones de una sola base y fusiones.
- Exportación de todas las mutaciones de una sola base.
- Exportación de las reorganizaciones genómicas estructurales.
- Listado de todos los genes en Cosmic, con referencias a otras fuentes.
- Conjunto de todas las mutaciones en el CGP Cell Line Project.

Como punto de interés cabe destacar que también se realizan extracciones de datos bajo demanda, pudiendo realizar peticiones para un conjunto de datos que se ajuste exactamente a las necesidades de cada proyecto.

Análisis razonado del interés de la base de datos

Esta base de datos nos puede resultar interesante porque se centra explícitamente en las mutaciones que provocan cánceres, identificando a qué tejidos u órganos afecta dicha mutación. Esto puede resultar ser muy útil, dado que la base de datos posee numerosa información sobre mutaciones relacionadas con el cáncer de mama. La información de cada uno de los genes está sacada en muchos casos de un conjunto de artículos seleccionados, siendo normalmente usada la información de los últimos artículos y estudios relacionados con el gen (si bien la propia web avisa de que para algunos casos no se trata de un estudio exhaustivo). La información de cada uno de los genes está relacionada con otras bases de datos como Ensembl, NCBI Entrez Gene, CCDS, HGNC, Swiss-Prot, OMIM y Atlas Genetics Oncology.

Una simple búsqueda sobre mutaciones que afectan a la mama devuelve 43,749 resultados en la base de datos Cosmic, lo que a priori sugiere la existencia de gran cantidad de información relevante para el proyecto de esta tesis.

Un aspecto relevante que debe ser explorado es el de la interrelación de la información sobre variaciones proporcionadas por las distintas bases de datos analizadas al respecto. Este punto vuelve a poner de manifiesto el problema de la complejidad de una gestión de datos asociada a un contexto tan sensible como lo es el clínico en general –y el Cáncer de Mama en particular-, objeto de este trabajo de Tesis Doctoral.

La información de esta base de datos está accesible vía internet de un modo completamente público y gratuito. Las mutaciones están organizadas dando especial importancia al tejido al que afecte la mutación. La interfaz web permite hacer búsquedas sobre gen, muestra, tejido, identificador de PubMed o descripción de la mutación. Además, permite explorar la lista de todos los genes disponibles y la lista de todos los tejidos.

GWAS Catalog

Este proyecto está financiado por el *National Human Genome Research Institute* (NHGRI) pertenenciente al *National Institutes of Health* (NIH) de EEUU. Su web con el logo identificativo que vemos en la Figura 27 puede encontrarse en http://www.ebi.ac.uk/gwas/ y comenzó en el año 2009 con el objetivo de recopilar todos los SNPs asociados con fenotipos o enfermedades y así poder relacionar distintos *loci* con diferentes rasgos. Desde septiembre de 2010 la distribución y desarrollo de Catalog ha sido un proyecto colaborativo entre el EMBL-EBI (*European Bioinformatics Institute*) y el NHGRI. En marzo de 2015 la infraestructura del Catalog se movió a EMBL-EBI para mejorar la interfaz de usuario, incluyendo la búsqueda dirigida por ontologías y una nueva infraestructura de revisión, mejorando la calidad de los procesos [39, 40].

Este proyecto recopila las variantes (SNPs) de más de mil artículos científicos donde se estudian más de 100.000 SNPs en alguna enfermedad o fenotipo. Aquellos SNPs que han resultado asociados a algún fenotipo o enfermedad con un p-valor menor de 10^{-5} son almacenados en este sistema de información.

Este catálogo de SNPs no se actualiza de manera periódica, sino a medida que se van publicando estudios que relacionen estas variantes con distintos fenotipos. Esta información pasa por un proceso de revisión de dos niveles, haciendo especial hincapié en la extracción y verificación de los datos (para evitar duplicidades) y en la descripción del fenotipo y obtención de los datos étnicos [40].

Figura 27. Logo de GWAS Catalog

Mecanismos de acceso a la base de datos

Desde la misma página web, hay varias maneras de acceder a la información. La más visual es un diagrama donde muestra todos los SNPs situados en los distintos cromosomas y, a su vez, relacionados con los distintos fenotipos. Es posible filtrar la información para mostrar únicamente los SNPs relacionados con un determinado fenotipo. También disponen de una herramienta de búsqueda donde se puede acceder a la información de Catalog introduciendo la característica fenotípica, identificador del SNP, estudio o gen relacionado. Además, también tenemos la información disponible en una página de descargas donde encontraremos toda la información de Catalog, de la versión actual y versiones anteriores, en varios ficheros tabulados o en otros formatos como RDF o OWL.

Análisis razonado del interés de la base de datos

El interés en la información de Catalog radica en la posibilidad de secuenciar variantes alélicas en los exomas que previamente han sido asociadas o relacionadas con ciertas enfermedades. De esta forma tendríamos accesible la información conocida de más de mil artículos, lo que nos permitiría realizar un mejor diagnóstico.

Además, debido a la selección y revisión manual de las variaciones que se incluyen en su base de datos, la información que ofrece se caracteriza por su consistencia y fiabilidad, características valoradas por los expertos. Sin embargo, esta selección minuciosa puede llevar a retrasos en la actualización de los datos o falta de información, que han de ser tenidas en cuenta y solventadas utilizando otros recursos adicionales.

OMIM

Este proyecto comenzó al principio de la década de 1960 por el Dr. Victor A. McKusick como un catálogo de fenotipos y desórdenes. En 1985 se creó la versión online mediante la colaboración de la *National Library of Medicine* y la *Johns Hopkins University*. Finalmente, en 1995, el *National*

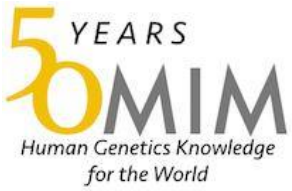

Figura 28. Logo de OMIM

Center for Biotechnology Information (NCBI) del *National Institute of Health* (NIH) creó la base de datos que conocemos actualmente como *Online Medelian Inheritance in Man* (OMIM) [41, 42]. La página web del proyecto es http://omim.org/, presidida por el logo creado para celebrar el 50 aniversario del proyecto, que aparece en la Figura 28.

El objetivo de esta base de datos es la creación de un catálogo con todos los genes humanos y todos sus fenotipos, con una atención especial en relacionar la variante genética con el fenotipo expresado. Actualmente contiene información sobre todos los desórdenes mendelianos conocidos, así como 13700 genes humanos y más de 3000 fenotipos.

Se actualiza diariamente y está relacionada con muchos otros recursos como Ensembl o Uniprot, ofreciendo enlaces a la información relacionada con la entrada en cuestión disponible en cada uno de estos recursos. Además, cada variación almacenada está relacionada con toda la bibliografía referente a la misma.

Mecanismos de acceso a la base de datos

Al igual que ocurre con otros recursos bioinformáticos, lamentablemente no existe una base de datos propiamente dicha. La información se parece más a un catálogo que a una base de datos. De hecho, la información se suministra exclusivamente en un fichero de texto plano, donde está escrita como un texto de lenguaje natural, lo cual complica la extracción de la información.

Para solventar este problema una interfaz web muy sencilla permite realizar búsquedas a partir de nombres de genes, texto, cromosoma, enfermedades,

fenotipos... pero no existe ninguna herramienta que permita hacer una minería de datos exhaustiva de la información contenida en este catálogo.

La información de este catálogo está dirigida a médicos y profesionales relacionados con los desórdenes genéticos. Está información es pública y descargable desde un FTP accesible desde la propia web bajo un previo registro.

Análisis razonado del interés de la base de datos

Se trata de un recurso muy interesante, ya que contiene información sobre variantes genéticas y enfermedades recopiladas desde hace más de 30 años. Además, las enfermedades y variantes genéticas están ampliamente descritas y relacionadas con muchos otros recursos bionformáticos como Ensembl, UniProt, Pubmed o UCSC entre otros.

Esta información será de gran interés cuando se localicen las mutaciones o variantes en las muestras de los pacientes, ya que nos permitirá relacionar los distintos locus con los fenotipos o enfermedades.

El único problema de este recurso tan valioso es que no existe actualmente ninguna herramienta que permita su explotación de una forma automática.

KEGG

Desde 1995, la base de datos KEGG ha sido desarrollada en la Universidad de Kyoto y en la Universidad de Tokio gracias a la financiación del gobierno. El acceso a esta base de datos se realiza a través de su página web http://www.genome.jp/kegg/ donde se encuentra el logo representativo que vemos en la Figura 29.

Figura 29. Logo de KEGG

KEGG es una base de datos integrada formada por la unión o integración de varias bases de datos, ampliamente categorizadas en sistemas de información, información genética y química [43]. KEGG ha sido ampliamente usada como referencia base de conocimiento para la interpretación biológica de conjuntos de datos de gran escala generados por la secuenciación y otras tecnologías experimentales de alto rendimiento. Presenta información variada y dispersa, entre la que se destaca la relacionada con enfermedades humanas, medicamentos, genes, genomas y rutas metabólicas. Organiza la información en forma de grafos, relacionando conceptos y entradas de la base de datos entre sí.

KEGG es actualmente una de las más usadas según estadísticas de acceso web (entre 150 y 200 mil visitantes únicos por mes) y el número de citas en documentos (alrededor de mil al año).

KEGG presenta muchos recursos disponibles para su uso desde la web y aplicaciones de escritorio.

Los más destacados son:

- Búsqueda en la base de datos interna (DBGET)
- Búsqueda externa en bases de datos enlazadas con KEGG (LinkDB)
- Comparación de secuencias y estructuras moleculares (BLAST y SINCOMP)
- Herramientas para el desarrollo de software (xml, soap API, peticiones directas por url)
- Herramientas de escritorio (KegHier, KegArray, KegDraw)

La ayuda en la web no es de fácil acceso y la búsqueda de información puede ser un poco complicada si no se está familiarizado con esta web y formato.

No se proporciona la fecha de actualización de ninguna de las bases de datos integradas.

Mecanismos de acceso a la base de datos

Dispone de un punto de entrada desde la que se centraliza la información y las búsquedas. Además, proporciona acceso a búsquedas más específicas organizadas por categorías, por ejemplo, Medicamentos o Enfermedades Humanas.

Proporciona además herramientas de análisis, como comparación de secuencias con BLAST o en formato FASTA (uno de los formatos más populares para representar secuencias de nucleótidos y aminoácidos).

Una característica muy interesante de esta base de datos es que presenta una API para su acceso a través de servicios web, tanto SOAP como REST. Es posible descargar los archivos wsdl, javas, etc., necesarios para el desarrollo de clientes a KEGG, aunque no es posible descargar masivamente toda la información de KEGG, para lo cual se requiere una licencia de pago.

Otra característica de esta web es que permite peticiones a la base de datos vía *url* siguiendo unos patrones definidos en http://www.genome.jp/kegg/docs/weblink.html.

KEGG ofrece ya una librería java de servicios web para su uso, implementada con Axis 1.4. Además, ofrece el archivo WSDL que describe cada servicio web y su tipo de retorno, además del punto de entrada a los servicios web, para una implementación de clientes a estos servicios en cualquier otro lenguaje de programación que soporte la creación de clientes de servicios web.

KEGG también ofrece software para al análisis y trato de la información, como por ejemplo la aplicación KegArray, un programa diseñado para el análisis de datos provenientes de microarrays.

Otro mecanismo de acceso a la base de datos es mediante un servidor FTP, disponible desde Julio de 2011. Hay varias subscripciones y perfiles de usuario, todas ellas de pago.

Análisis razonado del interés de la base de datos

A pesar de que en un principio parece una interfaz complicada, después de unas pruebas iniciales, se puede conseguir un concepto general de realización de búsquedas y uso de las herramientas que proporciona KEGG.

Esta herramienta es interesante desde el punto de vista biológico porque centraliza varias bases de datos de diferente contenido y temática, desde *pathways* hasta medicamentos. Se trata de un nuevo tipo de información genómica (más ligado al metabolismo celular y los *pathways*) que no está contemplado en las bases de datos analizadas hasta ahora. También es importante porque es un proyecto vivo y en continuo desarrollo. Otra característica funcionalmente relevante es que posee un servidor FTP, en el que hay que registrarse y pagar una licencia para obtener acceso a los datos de la base de datos.

Desde un enfoque de desarrollo informático, gracias a los servicios web y su API, es posible desarrollar clientes que obtengan información de la base de datos y presentarla y formatearla a nuestra conveniencia.

La parte que más destaca respecto a otras bases de datos, además de su contenido de datos, es su alta interoperabilidad para integrarse con otros sistemas, desde capas de servicios web (ofrece ya un fichero *.jar* con los servicios web desarrollado en Axis 1.4) y peticiones por url hasta aplicaciones cliente de escritorio.

La descripción de todas las bases de datos a las que ataca y recoge información está detallada en http://www.genome.jp/dbget/.

2.2.2. Bases de datos con información sobre microARNs

Después de analizar una selección de las bases de datos genómicas más generalistas por su contenido, procedemos en esta subsección a continuar el análisis en un contexto más especializado centrado en un aspecto de gran

trascendencia en la investigación genómica actual: los microARNs, y su importancia en el estudio del cáncer. Este análisis es relevante porque en la parte experimental final de esta Tesis, usaremos datos de microARNs en un ambiente de investigación real con pacientes con cáncer de mama, y será esencial por lo tanto integrar estos datos con otros datos provenientes de esas bases de datos genómicas que hemos etiquetado como generalistas, que son las vistas hasta ahora.

miRBase

Esta base de datos fue fundada bajo el nombre de *microRNA Registry* en 2002 por el *Wellcome Trust Sanger Institute* [44-46]. En la actualidad se encarga de mantenerla el *Griffiths-Jones Laboratory* en la *Faculty of Life Sciences* de la *University of Manchester*, con financiación del BBSRC (*Biotechnology and Biological Sciences Research Council* del Reino Unido). Su objetivo inicial fue el de proporcionar a los investigadores de microARNs una nomenclatura estable y única para los nuevos microARNs descubiertos, así como un archivo de todas las secuencias de microARNs. Se puede acceder a ella a través de su URL http://www.mirbase.org/ donde se puede encontrar su logo corporativo que vemos en la Figura 30.

Sus funciones pueden resumirse en los siguientes puntos:

- Proporcionar un sistema coherente de nomenclatura para microARNs.
- Proporcionar un servicio que recoja todas las secuencias conocidas de microARNs y facilite la búsqueda online y descarga de todos los datos de microARNs.
- Proporcionar anotaciones legibles por el hombre y el ordenador

Figura 30. Logo de miRBase

para cada secuencia de microARN.
- Proporcionar acceso a una evidencia primaria que dé soporte a las anotaciones de microARNs.
- Agregar y vincular información de dianas de microARN.
- Expandir las anotaciones textuales y funcionales y un servicio de agregación para el creciente número de predicciones de dianas de microARNs y validaciones.

Mecanismos de acceso a la base de datos

Para acceder a los datos de secuencias, miRBase cuenta con varios métodos: mediante navegación por la base de datos, mediante búsqueda de similitud de secuencias, mediante intervalos de coordenadas genómicas, por búsqueda de palabras clave y mediante descarga [44].

A partir de los resultados y del tipo de búsqueda, haciendo clic sobre el nombre del microARN se accede a la página de entrada. En ella se incluye información de la secuencia precursora y la secuencia madura, así como enlaces a publicaciones que describen la identificación de los microARNs.

La base de datos ofrece la posibilidad de descargar la base de datos relacional completa en un fichero .sql para poder instalarla como una instancia de base de datos local. Además, es posible descargar las secuencias localmente a través de un servidor FTP en una amplia variedad de formatos. Finalmente, también dispone de una interfaz web de consulta para poder realizar consultas vía online.

Análisis razonado del interés de la base de datos

MiRBase nos permite obtener información sobre las alteraciones en los niveles de expresión de los microARNs en determinadas muestras, principalmente tumorales. Además, proporciona niveles de confianza de cada microARN y permite introducir una secuencia para comparar su similitud con las secuencias disponibles en la base de datos. Contiene secuencias de microARNs, tanto del microARN maduro como del pre-microARN, así como una serie de anotaciones.

Esta base de datos es considerada en el mundo de la biología como una base de datos de referencia para los microARNs, ya que es la que establece la nomenclatura de los microARNs que el resto de fuentes de datos utilizan comúnmente.

Es interesante resaltar que este recurso pone a disposición del usuario de forma gratuita su base de datos en .sql y ficheros en múltiples formatos con la información contenida en sus bases de datos con el fin de que el usuario pueda integrarlas en sus propias bases de datos y utilizar la información de la forma que le resulte más cómoda.

DIANA Tools

El objetivo de DIANA Tools es proporcionar algoritmos, bases de datos y software para interpretar y archivar datos en un entorno sistemático procedentes de análisis de la expresión de microARNs, la anotación de los elementos reguladores de microARNs y la interpretación del papel de ncARNs (ARNs no codificantes) en diversas enfermedades y rutas metabólicas. La página web de este recurso es http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=site/index y está encabezada por el logo que aparece en la Figura 31.

Entre los recursos disponibles de DIANA Tools se encuentran algoritmos de predicción de dianas (microT v4 y microT-CDS) [47], bases de datos de dianas de microARNs verificadas experimentalmente para ARNs codificantes y no codificantes (TarBase v7.0 y LncBase) [48, 49] y un software capaz de identificar las rutas metabólicas potencialmente alteradas por la expresión de un único o múltiples microARNs (mirPath) [50]. Además, un nuevo Servidor Web (v5.0) [47] soporta varios de flujos de trabajo simultáneos, permitiendo a los usuarios realizar complejos análisis funcionales de microARNs sin necesidad disponer de la infraestructura bioinformática necesaria para llevar a cabo este tipo de análisis.

Figura 31. Logo DIANA Tools

De los recursos nombrados, las bases de datos más relevantes son las siguientes:

Diana-microT-CDS, que es la 5ª versión del algoritmo microT [47]. Está programado específicamente con un conjunto de Elementos de Reconocimiento de microARNs (MRE) situados tanto en las regiones 3'-UTR como en las CDS, que permiten predecir la relación de los microARNs con sus respectivos genes diana. Diana-microT-CDS se actualizó por última vez en julio de 2012 a miRBase 18 y es totalmente compatible con la nueva nomenclatura de microARNs introducida en esta versión. También proporciona enlaces a servidores en línea, como iHOP, y a los datos de expresión de los microARNs seleccionados en tejidos y líneas celulares. DIANA-microT-CDS puede ser accedido desde la siguiente dirección: http://www.microrna.gr/microT-CDS, y toda la información que maneja la base de datos es fácilmente descargable desde el área de descargas de DIANA Tools.

TarBase v7.0 es la mayor base de datos de dianas revisada de forma manual, con más de 65.000 interacciones indexadas[48]. La base de datos incluye dianas extraídas de estudios específicos, así como de experimentos de alto rendimiento, tales como microarrays y proteómica. Además, incluye dianas derivadas de experimentos de secuenciación, como HITS-CLIP y PAR-CLIP. TarBase contiene la mayor cantidad de información disponible online procedente de estudios 3 CLIP-Sec y 12 degradoma-Seq. La base de datos está perfectamente conectada con otras DIANA-lab Tools, tales como DIANA-microT, lo que permite ampliar la información de cada interacción validada con la información predicha. DIANA-TarBase ofrece una cantidad significativa de información crucial para el usuario, incluyendo la descripción detallada de los genes implicados y microARNs, una lista de publicaciones que referencian cada interacción y los métodos experimentales utilizados para las validaciones junto con sus resultados. La base de datos ofrece también enlaces a las rutas metabólicas de KEGG, así como a otras bases de datos externas, como Ensembl, Uniprot y RefSeq. También está equipada con potentes funciones de búsqueda y filtrado.

TarBase v7.0 se puede acceder desde la siguiente dirección: http://www.microrna.gr/tarbase.

Mecanismos de acceso a la base de datos

DIANA Tools dispone de un servicio de descargas donde pueden descargarse en un fichero comprimido los datos de cualquiera de sus bases de datos, previo envío de un formulario con datos personales y profesionales para su registro, indicando además, el propósito de los datos que van a ser descargados.

Análisis razonado del interés de la base de datos

Se trata de una plataforma ampliamente utilizada en la comunidad científica. Dispone de herramientas muy útiles para descubrir las consecuencias de las alteraciones en la expresión de los microARNs y bases de datos muy completas que se pueden descargar en ficheros tabulados y gestionar localmente de forma gratuita.

Además, la información contenida en estas bases de datos se identifica utilizando los identificadores de miRBase, por lo que es muy sencillo conectar la información de ambas bases de datos.

Sin embargo, la estructura de su base de datos no se encuentra disponible ni para ser descargada ni consultada, por lo que requiere un proceso de extracción, transformación y carga en una base de datos propia de manera manual.

TargetScanHuman

TargetScan es un servidor web que predice dianas biológicas de microARNs mediante la búsqueda de la presencia de sitios que coinciden con la región de la semilla de cada microARN (TargetScan define la región de la semilla como las posiciones 2-7 del microARN maduro). En comparación con otras herramientas de predicción de dianas, TargetScan ofrece rankings precisos de las dianas predichas para cada microARN [51, 52]. Estas clasificaciones se basan ya sea en la probabilidad de tener dianas

Figura 32. Logo de TargetScan Human

conservadas evolutivamente (*PCT scores*), o en la eficacia prevista de la represión (*context++ scores*). Su página web se encuentra accesible en http://www.targetscan.org encabezada por el logo de la Figura 32.

Creada por el laboratorio BartelLab en el año 2003, esta herramienta tiene como objetivo principal identificar los genes diana de los microARNs de animales vertebrados. Para ello, realiza una predicción de las dianas biológicas de los microARNs mediante la búsqueda de la presencia de sitios 8mer, 7mer y 6mer que coincidan con la región de la semilla de cada microARN.

TargetScanHuman 7 contiene todos los microARNs maduros procedentes de la versión 21 de miRBase (la última disponible), pero para algunos microARNs conservados hace algunas modificaciones basadas en el análisis de datos procedentes de la secuenciación de alto rendimiento [51]. En estos casos se han incluido algunos microARNs adicionales debido a las terminaciones 5' alternativas observadas. Estos microARNs están indicados con un sufijo de .1 o .2.

Además, TargetScan hace su propia clasificación de los microARNs en familias, basándose en las mismas regiones de origen, por lo que la clasificación por familias de microARNs e incluso la nomenclatura pueden variar entre miRBase y TargetScan.

Mecanismos de acceso a la base de datos

Para acceder a esta base de datos disponemos de dos vías. La primera es el software en línea al que podemos acceder desde http://www.targetscan.org/vert_70/.

Por otra parte, desde la página principal de TargetScan encontramos un link ("*Download data or code*") de acceso al sitio web de descarga. Desde

aquí podemos descargar un conjunto de ficheros tabulados donde queda distribuida toda la información que manejan en su sistema de información. Además, desde esta misma página podemos descargar unos scripts en Perl para realizar consultas y análisis sobre los datos descargados.

Análisis razonado del interés de la base de datos

Es una base de datos sencilla y cómoda de utilizar muy útil para conocer las predicciones de asociaciones de genes diana con los microARNs. Por lo tanto, estamos hablando de predicciones, no de información validada, lo que puede ser interesante o no, dependiendo de la finalidad del análisis realizado y del tipo de información buscada.

Un interesante punto a su favor es que se encuentra actualizada a la última versión de las nomenclaturas de miRBase, lo que ayuda mucho al manejo de sus datos y permite conectar la información de predicciones de asociaciones de genes diana con los datos de esta base de datos de referencia.

Además, tener acceso a la información en formato descargable supone una facilidad de uso para los usuarios que prefieran manejar los datos en sus propios sistemas de gestión de base de datos.

2.2.3. Bases de datos con información específica del cáncer de mama

Pretendiendo completar nuestro estudio de las principales bases de datos de información genómica utilizadas en el entorno clínico y de investigación del cáncer de mama pasaremos a centrarnos en las bases de datos específicas de la enfermedad, aquellas que contienen información seleccionada sobre el cáncer de mama. Es interesante tener en cuenta estas bases de datos en nuestro estudio, puesto que preseleccionan la información del caso de estudio que tratamos en esta tesis.

Genes-to-Systems Breast Cancer Database

La base de datos Genes to Systems Breast Cancer (G2SBC) es un recurso bioinformático creado por el "Institute for Biomedical Technologies, National Research Council, Segrate (Milan), Italy" que recoge e integra datos sobre genes, transcritos y proteínas que han sido reportadas como alteradas en células de cáncer de mama. Su propósito es almacenar y centralizar información de otras fuentes de datos específicas sobre el cáncer de mama y hacerla accesible a la comunidad científica con un sistema unificado de consultas vía interfaz web [53] desde http://www.itb.cnr.it/breastcancer/. La imagen corporativa queda representada por el logo que aparece en la Figura 33.

Sobre la arquitectura de la base de datos cabe mencionar que trabaja bajo Mysql Server y que recolecta y formatea datos de otras fuentes de datos dedicadas. Se accede a los datos mediante scripts realizados en Perl.

A partir de las búsquedas, podemos acceder por ejemplo, a la secuencia FASTA de un gen en concreto, tanto de nucleótidos, ARN y Aminoácidos, a la ontología (GeneOntology [54]), los SNPs (obtenida de dbSNP [19]), *pathways* (obtenidos de KEGG [43]), e incluso a medicamentos y artículos relacionados con dicho gen.

Se integra con otras herramientas de otras plataformas, por ejemplo, con el visor de UCSC [55], que permite una navegación del cromosoma dividido por pistas o tracks, en los que es posible seleccionar la información asociada a esa región de cromosoma que se encuentra en estudio.

Además, contiene información acerca de más de 2000 genes. Entre sus principales fuentes de datos destacan:

Figura 33. Logo de G2SBC

- NCBI
- Oncomine
- Tumor gene family of databases
- Breast Cancer Database
- Breast Cancer Information Core Database
- dbSNP
- RefSeq
- Standford MicroArray Database
- Protein Data Bank
- Human Protein Atlas

Esta fuente de datos tiene la novedad respecto a otras fuentes de la misma temática, de que presenta herramientas de cálculo de modelos matemáticos. En concreto, ofrece una herramienta para obtener genes asociados al cáncer involucrados en el control del ciclo celular y simular modelos matemáticos y otra también de modelos matemáticos relacionada con el crecimiento de tumores, cancerogénesis y respuesta a tratamientos. Además, permite la caracterización de resultados mediante anotación enriquecida de genes. También ofrece la comparación de cadenas con BLAST y predicción de resultados.

Mecanismos de acceso a la base de datos

El único mecanismo de acceso a la base de datos de una forma externa es a través de una interfaz web implementada en PHP y javascript que utiliza scripts en Perl (lenguaje comúnmente utilizado entre la comunidad bioinformática) para el acceso a los datos.

Los datos son únicamente de consulta, no se pueden descargar. El único dato que es posible guardar es la secuencia de nucleótidos, ARN o aminoácidos del gen sujeto a estudio. El formato en que se presentan dichos datos es la secuencia FASTA en texto plano dentro de un archivo html que, de nuevo, no es posible descargar, hay que copiar el texto y pegarlo en un fichero manualmente.

Análisis razonado del interés de la base de datos

G2SBC es una fuente multinivel dedicada a la biología y sistemas moleculares del cáncer de mama, incluyendo niveles estructurales (genes, transcritos y proteínas) y niveles de sistemas (sistemas moleculares y celulares, poblaciones de células). Tomando ventaja de una aproximación multinivel y de una serie de herramientas para el análisis del contenido de la base de datos, G2SBC proporciona predicciones, que pueden inducir a nuevos experimentos.

La unificación de datos que proporciona esta base de datos resulta interesante, teniendo accesible en un mismo recurso información sobre genes, transcritos y proteínas sobre el cáncer de mama procedentes de múltiples fuentes.

Sin embargo, esta base de datos no presenta una capa de servicios web, ni otro sistema de consulta o intercambio de información aparte de la proporcionada por su interfaz web. Estos motivos hacen que la descarga de los datos a nivel local, aspecto de gran interés para el trabajo de esta tesis, resulte imposible.

The Breast Cancer Information Core Database (BIC)

El Instituto Nacional de Investigación del Genoma Humano (NCHGR acrónimo en inglés), fue fundado en 1989 para llevar a cabo el papel del Instituto Nacional de Salud (NIH) de los Estados Unidos en el Proyecto Internacional del Genoma Humano (HGP). Con la secuencia del genoma humano completada desde abril del 2003, cualquier científico tiene acceso a una base de datos que facilita y acelera el camino de la investigación biomédica. La URL de acceso a su portal web es https://research.nhgri.nih.gov/projects/bic/Member/index.shtml. En este caso, la base de datos en si misma carece de imagen corporativa.

Esta base de datos se centra en los dos genes con más información sobre el cáncer de mama hereditario, el BRCA1 y BRCA2. La reciente identificación de mutaciones en los genes susceptibles de causar cáncer de mama, ha

proporcionado la oportunidad de ayudar a identificar mujeres con riesgo alto de padecer cáncer de mama. Esta nueva información sobre mutaciones debe ser coordinada y centralizada. Este proyecto se centra en ese problema, la colección y centralización de información sobre las mutaciones y polimorfismos en los genes BRCA1 y BRCA2 [56, 57].

Mecanismos de acceso a la base de datos

El acceso a la base de datos se hace vía web. Hay que registrarse para obtener acceso, aunque este registro es gratuito y en pocos segundos se obtiene un usuario y contraseña vía e-mail.

Los datos se obtienen a través de su interfaz web. Además, únicamente es posible descargar la información de la base de datos en un fichero tabulado en formato texto plano. Los enlaces se dividen en 2 secciones, una para cada gen. Se pueden hacer varios tipos de búsquedas, desde un mapa de exones se puede acceder a la información del exón seleccionado y también se puede acceder a un formulario de búsqueda desde otro enlace.

No existen, o no se citan ni en literatura buscada en la web ni en la página principal de la base de datos, mecanismos de interoperabilidad y conexión con otras plataformas.

En cualquier caso, es posible contactar vía email con el creador de la web y el gestor de la base de datos para evaluar la posibilidad de conexión directa hacia su SGBD.

Análisis razonado del interés de la base de datos

El interés de esta web radica en la información relevante que se obtiene de los dos genes citados anteriormente y en la importancia de este organismo como centro investigador del genoma humano.

Los genes BRCA1 y BRCA2 son considerados los dos genes con más relevancia en el cáncer hereditario de mama e indispensables para trabajar en la prevención de la enfermedad. En el caso de la base de datos G2SBC,

recopila información de muchos otros genes relacionados con la enfermedad, principalmente cuando el tumor ya se ha desarrollado.

Por otra parte, teniendo en cuenta el organismo fundador de la base de datos (el NCHGR), su carácter investigador y de condición gubernamental, y su ánimo de centralizar y organizar la información descrita anteriormente, podemos considerar a esta base de datos de alta relevancia para la comunidad científica que estudia el cáncer de mama.

The Tumor Gene Family of Databases

Esta fuente de datos fue creada por el *Biomedical Computing Inc.* en colaboración con la *University of Texas Health Science Centre* y el *Baylor College of Medicine*, todos ellos localizados en Houston (Texas, EEUU). El propósito de la misma fue reunir información de artículos relacionados con varios tipos de cáncer, entre los que se encuentra el cáncer de mama seleccionando la base de datos *Breast Cancer Gene Database* al realizar la búsqueda del gen en cuestión [58]. Podemos acceder a ella a través de su página web http://www.tumor-gene.org/tgdf.html, la cual es bastante sencilla y carece de imagen corporativa.

Contiene información sobre genes que son objetivos de causas de cáncer, oncogenes y genes supresores de tumores. Según especifica la página web, contiene más de 2600 hechos relacionados con más de 300 genes [59].

Mecanismos de acceso a la base de datos

Se accede mediante interfaz web, no se encuentra un mecanismo alternativo tipo servicio web. Se presenta un formulario simple con una entrada de texto y un desplegable para seleccionar la base de datos, en este caso, es de interés la de cáncer de mama. En la página principal hay un pequeño ejemplo de cómo hacer búsquedas en la base de datos.

Los resultados de la búsqueda se presentan en un patrón maestro / detalle, una tabla con información breve y un enlace para ampliar esa información.

En la página con información detallada únicamente se presentan citas a artículos sin enlaces a ellos, organizados por temática o tipo de información. Todo ello en texto plano y sin posibilidad de descarga automatizada.

Análisis razonado del interés de la base de datos.

El interés de esta base de datos radica en que tiene un apartado específico para el cáncer de mama y que además forma parte de las bases de datos que se consultan en G2SBC Database.

La información se presenta de forma sencilla y simple siendo ésta brece y escueta y no aportando a su vez novedad alguna respecto a otras bases de datos. Tampoco presenta información sobre los datos que ofrece, ni bibliografía asociada que explique su origen de datos o tratamiento de los mismo.

2.2.4. Conclusiones

El objetivo alcanzado con este análisis del estado del arte es triple:

- Identificar las fuentes de datos genómicas más relevantes, incidiendo en el ámbito de este trabajo de Tesis que es el Cáncer de Mama.
- Poner de manifiesto la complejidad del entorno de trabajo analizado, donde convive una gran cantidad de información, de origen diverso, heterogénea, presentada bajo formatos diferentes, con problemas de consistencias de datos al comparar unas fuentes con otras, y en el que asegurar la calidad de los datos que deben ser gestionados se convierte en un proceso altamente exigente.
- Seleccionar un conjunto de bases de datos tomadas como las de referencia a la hora de proveer los datos para el sistema de información de gestión de datos para cáncer de mama que se va a presentar

Después de revisar las características más importantes de cada plataforma estudiada, basándonos en los requerimientos (interoperabilidad, volumen de datos, especificidad de datos, calidad de datos y frecuencia de actualización de datos), para seleccionar las plataformas de trabajo más adecuadas se ha seguido el criterio de hacerlo diferenciando las de información general de las específicas de cáncer de mama. Entre las bases de datos de información genética general más relevantes englobaríamos las pertenecientes a NCBI, como son dbSNP, GenBank y OMIM, así como HGMD, COSMIC o KEGG. Otras bases de datos a tener en cuenta serían tanto Genes-to-Systems Breast Cancer database (G2SBC) como Breast Cancer Information Core Database (BIC). Ambas son específicas del estudio de la enfermedad y nos permitirían sacar el máximo partido a la interoperabilidad de las bases de datos genéticos generales para completar la información o contrastarla.

dbSNP es la base de datos de referencia por excelencia en cuanto a variaciones genéticas se refiere. Pertenece al conjunto de bases de datos de NCBI y la información que contiene es muy valiosa y ampliamente aceptada por la comunidad científica. Además, tiene múltiples posibilidades de descarga de datos, por lo que se pueden manejar cómodamente. La información contenida en esta base de datos es de gran relevancia para esta tesis.

GenBank es otro componente de las bases de datos de NCBI que debemos integrar en el sistema de información que se va a presentar. Contiene información sobre genes y secuencias de referencia. No ofrece servicios web, pero tiene implementado software cliente para la consulta de sus datos, FTP disponible y público y también ofrece varias alternativas de integración e interoperabilidad como *BLAST* y consultas vía *URL*.

El otro recurso de NCBI analizado es OMIM, con información muy interesante sobre enfermedades, pero poco accesible, ya que únicamente contiene información en un texto plano muy difícil de manejar.

Por otra parte, HGMD es una base de datos privada y revisada por profesionales que aseguran la validez de los datos. A pesar del

inconveniente de tener que comprar la licencia profesional para el uso de la base de datos completa, la información que contiene es de alta calidad y se encuentra perfectamente estructurada.

COSMIC también contiene una gran cantidad de información relevante, ya que se centra explícitamente en las variaciones somáticas, es decir, aquellas que aparecen en los tejidos tumorales, identificando a qué tejidos u órganos afecta dicha mutación, lo que nos facilita la tarea de relacionar las variaciones con el cáncer de mama.

Respecto a KEGG, podría usarse como complemento, ya que ofrece buena interoperabilidad, presentando una capa de servicios web y demás herramientas de integración descrita en su sección. A pesar de no ser específica del cáncer de mama, es un muy buen recurso para analizar gracias a todos sus mecanismos de acceso, una secuencia o gen de interés en nuestro estudio, una vez identificado. Además, contiene información sobre medicamentos, de gran importancia para los servicios que tenemos que ofrecer, y de la que carecen la mayoría de bases de datos.

G2SBC es una base de datos que recopila y aúna información de muchas otras fuentes de información. Además, su grupo de investigación ofrece una rápida respuesta ante consultas sobre su plataforma. Esto último nos indica que su grupo de trabajo está a pleno rendimiento. Además, el acceso a su base de datos no es sencillo, ya que habría que contactar previamente con los administradores para solicitar acceso a su SGBD.

Por otra parte, BIC es un sistema en actualización continua, por lo que registra los últimos cambios o hallazgos, y permite la descarga de sus datos, con lo que podríamos hacer pruebas offline. La desventaja de BIC es que únicamente almacena información respecto a dos genes relacionados con el cáncer de mama, BRCA1 y BRCA2, que a su vez son los genes más estudiados y los más analizados para esta enfermedad.

Si pretendemos integrar información sobre estudios de asociación del genoma completo (GWAS) en este capítulo se han analizado dos bases de datos al respecto: Open Access GWAS Database y NHGRI_GWAS_catalog. La primera de ellas contiene información sobre cáncer de mama, por lo que

podría ser muy interesante. Sin embargo, hay que tener en cuenta que no está publicada la frecuencia de actualización de la misma, por lo que podría estar desactualizada y obsoleta. La segunda de ellas tiene una mayor frecuencia de actualización (dos veces al año aproximadamente), sin embargo, no se trata de una base de datos real, sino más bien de un catálogo disponible en un fichero de texto, tabular o en formato Excel. Por tanto, estas dos bases de datos podrían ser complementarias para cubrir la información sobre GWAS.

Respecto a las bases de datos de microARNs, la más importante de todas y referencia de las demás es la base de datos miRBase. Esta base de datos se utiliza como referencia, almacena las secuencias de referencia de los microARNs y establece la nomenclatura estándar que el resto de bases de datos utilizan para referirse a los microARNs.

Las otras dos bases de datos de microARNs se encargan de almacenar las relaciones entre microARNs y sus genes diana. En este caso, hemos analizado las dos fuentes de datos más importantes. La primera, DIANA Tools, es un conjunto de herramientas que dispone de dos bases de datos, ambas descargables gratuitamente, con distinto contenido: una de ellas, microT-CDS, contiene predicciones de microARNs con genes diana y la otra, TarBase, también contiene asociaciones de microARNs con genes diana, pero en este caso validadas experimentalmente. TargetScanHuman tiene la misma funcionalidad que microT-CDS, pero dispone de más microARNs que han sido detectados utilizando la técnica de secuenciación de alto rendimiento, además de disponer de una clasificación por familias distinta a la de miRBase. Cabe destacar que DIANA Tools trabaja con la versión 18 de miRBase, mientras que TargetScanHuman trabaja con la versión 21, la última versión disponible.

Finalmente, se podían descartar tres de las bases de datos analizadas por problemas varios. Una de ellas es Ensembl, que pese a ser una base de datos bastante clara y estructurada, tiene su información referida a distintos transcritos, lo que incrementaría la dificultad de unificar los datos con los provenientes de otras bases de datos con referencias génicas. Además, esta base de datos contiene información proveniente de otras bases de datos,

por lo que podemos obtener la información de otras fuentes sin necesidad de realizar el cambio de posicionamiento. Otra base de datos con el problema de la diferencia de posicionamiento es Uniprot. En esta base de datos las posiciones de las variaciones están referidas a la cadena de proteínas, por lo que todavía es más complicada la conversión. Finalmente, la última de las bases de datos con posibilidad de ser descartada es la Tumor Gene Family of Databases, que no ofrece ninguna información sobre el grupo de investigación que la mantiene, ni qué actualizaciones tiene, ni cuando fue creada. La única información que tenemos sobre ella la hemos encontrado rebuscando entre las publicaciones científicas que se encuentran disponibles en la red. Esta falta de información nos impide saber si los datos que contiene son lo suficientemente fiables como para utilizarlos en procesos de diagnóstico genético.

2.3. Herramientas o tecnologías existentes para la gestión de datos clínicos

Tanto en el entorno clínico como en el investigador, el estudio del cáncer de mama no conlleva únicamente el manejo de datos genéticos, sino también datos clínicos de los pacientes. Poder conectar la información clínica con la información genética es esencial para relacionar los perfiles clínicos de los pacientes con los datos genéticos de las muestras e inferir nuevas asociaciones que ayuden a mejorar la prevención, diagnóstico y tratamiento de la enfermedad.

Como parte del Estado del Arte de esta tesis, es importante conocer las soluciones existentes en la actualidad para gestionar los datos clínicos en los hospitales y centros de investigación para poder detectar en qué y cómo se puede mejorar la gestión de datos desde el punto de vista de la teoría de Sistemas de Información y Modelado Conceptual.

2.3.1. Historia Clínica Electrónica: Estándar ISO 13606

La norma ISO 13606, propuesta como CEN/TC251 EN 13606 por el Comité Técnico 251 del CEN (Comité Europeo para la Estandarización), fue publicada por la ISO (Organización Internacional para la Estandarización) entre los años 2008 y 2010 con la finalidad de establecer pautas para la comunicación de la historia clínica electrónica.

El motivo de su creación fue responder a las necesidades de normalización de los datos clínicos para compartir información entre diferentes sistemas, ya que en la actualidad impera la diversificación de sistemas de información dentro (sistemas de información de laboratorio o radiología) y fuera del centro de atención sanitaria (entre diferentes instituciones).

Los objetivos que piensa cubrir la norma incluyen definir una estructura de información estable y rigurosa para comunicar partes de la historia clínica

electrónica (HCE) de un paciente y soportar la interoperabilidad de sistemas y componentes que necesitan comunicar (acceder, transferir, modificar o añadir) datos de HCE. Todo ello conservando el significado clínico original y reflejando la confidencialidad de los datos. Además, la norma está orientada a la separación de la información y el conocimiento (modelo dual de datos), hecho que dota al sistema de información basado en la ISO 13606 de capacidad evolutiva: la información permanece estable, mientras que el conocimiento evoluciona [60].

Se compone de cinco partes:

- **Parte 1: modelo de referencia.** Establecida en el año 2008, se relaciona con la interoperabilidad sintáctica, es decir, describe la estructura de mensajes y ficheros para que los diferentes sistemas de información puedan interpretar su contenido. En definitiva, presenta las bases para el intercambio de información [60].
- **Parte 2: especificación de intercambio de arquetipos.** Publicada en el año 2008, se relaciona con la interoperabilidad semántica, esto es, presenta unas pautas para intercambiar conocimiento. Si la parte 1 permitía el intercambio de información, la parte dos asocia a esta información un significado que no debe verse alterado en la comunicación entre sistemas de información y, para ello, se combina con terminologías como SNOMED-CT [61].
- **Parte 3: arquetipos de referencia y listas de términos.** Publicada en el año 2009, abarca listas de términos con vocabulario controlado referido a cada atributo del modelo de referencia. Además, describe algunos arquetipos de referencia que especifican la forma en que el modelo de referencia debe utilizarse para representar información de la clase acto de HL7 Version 3 y especializaciones de la clase entry de openEHR [62].
- **Parte 4: seguridad.** Establecida en el año 2009, forma parte de la arquitectura formulada en la parte 1 y presenta un marco básico que especifica los requisitos de acceso a la información de la HCE y el control de su uso para garantizar la confidencialidad de la

información [63].

- **Parte 5: especificación de interfaz.** Presentada en el año 2010, describe las interfaces que deben poseer los sistemas de información para comunicar extractos de HCE estandarizados con toda o parte de la HCE (REQUEST_EHR_EXTRACT), arquetipos (REQUEST_ARCHETYPES) y la historia de actividad de los registros realizados para llevar a cabo auditorías (REQUEST_EHR_AUDIT_LOG_EXTRACT) [64].

Algunos de los **recursos disponibles** relacionados con la norma incluyen:

- **CIMM**. Es un gestor de modelos de información clínica que permite publicar y almacenar arquetipos. Su propósito es que los usuarios compartan los arquetipos ya existentes para crear un repositorio [65].
- **LinkEHR**. Es una plataforma para modelar, integrar, normalizar y dotar a los datos de interoperabilidad semántica [66, 67]. Se compone de cinco módulos que podemos ver representados en la Figura 34:
  - **LinkEHR Editor**. Se trata de una herramienta orientada a la clínica que permite modelar arquetipos (crear, especializar y revisar).
  - **LinkEHR Studio**. Posibilita la transformación de los datos basados en arquetipos y utilizarlos en sistemas y datos existentes.
  - **LinkEHR Concept Manager**. Se trata de una aplicación web destinada a la gestión de modelos de información clínica y arquetipos.
  - **LinkEHR Integration Engine**. Permite la integración de los datos pertenecientes a diversas fuentes en un documento XML y su visualización personalizada según el usuario.

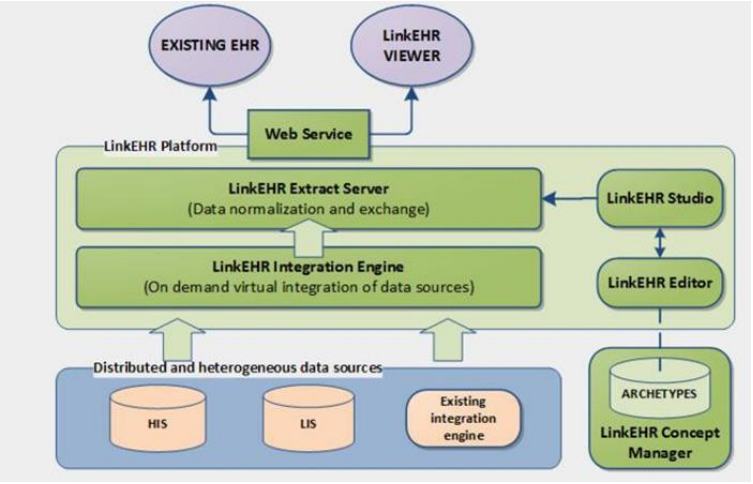

Figura 34. Estructura de la herramienta LinkEHR

- **LinkEHR Extract Server**. Basado en la parte 5 de la norma ISO 13606, es una plataforma de comunicación de extractos de la HCE. Junto con LinkEHR Viewer permite compartir y visualizar datos.

- **Repositorio de arquetipos del Ministerio de Sanidad, Servicios Sociales e Igualdad (MSSSI)**. El MSSSI de España ha creado un repositorio de recursos de modelado clínico que se pueden extraer en formato ADL (lenguaje de definición de arquetipos). Para posibilitar la mayor cantidad de combinaciones posible, los arquetipos se han diseñado de forma modular [68].

Estructura de los datos

El modelo dual, vinculado a la norma ISO 13606, indica como separar "información" (definida en la norma como los hechos que no varían en el tiempo, como el valor de glucosa en un análisis de sangre realizado en una fecha determinada) de "conocimiento" (que se define como los hechos que

sí varían, como las nuevas interpretaciones que podrían deducirse de esa prueba diagnóstica con los avances científicos) [1] [69].

La estructura de datos de cualquier herramienta basada en la norma partirá, por lo tanto, del modelo dual, que permite establecer cierta flexibilidad para posibilitar la evolución del sistema de información.

Para describir la estructura de la información y una semántica mínima, la ISO redactó las pautas del Modelo de Referencia (parte 1), mientras que, para la documentación del conocimiento, se creó el Modelo de Arquetipos (parte 2). A continuación, se describen ambos.

Modelo de referencia

El modelo de referencia [60] que podemos ver en la Figura 35 (para simplificar, se muestra sin atributos) aporta una estructura y un significado genéricos a los datos, definiendo su organización en clases (objetos), la información de contexto necesaria en cada clase y sus características generales. Su finalidad es que el significado de los datos clínicos (que son instancias del modelo) sea lo más claro posible al intercambiar HCE entre distintos sistemas.

En el modelo, cada clase pertenece a distintos niveles de una jerarquía en la que nodos padre e hijo se relacionan. Estas clases son: *folder, composition, section, entry,* cluster y *element.* Además, se define un tipo especial de objetos: los abstractos, que no se utilizan específicamente para almacenar u organizar datos dentro de la jerarquía, pero que aportan atributos al resto de clases. Las clases abstractas son: *Record_Component*, *content* e *ítem.* A continuación, se definen cada una de las clases del modelo:

- ***Extract*** es el componente más elevado en la jerarquía y abarca

[1] En este capítulo 2.3.1 se mantienen estas definiciones de "información" y "conocimiento" con el objetivo de seguir a la nomenclatura definida por la norma ISO13606.

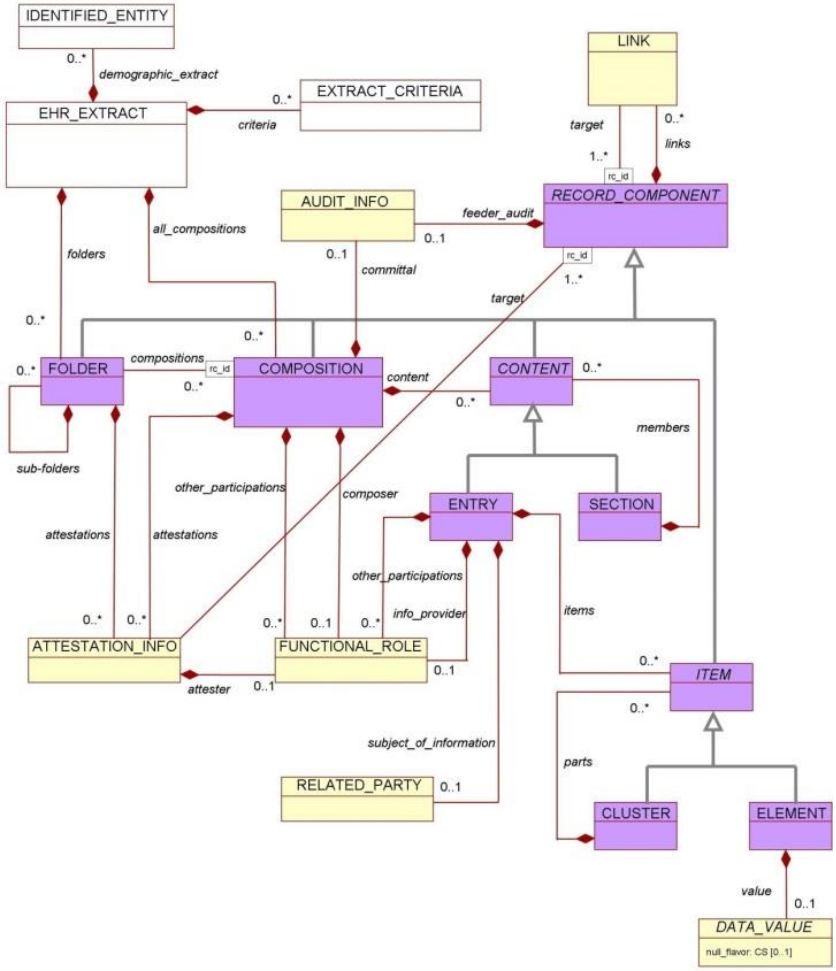

Figura 35. Modelo de Referencia resumido del estándar ISO 13606

esencialmente *compositions*. De forma adicional, puede contener *folders* e información demográfica acerca de cada persona, organización, dispositivo o software relacionado con el contenido del extracto. En general, se trata del nivel con parte de o toda la historia clínica de un paciente que se intercambia entre un sistema proveedor y otro receptor de información.

- ***Record_Component*** es el objeto abstracto más elevado en la jerarquía. Los componentes que se sitúan por debajo (*folder, composition, content, section, entry, ítem, entry, elemen*), heredan

de él sus atributos.

- ***Folder*** es opcional y consiste en un nivel de organización superior para agrupar la información según diferentes criterios como episodios clínicos, períodos de tiempo, organización o servicio... Está contenido en el *extract* y como atributos puede tener otras *folders*, referencias a composiciones y confirmaciones que pertenezcan a ella o a su contenido.
- ***Composition*** contiene la información creada durante una sesión clínica por un profesional de la salud. Ejemplos de *compositions* son la historia clínica resumida o el informe de resultados de las pruebas de laboratorio. Las *compositions* se pueden agrupar en *folders* y como atributos poseen *sections* y *entrie*s.
- ***Content*** es el objeto abstracto que contiene a los objetos *section* y *entry.*
- ***Section*** es el nivel de organización opcional intermedio que recoge datos que se encuentran bajo un mismo encabezamiento, procedentes de un proceso de captura de información o que facilita su comprensión. Ejemplos son alergias, examen abdominal o

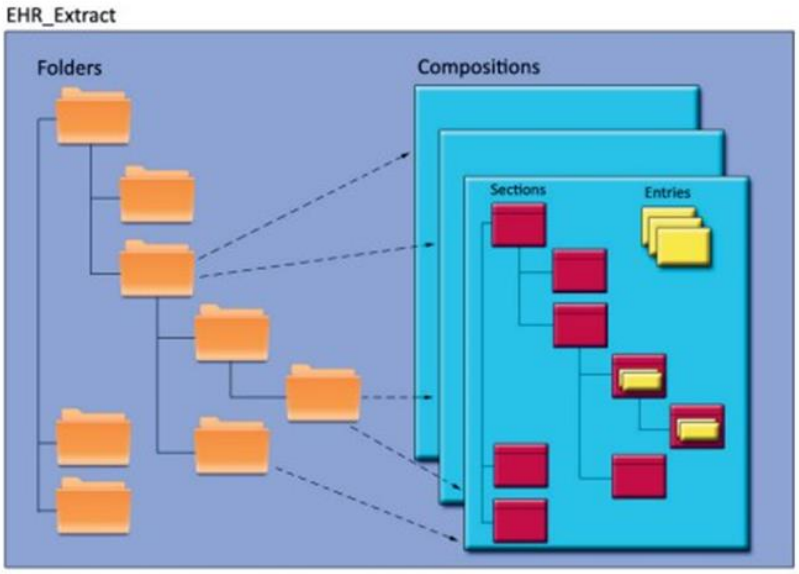

Figura 36. Representación gráfica de la jerarquía de clases del Modelo de Referencia

síntomas. Además de contener los atributos heredados por los objetos abstracto, abarca *entry*s o *sections*.

- ***Entry*** contiene hechos clínicos registrados (observaciones, diagnósticos, prescripciones) y se trata de la unidad mínima con significado clínico. De ella pueden surgir tanto *element* como *clusters*.
- ***Item*** es el objeto padre de los objetos *cluster* y *element*.
- ***Cluster*** es opcional y, al igual que *sections* y *folders*, consiste en un nivel organizador utilizado para organizar entradas múltiples en forma de series temporales, tablas o listas de datos. Permite agrupar múltiples *elements* o *cluster*s dentro de una *entry*.
- ***Element*** es el contenedor para los tipos de datos primitivos, como puede ser una cadena de texto, un número entero o un elemento multimedia.

En la Figura 36 vemos representada de una forma gráfica más sencilla la jerarquía de clases que acabamos de definir.

Modelo de arquetipos

Las clases del modelo de referencia permiten modelar conceptos organizándose en estructuras de mayor riqueza semántica. Estas estructuras de clases definidas en situaciones clínicas concretas se denominan arquetipos y expresan el conocimiento dentro del modelo dual [61]. A su vez, constituyen las instancias del modelo de arquetipos.

Los arquetipos se caracterizan, además de por su organización de los objetos, por restringir las características de cada posible dato en el modelo. Algunas cualidades que se pueden establecer para cada nodo del arquetipo son:

- Tipo de clase en el modelo de referencia: folder, composition, section, entry, cluster o element.
- Identificación de cada nuevo nodo del arquetipo. Puede referirse a nodos que ya existen en otros arquetipos o crearse explícitamente

en el arquetipo.

- Rango con el número de ocurrencias que se pueden instanciar. Si son múltiples las instancias que se pueden realizar, se podrá indicar si deben estar ordenadas o si cada instancia debe ser única.
- Valores permitidos a atributos y datos de cada nodo. Además, se pueden relacionar los valores permitidos entre diferentes nodos, es decir, referir si un nodo admite los mismos valores que otro nodo.
- Obligatorio u opcional.
- Cada nodo debe asociarse a un término/s clínico/s que indiquen el concepto al que se puede referir cualquier instancia del nodo.

Como el conocimiento es cambiante, estas estructuras pueden ir modificándose o sustituyéndose. También pueden reutilizarse para componer nuevos arquetipos o versionar los existentes.

Los arquetipos se pueden definir con el modelo de objetos de arquetipos (AOM) y el lenguaje de definición de arquetipos (ADL) (ambos definidos en [61]), dando lugar a arquetipos como el del siguiente ejemplo. La Figura 37 representa el arquetipo "Historia Clínica Resumida" del MSSSI [69]. Se

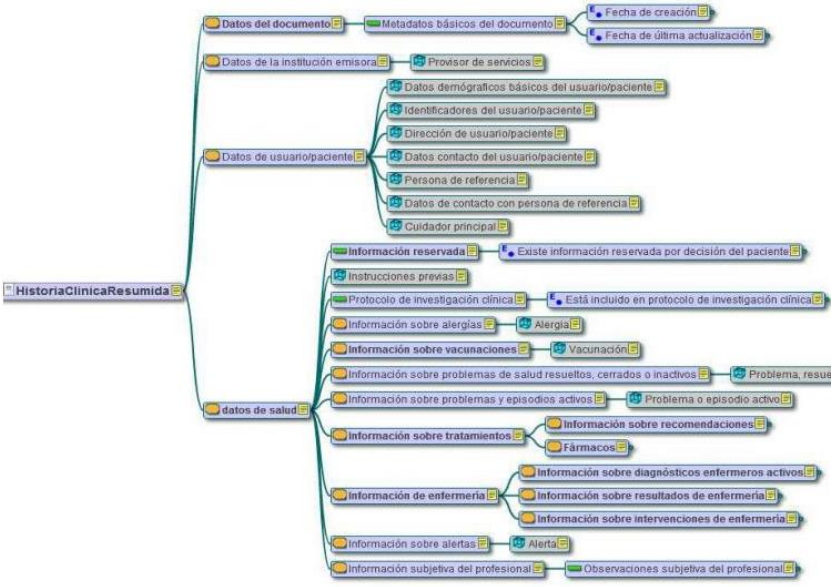

Figura 37. Arquetipo Historia Clínica Resumida creado por el MSSSI

trata de un arquetipo que documenta el conocimiento sobre el área referida por su título.

Análisis del interés del estándar

La norma ISO 13606 resulta interesante en primer lugar, porque se trata de una norma que pretende regularizar la HCE para permitir la interoperabilidad de datos entre diferentes sistemas. Esto permitiría que un paciente pudiera tener su HCE distribuida en distintos centros de atención sanitaria o incluso entre distintos servicios especializados.

En segundo lugar, para permitir la interoperabilidad, la norma ISO 13606 utiliza el modelo dual. Esta separación de información y conocimiento es muy interesante en el uso clínico, ya que es un campo de saber en constante evolución donde tanto la organización de los datos como la necesidad de cambio en los modelos de información son necesarios para una atención eficiente al paciente.

En tercer lugar, y también relacionado con la versatilidad, la capacidad de versionar, combinar y especializar arquetipos permite adaptar los modelos a cada necesidad concreta de información.

En cuanto a fiabilidad y seguridad, en su parte 4, la norma aporta un marco básico con directrices sobre seguridad.

2.3.2. Bases de datos para la investigación traslacional

Teniendo en cuenta que compartir la información de los estudios e investigaciones llevadas a cabo por los médicos e investigadores es un factor clave para avanzar en el conocimiento de las enfermedades y en la mejora de su diagnóstico y tratamiento, surge la necesidad de crear bases de datos de acceso libre donde poder compartir todos esos datos que se van generando y facilitar, de esta manera, la investigación traslacional. En este apartado vamos a comentar dos casos publicados de esta índole.

Base de datos del cáncer colorrectal

Esta base de datos fue presentada en 2011 en la 4ª Conferencia Internacional de Ingeniería e Informática Biomédica (BMEI), siendo elaborada por la Universidad de Zhejiang y el hospital *Sir Run Run Shaw*, ambos situados en Hangzhou (China) [70].

La finalidad del proyecto consistía en crear una *plataforma de investigación traslacional* relativa al cáncer colorrectal; es decir, una herramienta que integrase los datos provenientes tanto de la investigación como del ámbito clínico para obtener una mayor información de la patología, y permitiese traducir los resultados de investigación a la práctica clínica. Los usuarios de la herramienta serían, por lo tanto, especialistas médicos y biólogos moleculares, entre otros.

Los datos que almacena la herramienta son variados: ómicos, clínicos y epidemiológicos. Se han escogido estos tres tipos de datos para relacionar y estudiar el genotipo, el fenotipo y los factores de riesgo (ambientales, estilo de vida, antecedentes familiares relativos al cáncer...), considerándose importantes para el diagnóstico y tratamiento del cáncer, ya que su causa es multifactorial.

Los datos se organizaron es un modelo conceptual mixto Entidad- Relación y Entidad-Atributo-Valor atendiendo a la naturaleza cambiante (necesidad de añadir nuevos atributos) y escasa (no todos los campos se pueden llenar ya que dependen del estadio del cáncer) de la información. De esta manera, el modelo entidad- atributo- valor responde a la dinamicidad de los datos biomédicos (se pueden crear nuevos atributos sin cambiar el esquema conceptual), mientras que el modelo entidad- relación permite almacenar los datos de una manera similar a como se visualizan y facilita la consulta de atributos.

Cabe destacar que para el caso de los datos epidemiológicos se utilizó únicamente un modelo entidad-relación debido a la naturaleza estable de dichos datos.

Las funciones que permite la herramienta son consulta, visualización integrada y análisis de datos. La consulta puede ser básica (mediante palabras clave) o avanzada (estableciendo criterios de búsqueda), ordenándose los resultados en función del identificador del paciente. La visualización de los resultados se puede personalizar y, además, la herramienta posee una función de navegador para realizar análisis estadísticos y presentar los datos en forma de diagramas circulares. A partir de estas representaciones el usuario puede acceder a los datos de los pacientes pertenecientes a cada categoría a partir de un árbol jerárquico.

Observaciones

Este sistema de información fue presentado como la fase inicial del proyecto, es decir, aún no cumple las especificaciones más avanzadas y se desea resolver el problema de la inestabilidad del modelo de datos. En el artículo se expresa la voluntad de añadir en el futuro bases de datos biomédicas (literatura y biología molecular) y mejorar el análisis de la información, así como finalizar el desarrollo de cuestionarios online para recopilar datos epidemiológicos.

Sistema de información para el cáncer de próstata

Esta herramienta, creada por *L'Hospital Clinic Thonon* (Francia), el *IPA S.A. R&D Institute (*Cluj-Napoca, Rumanía) y la *Iuliu Hatieganu University of Medicine and Pharmacy (*Cluj-Napoca, Rumanía), tiene como objetivo general optimizar la gestión y análisis de los parámetros utilizados en la detección precoz del cáncer de próstata [71]. Para ello, se complementan una base de datos y una función que procesa y analiza las imágenes por ultrasonidos, aunque para este trabajo sólo nos concierne la base de datos. Para el diseño de la base de datos, se procuró que el DBMS permitiera el análisis de los datos por medios estadísticos.

Este sistema de información ha sido diseñado para incluir distintos tipos de datos:

- Información clínica, entre la que destaca la edad del paciente y los

resultados del examen rectal digital.

- Información biológica. En esta sección, resultan importantes los niveles de PSA.
- Información histológica, siendo relevante la escala de Gleason
- Características de las imágenes por ultrasonidos. Entre ellas encontramos parámetros como el tamaño del tumor o anomalías.

Todos ellos son utilizados en el diagnóstico, seguimiento y planificación de la terapia.

La base de datos en la que gestionan los parámetros es relacional y se desarrolló en el marco del proyecto PROSON, siendo implementada en Microsoft SQL Server 2000. Sus funciones son la gestión de pacientes y médicos, de exámenes médicos y la visualización integrada de datos (por ejemplo, observar para un mismo período de tiempo la imagen de ultrasonidos y los datos clínicos correspondientes).

2.3.3. Registro del cáncer

A pesar de que no se usa como sistema de gestión de la información en la práctica clínica, los registros del cáncer son relevantes para realizar estudios de salud pública orientados a esta patología. Según el Programa Nacional de Registros del Cáncer (NPCR, EEUU), un registro del cáncer "recopila información sobre pacientes con cáncer y los tratamientos que reciben, para almacenarlos en una computadora" con la finalidad de relacionar, por ejemplo, grupos de personas con determinadas características (grupos de riesgo en la misma área geográfica o con similares hábitos de vida) con la incidencia del cáncer. Gracias a las correlaciones obtenidas se pueden establecer políticas de prevención y mejorar los sistemas de diagnóstico y tratamiento.

A continuación, como ejemplo, se analizará brevemente el sistema de información del registro danés del cáncer.

Registro danés del cáncer

El registro danés del cáncer fue fundado en 1942 por el doctor Johannes Clemmensen, siendo actualmente gestionado por la Sociedad Danesa del Cáncer [72].

El motivo de su creación fue la recopilación sistemática de datos relativos al cáncer, mientras que sus objetivos son proveer de datos a la investigación y a las estadísticas para estudiar la prevalencia del cáncer y planear la provisión de servicios de salud daneses.

La función que tiene es la de aportar datos para investigar las causas y el transcurso del cáncer y, para ello, utiliza un registro poblacional de las incidencias de neoplasmas malignos y ciertas lesiones cancerígenas y benignas, con características del paciente en la fecha del diagnóstico.

El registro danés del cáncer se integra con el registro nacional danés del paciente y el registro danés de causas de muerte.

No se ha encontrado información acerca de la estructura de la base de datos (relaciones entre los diferentes campos), pero sí sobre los datos almacenados. Estos se pueden dividir tal y como se muestra en la Tabla 2.

Características personales	Características tumorales
Identificador personal.	Diagnóstico en el código ICD-10
Sexo	Diagnóstico en el código ICD-7 modificado.
Edad en el diagnóstico	Morfología
Municipalidad	Topografía
Región	Comportamiento
County	Lateralidad
Fecha de nacimiento	Estadio
Fecha de fallecimiento o emigración	Grado
	Datos de diagnóstico
	Base para el diagnóstico
	Número de tumor

Tabla 2. Datos incluidos en el Registro Danés del Cáncer

2.3.4. Orion Clinic

Orion Clinic es un sistema de información para la gestión de la atención hospitalaria, creado para la Agencia Valenciana de Salud (AVS) por la empresa Everis. El proyecto comenzó en 2006 [16, 73].

El motivo de su creación fue que la AVS se encontró con que sus sistemas de gestión de la atención especializada eran variados, basados en tecnologías desfasadas y no permitían cubrir las necesidades demandadas, por lo que se inició el desarrollo de una herramienta especializada con el fin de implantarla en todos los centros de la comunidad [73].

Los objetivos incluyen soportar la asistencia sanitaria en centros de atención hospitalaria, mejora de la eficacia del sistema, capacidad de evolución e integración y, sencillez y ergonomía del sistema [17].

Entre sus funcionalidades se encuentran [74]:

- Asistenciales. Propias de esta área son las funciones administrativas, de atención social y rehabilitación. En conjunto

Figura 38. Captura de la pantalla de entrada de Orion Clinic

con el ámbito clínico, se soportan las urgencias, la hospitalización, las agendas de actividad, el hospital de día, el bloque quirúrgico y de intervencionismo, los partos y al nenonato sano.

- Relacionadas con la farmacia. Propias de este sector son las acciones de validación de la prescripción, la elaboración, la atención farmacéutica y la dispensación. Relacionadas con el ámbito clínico, se definen las funciones de prescripción y administración de fármacos.
- Documentación. Orion Clinic permite el archivo, codificación y la creación de informes hospitalarios.
- Gestión de solicitudes administrativas, de laboratorio, morfológicas, de procedimientos y de consultas.
- Prevención y seguridad clínicas.
- Farmacovigilancia.
- Explotación de los datos.
- Identificación del paciente
- Programación y citación.
- Gestión del paciente.
- Organizador.
- Captura de actividad.
- Eventos y alertas clínicas

Como en la AVS subsisten diferentes sistemas de información, es necesaria la integración de Orion Clinic con ellos. Algún ejemplo de estos sistemas son Abucasis [75] (sistema de información de atención primaria de la comunidad) y el de prescripción farmacéutica [74].

En cuanto a la información que almacena es fundamentalmente sanitaria y orientada al paciente en todos sus episodios de salud en centros de atención secundaria. Abarca, por lo tanto, desde patologías y tratamientos hasta información de contacto o imágenes para el diagnóstico y seguimiento.

Las consultas de datos se realizan accediendo a la historia clínica de cada paciente según su número de historia, identificación SIP u otro dato

identificativo. En cuanto a la introducción de información, esta se realiza generalmente mediante la escritura de texto, existiendo un diccionario de términos que se pueden seleccionar. Además, en lo relativo a datos pertenecientes a la toma de variables fisiológicas, estos se podrán volcar directamente de los dispositivos y sensores al sistema de información en función del área y del aparato.

Tanto para la consulta como para la introducción de información, será necesario verificar la identidad del usuario.

Centrándonos en su estructura, la herramienta contiene sendas facetas orientadas a:

- Profesional. Contiene las listas de trabajos del profesional, así como las del servicio y área.
- Paciente. Abarca los episodios de salud ocurridos en los centros de atención secundaria que utilizan Orion Clinic, así como los datos provenientes de los sistemas con los que interoperan. Además, permite llevar a cabo la prescripción electrónica y documentar mediante informes la mayoría de hechos clínicos.

Por otro lado, cabe destacar que para cada servicio y rol existe una interfaz de usuario concreta, por lo que no todo el personal puede acceder a todas las funcionalidades de Orion Clinic. En la Figura 38, se observa la pantalla de entrada al sistema con los apartados urgencias, hospitalización, consultas y técnicas, farmacia, admisión, archivo y codificación, hospitalización a domicilio, obstetricia, bloque quirúrgico, trabajo social, prevención y seguridad y explotación de la información.

Como ventajas, Orion Clinic presenta un diseño orientado al sistema de salud de la Comunidad Valenciana, adecuándose a las necesidades de información y de interoperabilidad de esta área. Sin embargo, este enfoque resulta inadecuado para comunicar y reunir la historia clínica electrónica de un paciente a nivel nacional, ya que resulta complicada la interoperabilidad entre éste y otros sistemas de la red nacional. Además, a

pesar de que se planteó para intercambiar información con el sistema de información de atención primaria de la comunidad (Abucasis), esta acción no es posible aún en su totalidad. Orión permite tener acceso a Abucasis desde un enlace para consultar información o realizar prescripciones médicas, pero gestionando la información de manera totalmente independiente. Haciendo referencia también a la estandarización, Orion Clinic induce al personal a introducir la información en campos de texto libre que, durante el intercambio de información puede ser difícil de interpretar.

Otros aspectos positivos del sistema de información son su diseño orientado al usuario, teniendo un contacto estrecho y tratando los problemas que pueda tener el personal; su evolución continua, publicando nuevas versiones; y el énfasis en la seguridad y confidencialidad.

Como inconvenientes, cabe destacar que aún no ha sido implantado en todos los hospitales de la Comunidad Valenciana y que ello contribuye a crear nichos de información y a dificultar la movilidad del paciente.

2.3.5. Conclusiones

Habiendo analizado distintas soluciones planteadas por la comunidad científica para solucionar el problema de la gestión de datos clínicos y de investigación, detectamos ciertas características relevantes.

La solución propuesta por la norma ISO 13606 resulta interesante porque se trata de una norma que pretende regularizar la HCE para permitir la interoperabilidad de datos entre diferentes sistemas. Plantea esta solución para generar informes clínicos desde un conjunto de modelos que sirven de base para un sistema configurable, adaptable a cada entorno clínico, seguro e interoperable entre todos aquellos sistemas diseñados según la misma norma. Otro aspecto a tener en cuenta en este sistema es que los modelos planteados, principalmente el modelo de arquetipos, están diseñados para almacenar información clínica en formato de informes clínicos, con sus cabeceras y sus contenidos, lo que lo hace adaptable a todos los dominios,

pero no específico para relacionar los datos conceptualmente y, por tanto, es apropiado para un correcto almacenaje de los datos, pero no tanto para la realización de análisis de los mismos en un entorno investigador, ya que no están relacionados entre sí conceptualmente.

En este estudio, también nos encontramos con un par de ejemplos de bases de datos diseñadas para la investigación traslacional en dos enfermedades: el cáncer colorrectal y el cáncer de próstata. Hay poca información disponible sobre ellas, por lo que podemos deducir que ambos proyectos están en fase de desarrollo o se han quedado sin desarrollarse totalmente. La base de datos de cáncer colorrectal fue presentada como la fase inicial del proyecto, es decir, aún no cumple las especificaciones más avanzadas planteadas inicialmente y sigue en fase de desarrollo. En este caso, sí que comentan la existencia de un modelo conceptual mixto Entidad-Relación y Entidad-Atributo-Valor para asegurar la consistencia de los datos, aunque no se encuentran resultados acerca del proyecto en su totalidad. Algo similar ocurre con el de cáncer de próstata. Se comenta en el artículo que la base de datos planteada ha sido diseñada utilizando un modelo Entidad-Relación, integrando información sobre la gestión de pacientes y médicos, de exámenes médicos y la visualización integrada de datos.

Otro ejemplo encontrado que gestiona este tipo de información es el Registro danés del cáncer. A pesar de que no se usa como sistema de gestión de la información en la práctica clínica, los registros del cáncer son relevantes para realizar estudios de salud pública orientados a esta patología. No se centra en un cáncer en concreto, sino que recoge en una base de datos información general sobre todos los tipos de cáncer. Sin embargo, no se ha encontrado información acerca del modelo de base de datos que sigue, pero sí sobre los datos que almacena.

Finalmente, la última de las herramientas de este estudio es la que se encuentra implementada en la mayoría de hospitales de la Agencia Valenciana de Salud: Orion Clinic. Esta herramienta fue diseñada adecuándose a las necesidades de información y de interoperabilidad de esta área. El modelo en el que se basa esta herramienta no está disponible

públicamente. Sin embargo, viendo el funcionamiento de la misma nos damos cuenta de que induce al personal a introducir texto aleatorio que, durante el intercambio de información puede ser difícil de interpretar, y conlleva muy poca especificidad de datos en su base de datos. Cabe destacar su diseño orientado al usuario, su evolución continua, y el énfasis en la seguridad y confidencialidad.

Viendo estos sistemas, detectamos la creciente necesidad de gestión de datos que se está dando en el sector médico, y el planteamiento de soluciones varias que se están ofreciendo, no siempre desde una correcta perspectiva de los Sistemas de Información. La falta de información sobre los esquemas conceptuales en los que se basan estos sistemas podría ser un indicador de la carencia de los mismos en el diseño del sistema de información.

Destacaríamos como conclusión que es evidente que existe una "asimetría conceptual" entre los intentos rigurosos de elaborar estándares de uso internacional (que no consiguen cruzar de una manera convincente la frontera del "ámbito teórico"), con las herramientas usadas en la práctica que hacen usable y gestionable toda la información clínica, pero de una manera mucho más desestructurada y desnormalizada. La existencia de fuentes independientes de información especializada para patologías individuales complica el escenario analizado y la posición de este trabajo de Tesis: hacer de los modelos conceptuales las componentes esenciales de los Sistemas de Información clínica que queremos instrumentalizar.

2.4. Justificación de la utilización de técnicas de Modelado Conceptual en el entorno clínico y biológico

El tratamiento eficaz de enfermedades de un perfil tan complejo como el Cáncer de Mama -o cualquier otro Cáncer- exige, actualmente, disponer de sistemas informáticos que permitan integrar y gestionar eficientemente toda la información relevante sobre esta enfermedad. La finalidad de esta tesis es demostrar que el uso adecuado de las buenas prácticas de la Ingeniería Avanzada en Sistemas de Información en el dominio del Cáncer de Mama, y en concreto el uso de técnicas de Modelado Conceptual, mejorará de una forma significativa los entornos actuales de trabajo de los profesionales dedicados al estudio y al tratamiento de esta enfermedad. Al hablar de mejoras significativas nos estamos refiriendo a mejorar tanto el diagnóstico como el tratamiento de los pacientes, a establecer perfiles específicos de pacientes, a determinar los tratamientos más apropiados para cada uno de estos perfiles, a generar nuevo conocimiento a partir de estudios sobre expresión de microARNs o estudios genómicos, a detectar nuevos marcadores genéticos, nuevas vías de señalización involucradas en este tipo tumoral y a diseñar estudios funcionales encaminados a mejorar las terapias existentes. En definitiva, solo si se dispone de un Sistema de Información diseñado para manipular de forma eficiente y efectiva toda la información relevante sobre esta enfermedad, se podrá avanzar de una manera continua y ordenada en el descubrimiento de todas las particularidades inherentes al Cáncer de Mama, y a su prevención, diagnóstico y tratamiento.

Abordar el estudio del Cáncer de Mama en su totalidad es una tarea ardua, por lo que es necesario delimitar el alcance de esta propuesta. La construcción de este sistema permitirá la integración de grandes volúmenes de datos de gran valor y de alto coste de obtención, como los resultantes de dichos trabajos, así como el desarrollo de procedimientos de explotación de estos datos que incidirán directamente en la generación de nuevos

resultados que mejorarán a su vez, el trabajo asistencial en este campo. La prevención, el diagnóstico precoz, el descubrimiento de tratamientos novedosos o de interrelaciones significativas con distintos tipos de cánceres se facilitará con el uso de este tipo de sistemas.

Este trabajo es esencialmente multidisciplinar. Los expertos en Sistemas de Información pueden diseñar y construir sistemas de gestión y explotación de datos altamente eficientes, mediante aproximaciones tecnológicas de amplia difusión como las plataformas "Big Data", los Almacenes de Datos, las herramientas de Minería de Datos, o las técnicas de Modelado Conceptual. Sin embargo, el contexto en el que se aplican estas técnicas requiere de un conocimiento adicional desde una perspectiva médica para poder diseñar un sistema que se adapte perfectamente a los datos médicos y biológicos. Este es el papel del Ingeniero Biomédico en este entorno multidisciplinar, gracias al cual puede plantearse de forma realista el reto de diseñar e implementar un Entorno para la Gestión Avanzada de datos asociados al Cáncer de Mama.

Como hemos visto en este capítulo, los antecedentes de proyectos que han abordado este problema raramente aportan esta doble visión. Es cierto que son muchos los trabajos relacionados con datos derivados de estudios clínicos, biológicos y genómicos, y que es mucha la información que se genera hoy en día. Pero es también habitual la carencia de un tratamiento de calidad desde el punto de vista "sistémico" de dichos datos. Es frecuente que los Sistemas de Información soporte de estos datos se limiten al uso de herramientas más bien convencionales como hojas de cálculo, paquetes estadísticos y otros programas software que con demasiada frecuencia no son las más adecuados para su tratamiento. Se echa en falta inmediatamente la utilización de técnicas de modelado conceptual para analizar y determinar qué datos son realmente significativos, el uso de técnicas de transformación de los esquemas conceptuales en bases de datos que aseguren una explotación eficiente (sea relacional, objeto-relacional, bases de datos No-SQL), plataformas de carga de datos fiables, plataformas de consulta adaptables a los intereses de los usuarios, estrategias de evolución

de esquemas que hagan posible la incorporación flexible de nuevo conocimiento, etc.

Es llamativa la falta de esa visión global que estamos llamando sistémica, y que hace referencia a la necesidad de abordar el problema desde una perspectiva multidisciplinar. Esta es la aportación esencial de esta tesis. En [76] se presentaba una cuestión que está en el centro de esta tesis: ¿"Big Data or Right Data"? No es un problema de almacenar una cantidad ingente de información de todo tipo sobre el fenómeno analizado (the "Big Data"), sino más bien un problema de saber qué información es realmente relevante (the "Right Data"), entender su semántica, diseñar un sistema de información adaptado a sus peculiaridades, y poder así disponer de una plataforma software que haga posible los objetivos comentados al principio de mejorar drásticamente la prevención, el diagnóstico y el tratamiento de la enfermedad estudiada de acuerdo con el proceso de los datos almacenados.

Respecto a por qué elegir el dominio del cáncer de mama, como se ha comentado en la motivación de esta tesis, huelga decir que el cáncer de mama es el tumor más frecuente en la mujer. Representando un 29% de los cánceres registrados en mujeres, el cáncer de mama es el tipo de cáncer con mayor incidencia en las mujeres españolas. La búsqueda de soluciones para reducir su frecuencia y severidad y mejorar la calidad de vida de los pacientes que sufren cáncer de mama es una prioridad del sistema público de salud y uno de los asuntos que más preocupan a la sociedad. Esta alta incidencia de la enfermedad en los últimos años y el interés de los sistemas públicos de salud en invertir en el estudio de esta enfermedad hace que la cantidad de información y datos referentes al cáncer de mama, tanto al tratamiento como a la prevención de la enfermedad, crezca exponencialmente día a día.

El hecho de poder trabajar con el grupo de la Dra. Gloria Ribas, perteneciente al Servicio de Oncología Médica y Hematología del Instituto de Investigación Sanitaria INCLIVA y especializado en el estudio de este problema, ha proporcionado el contexto ideal para plantear esta tesis

centrada en el cáncer de mama. Conforme a lo dicho anteriormente, partiendo de una Ingeniería Avanzada de Sistemas de Información, el primer punto importante a tener en cuenta son los estudios clínicos, biológicos y genómicos que se llevan a cabo actualmente y que permiten obtener datos referentes a las pacientes para realizar los análisis correspondientes. El siguiente punto será determinar qué tecnologías se van a aplicar sobre estos datos con el fin de extraer conclusiones relevantes.

Los datos clínicos de los pacientes son una fuente de información muy valiosa. En ellos se almacenan todos aquellos datos obtenidos en la consulta del médico o procedentes de análisis previos del paciente, necesarios para que el médico pueda establecer un diagnóstico o aplicar un tratamiento. En relación con los datos clínicos recopilados en casos de Cáncer de Mama, datos como los marcadores encontrados, el tratamiento, cómo ha reaccionado la paciente al tratamiento administrado, y la disponibilidad del tumor en los bancos de tejidos, son datos que nos pueden servir para relacionarlos con otros tipos de datos extrayendo conclusiones muy interesantes. Todos estos datos estarán incluidos en la Histórica Clínica Electrónica [77-79] del paciente (en el caso en el que el hospital posea este servicio), en la Historia Clínica clásica en formato papel, o incluso en algún sistema digital de almacenamiento de datos utilizado por los médicos para poder disponer de los datos en formato digital. En este último caso, la mayoría de los médicos que se aventuran a almacenar digitalmente los datos de sus pacientes utilizan -como señalábamos anteriormente- tecnologías convencionales como Microsoft Access o SPSS, que no han sido diseñadas para almacenar y gestionar información tan compleja y tan extensa [80]. Es cierto que resultan cómodas, sencillas y accesibles, y que su utilización para almacenar datos supone un avance a la hora de tener los datos integrados en un mismo soporte, tenerlos accesibles y poder efectuar comparaciones. Pero cuando se manejan volúmenes de datos muy grandes, interrelacionados, con necesidad de asegurar su integridad, consistencia, no redundancia y con la necesidad además de poder inferir nuevas informaciones a partir de ellos, esas herramientas convencionales se quedan muy cortas. Se hace imprescindible diseñar, desarrollar y utilizar

plataformas software avanzadas de gestión de datos que hagan posible -por ejemplo- la utilización efectiva de la HCE o de un sistema de información especialmente adaptado a sus necesidades para almacenar y gestionar los datos.

Otra fuente de datos que el Sistema de información resultante de esta tesis tiene que tener en cuenta afecta a los datos procedentes de los análisis genómicos de los pacientes realizados por biólogos moleculares. Estos estudios pueden llevarse a cabo a partir de análisis de expresión de microARNs o secuenciación genómica, entre otros. Todos los datos de estos estudios proceden en primera instancia de secuenciadores de nueva generación, que en un principio generan una gran cantidad de datos de difícil manejo, que necesita ser pre-procesada utilizando distintos algoritmos y herramientas específicos, que permitirán limpiar los datos de la muestra secuenciada extrayendo la información de interés antes de poder mostrársela al usuario para su interpretación [81, 82]. Esta información junto al resto de datos que se han ido recopilando mediante las diferentes técnicas, tratados conjuntamente y adecuadamente relacionados, pueden ayudarnos a extraer conclusiones relevantes y muy interesantes para poder avanzar en nuestra investigación.

En definitiva, toda esta gran cantidad de valiosa información se engloba en un contexto de manipulación de datos en el que sería altamente beneficioso disponer de un Sistema de Información correctamente estructurado y diseñado específicamente para gestionar de manera efectiva y eficiente tal información. Una alternativa es potenciar la aplicación de las técnicas de modelado conceptual en el ámbito de la información genética [83-87]. Actualmente la mayoría de las herramientas bioinformáticas que se utilizan en los diversos laboratorios genéticos han sido creadas con el fin de cubrir las necesidades que iban surgiendo de una manera ad-hoc, desestructurada.

Los Sistemas de Información Genéticos resultantes de aplicar las buenas prácticas propias de la teoría de Sistemas de Información y el Modelado Conceptual, con los Esquemas Conceptuales diseñados en esta tesis, permitirán explotar de forma correcta la información almacenada [88, 89].

Además, harían posible encontrar otras relaciones entre los datos clínicos y/o moleculares almacenados y la evolución de la enfermedad, muy útiles para el diagnóstico y tratamiento de las pacientes [90-92].

El diseño e implementación de Sistemas de Información a partir de modelos conceptuales y la implantación de los mismos en múltiples entornos donde funcionan de forma eficiente ha demostrado con creces desde hace décadas la imperiosa necesidad de diseñar los sistemas de información utilizando técnicas de modelado conceptual que aseguren la eficacia y la calidad de los datos [93-96]. Estos sistemas ayudan a los usuarios en la gestión de la información que manejan en su día a día de una forma mucho más eficiente. Además, estos sistemas de información permiten visualizar los datos desde una perspectiva holística, lo que en entornos sanitarios debería ser indispensable a la hora de trabajar con datos sobre enfermedades con alteraciones genómicas. Estos aspectos resaltan la importancia de esta tesis y refuerzan la gran utilidad de estos sistemas en entornos sanitarios desde una perspectiva de investigación traslacional esencial en su planteamiento.

3. Diseño de la solución

3.1. Modelado Conceptual y diseño del Sistema de Información

Diseñar un Sistema de Información correcto requiere usar métodos y tecnologías del ámbito de la Ingeniería de los Sistemas de Información, con el fin de asegurar la calidad de los datos y su óptima gestión y explotación. El desarrollo de un Sistema de Información en el entorno sanitario tiene que gestionar datos complejos y heterogéneos. Es un dominio en el que la utilización de técnicas de modelado conceptual como herramienta básica de trabajo se convierte en una necesidad si se quiere garantizar la corrección del diseño del sistema de información y la eficiencia en la gestión de sus datos. El aspecto más novedoso de esta Tesis es justamente demostrar cómo esa utilización de Modelos Conceptuales hace posible la creación de Sistemas de Información efectivos y eficientes, en los que distintas perspectivas de información (clínica y genómica) quedan integradas de manera precisa, estructurada y holística.

Teniendo en cuenta la información clínica de pacientes de cáncer de mama que se manejan en la Unidad de Oncología de un hospital, los datos genéticos y biológicos de esas pacientes obtenidos en un Laboratorio de Biología Molecular, y los datos de secuenciación y análisis genómico de pacientes de cáncer de mama obtenidos en un laboratorio de genómica, se

plantea el objetivo general de diseñar y desarrollar un Sistema de Información para la manipulación eficiente y fiable de la información sobre el cáncer de mama en mujeres jóvenes.

Utilizando técnicas de Modelado Conceptual, se ha diseñado un Esquema Conceptual del Cáncer de Mama, donde se relacionan de manera correctamente estructurada los datos clínicos, de expresión génica y de secuenciación masiva en pacientes de Cáncer de Mama.

Este Esquema Conceptual, dividido en tres perspectivas o vistas, servirá de base para implementar el sistema de información necesario para almacenar, gestionar y explotar los datos sobre cáncer de mama de forma eficiente generando resultados originales y de máxima calidad.

3.1.1. Perspectiva Clínica

Esta primera perspectiva representa los datos clínicos que se manejan en una Unidad de Oncología de un hospital. En ella quedan representados datos del paciente, de su historial reproductivo anterior al cáncer, antecedentes, episodios clínicos por los que pasa, tratamientos, pruebas y tumores, en forma de clases y relaciones utilizando el lenguaje de modelado UML [97]. En este apartado, vamos a describir las clases y relaciones que se representan en esta primera vista del Modelo Conceptual del Cáncer de Mama. Esta vista queda dividida en dos partes que podemos encontrar en la Figura 39 y la Figura 40.

La clase **Paciente** es la clase principal de este esquema conceptual. Contiene toda la información referente a la persona que llega a la consulta con sospechas de padecer la enfermedad. Esta clase contiene el siguiente conjunto de atributos que describen al paciente:

- ***NHC***: Número de Historia Clínica del Paciente. Es el número de identificación de la Historia Clínica del Paciente en el Hospital.
- ***num_SIP***: Contiene el número que identifica al paciente en el Sistema de Identificación de Pacientes de la Agencia Valenciana de

Salud.

- *fecha_nacimiento*: Es la fecha de nacimiento del paciente, en formato fecha.
- *edad_al_dco*: Es la edad del paciente en el momento del diagnóstico de la enfermedad.
- *nacionalidad*: Indica la nacionalidad del paciente, muy útil para estudios de población o aplicación de tratamientos.
- *peso*: Contiene el peso del paciente en kilogramos.
- *talla*: Contiene la altura del paciente en metros.
- *índice_masa_corporal (IMC)*: Representa el índice de masa corporal del paciente. Se calcula con la fórmula peso/talla^2 y corresponde a un número con decimales.
- *dieta*: Contiene una descripción del tipo de dieta que suele tomar el paciente.
- *teléfono*: Representa el teléfono de contacto del paciente.

El paciente puede provenir de un servicio de un determinado centro médico. Con este propósito tenemos la relación con la clase Servicio y esta, a su vez, con Centro Médico.[2]

La clase **Centro Médico** representa a cualquier centro del que puede provenir un paciente o al que puede estar adscrito un médico o clínico. Está compuesta por Servicios y tiene los siguientes atributos:

- *id_centro*: Identificador interno que representa de forma única al centro
- *nombre*: Contiene el nombre del centro médico en cuestión
- *ciudad*: Contiene el nombre de la ciudad donde se ubica el centro

[2] Con la intención de que la lectura de este capítulo resulte más amena se ha omitido la cardinalidad mínima y máxima de las relaciones, las cuales se pueden identificar fácilmente en la vista del esquema conceptual correspondiente.

La clase **Servicio** representa a los servicios médicos que forman parte de un centro médico y a los que puede estar adscrito un médico o clínico. Está relacionado con Paciente y con Personal y tiene dos atributos:

- ***id_servicio***: Contiene un identificador interno único que representa al servicio
- ***nombre:*** Incluye el nombre del servicio en cuestión.

La clase **Personal** representa a todos los clínicos o médicos que forman parte de un servicio y que pueden ser responsables de un paciente o de la información incluida en un episodio. Contiene tres atributos:

- ***id***: Identificador interno
- ***usuario***: El nombre de usuario que se le dará al personal para identificarlo en la aplicación
- ***contraseña***: La contraseña que le permitirá el acceso.

El paciente, además, a su llegada al centro tiene un único historial reproductivo, representado por la clase **Vida Reproductiva**. Esta clase está representada por varios atributos:

- ***gestaciones***: Contiene el número de gestaciones que ha tenido la paciente
- ***abortos***: Contiene el número de gestaciones que no han acabado en parto
- ***partos***: Contiene el número de gestaciones que han acabado satisfactoriamente
- ***edad_primer_embarazo (EPE)***: Representa la edad a la que la paciente fue gestante por primera vez
- ***lactancia***: Indica los meses que la paciente ha pasado por un periodo de lactancia. Si ha pasado varias veces, almacenamos el periodo mayor.
- ***menarquia***: contiene la edad de la primera menstruación de la paciente
- ***menopausia (FUR)***: indica la fecha de la última menstruación en

el caso de pacientes premenopáusicas y la edad en caso de pacientes postmenopáusicas.

- *duración_terapia_hormonal_sustitutiva (THS)*: Contiene los años o meses que la paciente estuvo bajo una terapia hormonal sustitutiva.
- *número_tratamientos_fertilidad*: Contiene la cantidad de tratamientos de fertilidad por los que ha pasado la paciente

Cada vida reproductiva de un paciente puede estar asociada a muchas clases **Anticonceptivo** que representan los tratamientos anticonceptivos que ha utilizado la paciente durante su vida reproductiva. El atributo ***duración*** asociado a la relación, contiene los años o meses que la paciente ha estado utilizando este método. La clase Anticonceptivo se representa por:

- *nombre*: contiene el nombre del fármaco o tratamiento
- *tipo*: indica el tipo de tratamiento anticonceptivo (oral, anillo, …)

Además, cada paciente puede haber tenido o no algún familiar que haya sufrido algún tipo de cáncer. Esta característica queda representada por la clase **Antecedente Oncológico**. Tiene dos relaciones con la clase Paciente. La relación *paciente* representa que un paciente puede tener muchos antecedentes oncológicos, mientras la relación *familiar* asocia el Antecedente Oncológico con un paciente presente en la base de datos, si este paciente ha sido tratado en el servicio de oncología previamente. Esta clase se define por los siguientes atributos:

- *id_ant_familiar*: contiene un identificador interno de la base de datos para ese antecedente familiar.
- *parentesco*: indica cual es el parentesco que tiene ese paciente con su familiar (padre, madre, …)
- *características_enfermedad*: contiene una descripción de la enfermedad que ha sufrido el familiar.

La clase **Antecedente Médico** representa todas aquellas enfermedades que haya podido sufrir el paciente antes de la primera visita al Servicio de Oncología. Pueden tener un valor asociado, si es necesario, para indicar aspectos como la hipercolesterolemia o enfermedades similares, reflejado en el atributo ***valor*** de la relación. Esta clase queda representada por un atributo:

- ***tipo_antecedente_médico***: contendrá el nombre de la enfermedad que ha sufrido o sufre el paciente. Tendrá un valor fijo entre los siguientes: *Diabetes Mellitus (DM), Alergias (RAMs), Hipertensión Arterial (HTA), Dislipemia (DL), Otras comorbilidades.*

De forma similar a la clase anterior, la clase **Fármaco habitual** representa aquellos fármacos que el paciente toma diariamente por motivos varios. La dosis y la pauta que sigue el paciente al tomar dicho fármaco se indicará en el atributo ***posología*** de la relación. Esta clase viene definida por un atributo:

- ***nombre***: contiene el nombre del fármaco que toma el paciente.

La clase **Estado** representa el estado de salud del paciente. La relación entre Estado y Paciente tiene un atributo para indicar la ***fecha*** en la que se modificó el estado del paciente. La clase estado viene definida por un atributo:

- ***tipo***: contiene el estado del paciente. Este atributo tendrá un número fijo de posibilidades, con unas opciones predefinidas, entre las que se encuentran *Vivo, Vivo con enfermedad, Éxitus* o *Pérdida de seguimiento.*

La clase **Performance Status** representa el estado de actividad de la vida cotidiana del paciente. Al igual que la clase anterior, el atributo ***fecha*** de la relación de las clases Performance Status y Paciente contendrá la fecha en la que se define un nuevo estado. Esta clase está definida por un atributo:

- *tipo*: contiene el estado de actividad del paciente. Este atributo tendrá un valor numérico que va definido de 0 a 4 con el siguiente significado:
  - 0: normal
  - 1: cierta limitación/vida normal
  - 2: limitación más aguda/vida normal
  - 3: vida cama-sofá
  - 4: encamado

La clase **Hábito Tóxico** representa los distintos hábitos o adicciones que puede tener el paciente, como por ejemplo el consumo de tabaco o alcohol. Se relaciona con la clase Paciente, indicando el paciente que tiene dicho hábito tóxico. Está representada por dos atributos:

- *tipo*: contiene el tipo de hábito tóxico adquirido por el paciente.
- *descripción*: contiene una descripción del consumo o frecuencia con la que el paciente se administra dicho tóxico.

La clase **Episodio** contiene información sobre los eventos que tienen lugar en la historia clínica del paciente. Se relaciona con la clase Paciente, indicando el paciente al que hace referencia el episodio, con la clase Personal, indicando la persona que lleva al paciente en ese episodio, y con las clases Prueba, Tratamiento, Síntoma, Muestra y Metástasis, representando los distintos eventos que pueden tener lugar en ese episodio. Esta clase está definida por varios atributos:

- *id_episodio*: incluye un identificador interno que representa al episodio
- *fecha_episodio*: contiene la fecha en la que se inició el episodio
- *plan_actuación*: contiene una descripción de las pruebas o tratamientos que tiene que es aconsejable que lleve a cabo el paciente en el siguiente episodio.
- *estado_tratamiento*: indica el tipo de tratamiento que está llevando el paciente en ese episodio. Puede tener uno de estos cuatro

valores:
- En tratamiento curativo
- Libre de enfermedad
- Tratamiento quimioterápico/paliativo
- Control de síntomas

Esta clase se especializa en otras subclases:

- **Urgencias**: Esta clase representa aquellos eventos en los que el paciente llega al hospital a través del servicio de urgencias. Tiene un atributo específico:
  - ***motivo_consulta***: Contiene una descripción del motivo por el cual el paciente llegó a urgencias.

  Además, esta clase se especializa en otras dos subclases:
  - **Alta domiciliaria**: Representa aquellos casos de urgencia en los que el paciente sale del servicio de urgencias y vuelve a su casa. Contiene un atributo específico:
    - ***tratamiento_pautado***: Contiene una descripción del tratamiento que se le ha dado al paciente desde el servicio de urgencias.

  - **Ingreso**: Representa aquellos casos de urgencia en los que el paciente se queda ingresado en el centro hospitalario. Contiene un atributo específico:
    - ***causa_ingreso***: Contiene una descripción del motivo por el que el paciente ha quedado ingresado en el centro hospitalario.

- **Seguimiento**: Esta clase representa aquellos eventos en los que el paciente vuelve a la consulta para hacerse un seguimiento de la enfermedad. Contiene un atributo específico:
  - ***descripción***: Contiene una descripción del estado del paciente en el momento del seguimiento.

- **Diagnóstico**: Representa el evento en el que el paciente llega por primera vez a la consulta y se le diagnostica la enfermedad. Tiene varios atributos específicos que lo definen:
  - ***descripción_dco***: Contiene una descripción del diagnóstico inicial del paciente
  - ***gestante_al_dco***: Indica si la paciente estaba en periodo de gestación cuando se le detectó la enfermedad
  - ***lactancia_al_dco***: Indica si la paciente estaba en periodo de lactancia cuando se le diagnosticó la enfermedad
  - ***motivo_consulta***: Incluye el motivo por el que la paciente ha acudido a la consulta.
  - ***descr_motivo***: Incluye una descripción del motivo de la consulta indicado en el atributo anterior.

- **Recaída**: Representa el evento en el que el paciente llega a la consulta por una recaída de la enfermedad. Los atributos que lo definen son los siguientes:
  - ***lugar***: Incluye el lugar donde aparece el tumor de nuevo
  - ***descripción***: Contiene una descripción de la recaída.
  - ***ILE***: (Intervalo libre de enfermedad) Representa los años que el paciente ha pasado sin enfermedad desde que se le da el alta hasta la aparición de nuevos síntomas.

En un episodio de la vida de un paciente puede recibir varios tratamientos. Estos tratamientos quedan representados en la clase **Tratamiento**, la cual se define con dos atributos:

- ***id_tratamiento***: Contiene un identificador interno único.
- ***fecha_tratamiento***: Contiene la fecha en la que se le empezó a aplicar el tratamiento al paciente.

La clase Tratamiento se especializa en varias subclases dependiendo del tipo de tratamiento:

- **Cirugía**: Representa aquellas intervenciones a las que se somete al paciente para extirpar un tejido. Está relacionada con la clase Muestra (la cual se describe más adelante), que es la pieza de tejido que se extrae en la intervención. Está definida por un atributo:
  - ***resultado***: Contiene el resultado final de la intervención, indicando la cantidad de masa tumoral extirpada. Contendrá uno de estos tres valores:
    - **R1**: Se ha quitado todo
    - **R2**: Ha quedado algún resto
    - **R3**: No se ha quitado nada

Además, la clase cirugía puede especializarse en cuatro subclases:

  - **Conservadora**: Representa a aquellas intervenciones quirúrgicas en las que se biopsia una parte del pecho.
  - **Mastectomía**: Representa aquellos tratamientos quirúrgicos en los que se elimina la mama entera. Está definida por un atributo:
    - ***izq/der/bilateral***: Indica qué mama ha sido extirpada.

  - **Ganglio Centinela**: Representa las cirugías en las que se extirpan los ganglios centinela.
  - **Linfadenectomía**: Son aquellas cirugías en las que se elimina la cadena ganglionar. Está definida por un atributo:
    - ***nivel***: Indica el nivel de extirpación de la cadena ganglionar, dependiendo de si ha sido parcial o completa.

- **Radioterapia**: Representa aquellos tratamientos en los que se somete al paciente a una radiación para eliminar las células malignas. Viene definido por dos atributos:
  - ***dosis***: Indica la dosis de radiación que se le ha dado al paciente mediante un valor numérico entero.
  - ***zona_irradiada***: Contiene la zona en la que se ha irradiado al paciente.

- **Hormonoterapia**: Representa los tratamientos cuya finalidad es la de evitar los influjos hormonales. Tiene un atributo que lo define:
  - ***tipo***: Indica el tipo de tratamiento que recibe el paciente. Contendrá uno de los valores siguientes:
    - **Tamoxifeno**: Es un fármaco administrado a pacientes premenopáusicas
    - **Letrozol**: Se trata de un inhibidor reversible de aromatasa no esteroideo indicado para pacientes postmenopáusicas
    - **Anastrozol**: Igual que el Letrozol, se trata de un inhibidor reversible de aromatasa no esteroideo indicado para pacientes postmenopáusicas
    - **Exemestano**: Se trata de un desactivador irreversible de aromatasa esteroideo indicado para pacientes postmenopáusicas
    - **Fulvestran**: Es un fármaco administrado a pacientes metastásicas postmenopáusicas

- **Biológico**: Representa los tratamientos basados en un fármaco dirigido a una diana terapéutica. Está definido por un atributo:
  - ***tipo***: Indica el tipo de tratamiento biológico que se le va a dar al paciente.

- **Quimioterapia**: Representa los tratamientos que se basan en fármacos químicos. Tiene varios atributos que lo definen:
  - ***esquema***: Contiene el esquema de quimioterapia que se le ha administrado al paciente.
  - ***num_ciclos***: Contiene el número de ciclos de quimioterapia que se le han administrado.

La clase Quimioterapia se especializa en dos subclases:

  - **Adyuvante**: Es aquella quimioterapia que se administra después del tratamiento principal para aumentar la posibilidad

de una supervivencia prolongada.

- **Neoadyuvante**: Es aquella quimioterapia que se aplica antes del tratamiento principal. Este tipo de quimioterapia irá asociada a una clase Respuesta.

La clase **Respuesta** representa la respuesta del paciente a un determinado tratamiento neoadyuvante. Se define por un atributo:

- ***completa/parcial***: Indica si la respuesta al tratamiento ha sido completa o no.

La clase Respuesta puede especializarse en 3 subclases:

- **Clínica**: La que se determina por la palpación de la mama realizada por un oncólogo.
- **Radiológica**: La que se observa mediante el uso de un ecógrafo o un sistema de resonancia magnética.
- **Patológica**: Se determina mediante una cirugía y el análisis posterior de la muestra realizado por el anatomopatólogo. Esta clase tiene dos atributos que son medidores de la respuesta y se representa en estadíos:
  - *RCB*
  - *Miller_Payne*

Además, un tratamiento puede ir asociado a la clase **Toxicidad.** Esta clase representa la toxicidad que puede tener un determinado tratamiento en un paciente. Está definido por tres atributos:

- ***id_toxicidad***: Contiene un identificador interno único.
- ***grado:*** Indica el grado de toxicidad que el tratamiento ha producido en el paciente. Esta graduación tiene valores de 1 a 4.
- ***descripción***: Contiene una descripción de los efectos de la toxicidad del fármaco en el paciente.

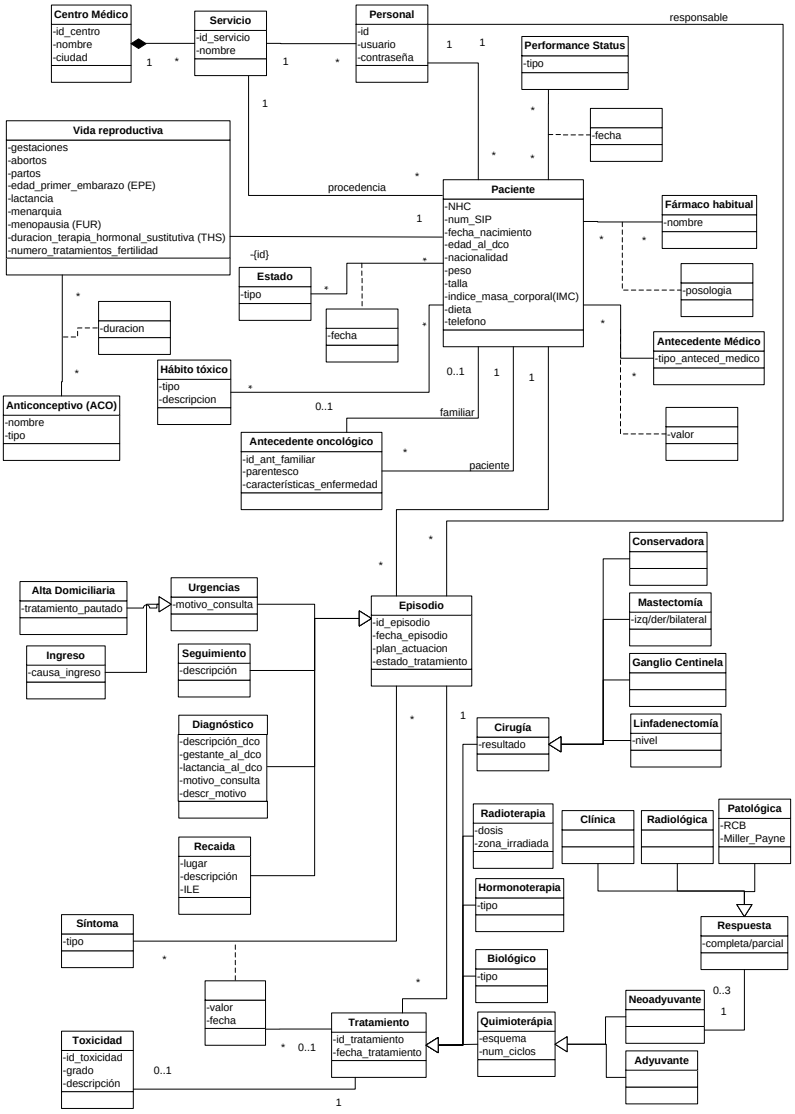

Figura 39. Perspectiva Clínica. Parte I.

La clase **Síntoma** representa los distintos síntomas que puede padecer un paciente durante un episodio. En la relación con la clase episodio hay dos atributos, ***valor*** y ***fecha***, los cuales indican el valor del síntoma y la fecha en la que se tomaron las medidas para ese síntoma en ese episodio. Estos dos

datos, además, pueden estar relacionados con el tratamiento que se le esté administrando al paciente. Esta clase está representada por un atributo:

- *tipo*: Contiene el síntoma que padece el paciente.

La clase **Prueba** representa las distintas pruebas que se le pueden realizar a un paciente durante un episodio. Esta clase está definida por tres atributos:

- *id_prueba*: Contiene un identificador interno único.
- *fecha_prueba*: Indica la fecha en la que se realizó la prueba.
- *descripción*: Contiene una descripción de los resultados obtenidos al realizar la prueba.

La clase prueba puede ser de varios tipos. Estos tipos se representan en varias subclases:

- **Radiología:** Representa aquellas pruebas en las que se utilizan herramientas radiológicas de imagen para llevarlas a cabo. Está relacionada con la clase Metástasis, ya que a partir de este tipo de pruebas se pueden detectar metástasis en la paciente. Esta clase contiene un atributo que la define:
  - *imagen*: Contiene el fichero con la imagen resultante de la prueba radiológica.

  Además, esta clase se especializa en varias clases:

  - **Mamografía:** Representa a las radiografías realizadas en la mama.
  - **Ecografía:** Representa aquellas pruebas realizadas con un ecógrafo.
  - **TAC**: Representa las pruebas de tomografía axial computarizada.
  - **Densitometría Ósea:** Representa aquellas pruebas que determinan la densidad de los huesos. Está definida por dos atributos:

        - ***cadera (t-score)***: Indica el t-score de la densitometría realizada en la cadera.
        - ***columna (t-score)***: Indica el t-score de la densitometría realizada en la columna.

    - **Resonancia Mamaria:** Representa aquellas pruebas en las que se utiliza un equipo de resonancia magnética para obtener imágenes médicas.
    - **BAG/PAAF:** Representa aquellas Biopsias de Aguja Gruesa o las de Punción Aspiración con Aguja Fina que son realizadas por el radiólogo guiadas por una herramienta de imagen médica. Está relacionada con la clase Muestra que representa la muestra extraída mediante esta prueba.

- **Exploración física:** Representa las exploraciones físicas que pueden llevarse a cabo desde el Servicio de Oncología del centro.
- **Laboratorio:** Representa aquellas pruebas que se llevan a cabo en un laboratorio clínico a partir de una muestra de sangre del paciente. Está relacionada con la clase Sangre, representando la muestra sanguínea a la cual se ha realizado el análisis. Esta clase se subdivide en varias clases:
    - **Marcador tumoral:** Representa aquellas pruebas que obtienen como resultado la presencia de una proteína presente en las personas con algún tumor. Está definida por un atributo:
        - ***nombre***: Contiene el nombre del marcador tumoral.
        - ***valor***: Indica la cantidad de ese marcador tumoral en la muestra mediante un valor numérico con decimales.

    - **Química:** Representa las pruebas de un análisis sanguíneo rutinario llevadas a cabo en el laboratorio. Cualquier valor fuera de los rangos de normalidad quedaría descrito en el atributo *descripción* de la clase Prueba.
    - **Hemograma:** Representa los niveles de células sanguíneas presentes en la sangre del paciente. Está definida por varios

atributos:

- ***leucocitos***: Contiene la cantidad de leucocitos en sangre por 10^9/L
- ***neutrófilos***: Contiene la cantidad de neutrófilos en sangre por 10^9/L
- ***hemoglobina***: Contiene la cantidad de hemoglobina en sangre medida en g/dL
- ***plaquetas***: Contiene la cantidad de plaquetas en sangre por 10^9/L

- **Biológicas:** Representa las pruebas relacionadas con la biología molecular. Esta clase está relacionada con la clase Muestra que representa la muestra sobre la que se están realizando las pruebas. Está definida por un atributo:
  - ***procedencia***: Contiene el nombre del laboratorio donde se realizaron las pruebas.

  Esta clase se subdivide en cinco subclases que representan el tipo de pruebas biológicas realizadas:

  - **Inmunohistoquímia:** Representa aquellas pruebas realizadas por el anatomopatólogo con técnicas de inmunohistoquímia. Está definida por los siguientes atributos:
    - ***receptores_estrógenos (RE)***: Contiene los niveles de receptores de estrógenos encontrados en la muestra tumoral representados por un porcentaje.
    - ***receptores_progesterona (RP)***: Contiene los niveles de receptores de progesterona encontrados en la muestra tumoral representados por un porcentaje.
    - ***HER2***: Indica si la proteína sintetizada por el gen HER2 está mutada, con los valores positivo, negativo o nulo si no se ha realizado la prueba.

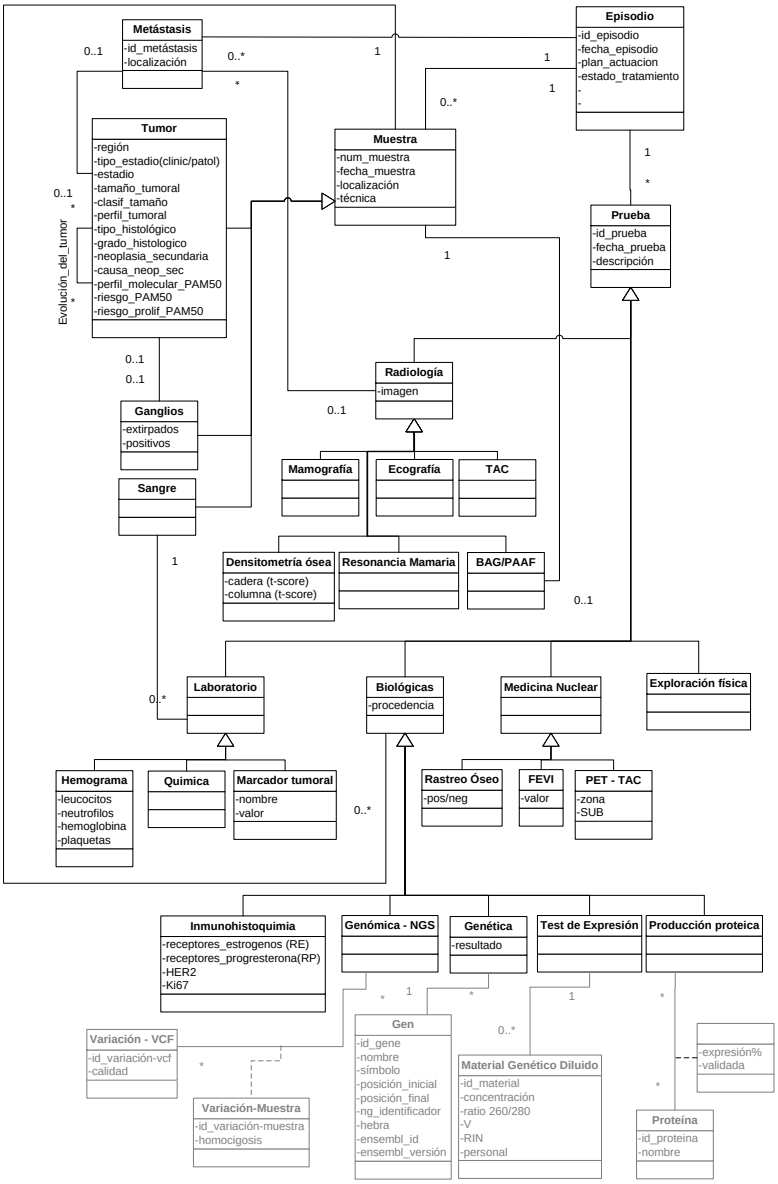

Figura 40. Perspectiva Clínica. Parte II.

- ***Ki67***: Indica si la proteína Ki67 está mutada. Esta proteína es un índice de proliferación celular. Se indica mediante un porcentaje.

- **Genética:** Representa las pruebas de análisis de un único gen realizadas sobre la muestra. Está relacionada con la clase Gen de la vista de Expresión Génica y está definida por un atributo:
  - ***resultado***: Contiene la variación encontrada en el gen en formato HGVS.

- **Test de Expresión**: Representa los estudios de expresión génica que se realizan en un laboratorio de Biología Molecular. Esta clase conecta con la vista Expresión Génica del Esquema Conceptual del Cáncer de Mama.
- **Genómica - NGS:** Representa los análisis genómicos realizados utilizando la técnica de Next Generation Sequencing. Está conectada con la vista Secuenciación Masiva de este esquema.
- **Producción proteica**: Representa los análisis de producción de proteínas realizados a las muestras. Se relaciona con la clase Proteína de la vista de Expresión Génica.

- **Medicina Nuclear:** Representa las pruebas llevadas a cabo en la unidad de Medicina Nuclear. Esta clase se subdivide en otras subclases que representan los distintos tipos de pruebas de medicina nuclear:
  - **Rastreo Óseo:** Representa aquellas pruebas realizadas para comprobar que el cáncer no ha producido metástasis óseas. Está definida por un atributo:
    - ***pos/neg***: Indica si el resultado de la prueba ha sido positivo (habiéndose encontrado metástasis) o negativo.

  - **FEVI:** (Fracción Eyección Ventrículo Izquierdo) Representa aquella prueba que comprueba el correcto funcionamiento del corazón. Está definida por un atributo:
    - ***valor***: Contiene un porcentaje que representa si el ventrículo izquierdo del corazón ha perdido fuerza de eyección.

- **PET-TAC:** Representa las pruebas realizadas utilizando la tecnología del TAC para descartar metástasis. Está definida por dos atributos:
  - ***zona***: Contiene la zona en la que se ha realizado la prueba
  - ***SUB***: Indica la intensidad de la fluorescencia de manera numérica.

La clase **Muestra** representa aquellas muestras biológicas extraídas de un paciente. Esta clase está relacionada con la clase Episodio, representando el episodio en el que se ha tomado la muestra, así como con las clases BAG/PAAF, Biológicas y Cirugía comentadas anteriormente. La clase Muestra se define con los siguientes atributos:

- ***num_muestra***: Contiene el identificador de la muestra.
- ***fecha_muestra***: Contiene la fecha en la que se extrajo la muestra del paciente.
- ***localización***: Contiene la localización inicial de la muestra antes de ser extraída del paciente.
- ***técnica***: indica la técnica utilizada para extraer la muestra

Esta clase se especializa en varias subclases que representan los distintos tipos de muestra:

- **Tumor:** Representa aquella pieza de tejido canceroso extraída de un paciente. La relación de la clase Tumor con la clase Ganglios representa aquellos ganglios que han sido invadidos por el tumor. Además, la relación con la clase Metástasis indica la metástasis que ha podido producir ese tumor en el paciente. Esta clase también se relaciona con ella misma representando la evolución del tumor. Está definida por los siguientes atributos:
  - ***región***: Indica la región mamaria de la cual se ha extraído el tumor. Viene definida por una de las siguientes opciones:
    - Retroareolar

- Línea intercuadrántica interna (LII)
- Línea intercuadrántica externa (LIE)
- Cuadrante superoexterno (CSE)
- Cuadrante inferoexterno (CIE)
- Cuadrante superointerno (CSI)
- Cuadrante inferointerno (CII)

- ***tipo_estadío***: Indica si el estadio del tumor es clínico (previo a la cirugía) o patológico (posterior a la cirugía), pudiendo incluir únicamente uno de estos dos valores.
- ***estadío***: Representa el estadio de evolución del tumor de forma numérica del 1 al 4.
- ***tamaño_tumoral***: Contiene el tamaño del tumor
- ***clasif_tamaño***: Indica en qué lugar de la clasificación por tamaños se encuentra el tumor. La clasificación tiene los siguientes niveles [98]:
  - **Tx**: No se puede evaluar el tumor primario
  - **T0**: No existe prueba de tumor primario
  - **Tis**: Carcinoma in situ
  - **Tis(CDIS):** Carcinoma Ductal In Situ
  - **Tis(CLIS)**: Carcinoma Lobulillar In Situ
  - **Tis(Paget)**: Enfermedad de Paget del pezón que no está relacionada con el carcinoma invasivo o carcinoma in situ (CDIS o CLIS) en el parénquima mamario subyacente.
  - **T1**: El tumor mide ≤ 20 mm
  - **T1 mi**: El tumor mide ≤ 1 mm
  - **T1a:** El tumor mide > 1 mm pero ≤ 5 mm
  - **T1b**: El tumor mide > 5mm pero ≤ 10 mm
  - **T1c**: El tumor mide > 10mm pero ≤ 20 mm
  - **T2**: El tumor mide > 20mm pero ≤ 50 mm
  - **T3**: El tumor mide > 50mm
  - **T4**: El tumor se ha extendido hasta la pared torácica o la piel
  - **T4a**: Extensión a la pared torácica que no sólo incluye

adherencia o invasión a los músculos pectorales
- **T4b**: Ulceración o nódulos satélites ipsilaterales o edema (incluyendo la piel de naranja) la cual no satisface el criterio de carcinoma inflamatorio
- **T4c**: Ambos, T4a y T4b
- **T4d**: Carcinoma inflamatorio

- ***perfil_tumoral***: Indica el perfil celular del tumor después de un examen histológico. Contendrá uno de los siguientes valores:
  - Luminal A
  - Luminal B
  - Her2
  - Her2 Luminal
  - Triple negativo

- ***tipo_histológico***: Información proveniente del análisis de Anatomía Patológica que indica el tipo de célula del tumor.
- ***grado_histológico***: Clasificación de tumor según los criterios de Scarff-Bloom Richardson con los siguientes valores:
  - **I** o bien diferenciado
  - **II** o moderadamente diferenciado
  - **III** o pobremente diferenciado
  - Desconocido

- ***neoplasia_secundaria***: Indica si el tumor extraído es el segundo tumor de un mismo paciente, pero no se corresponde con una metástasis del primero, habiendo sido provocado por los tratamientos administrados.
- ***causa_neop_sec***: Contiene la causa por la que se ha desarrollado el tumor secundario.
- ***perfil_molecular_PAM50***: Perfil molecular obtenido de la realización del análisis genómico PAM50 definido por la clasificación:
  - Luminal A

        - Luminal B
        - Normal
        - HER2
        - Basal

    - ***riesgo_PAM50***: Riesgo de recurrencia según PAM50:
        - Bajo
        - Medio
        - Alto

    - ***riesgo_prolif_PAM50***: Riesgo de proliferación según PAM50:
        - Bajo
        - Medio
        - Alto

- **Ganglios:** Esta clase representa aquella muestra que contiene los ganglios linfáticos extraídos de la paciente para poder observar si han sido invadidos por células tumorales. Se relaciona con la clase Tumor, indicando de qué tumor provienen las células que están infectados los ganglios linfáticos. Está definida por dos atributos:
    - ***extirpados***: Contiene el número de ganglios que le han sido extirpados a la paciente.
    - ***positivos***: Contiene el número de ganglios afectos que se han encontrado entre los extirpados.

- **Sangre:** Representa aquellas muestras sanguíneas extraídas del paciente. Puede estar relacionada con la clase Laboratorio, indicando las pruebas de laboratorio relacionadas con dicha muestra de sangre.

La clase **Metástasis** representa las posibles metástasis que se han desarrollado en un paciente. Estas metástasis pueden haberse localizado por pruebas radiológicas, lo que queda representado por la relación con la clase Radiología, o puede tener su origen en una muestra tumoral del paciente,

quedando representado por la relación con la clase Tumor. Además, esta clase se relaciona con la clase Episodio, representando el episodio en el que se ha encontrado la metástasis. Esta clase está definida por dos atributos:

- ***id_metástasis***: Contiene un identificador interno de la metástasis.
- ***localización***: Indica el lugar donde se ha encontrado la metástasis.

3.1.2. Perspectiva de Expresión Génica

Esta segunda perspectiva contiene los datos que se analizan en un Laboratorio de Investigación en Biología Molecular relacionados con la expresión génica en tumores mamarios. Relaciona información extraída de PCRs-Cuantitativas y Chips de Expresión sobre microARNs y mARNs. En los siguientes párrafos se definen en detalle las clases y relaciones que representan estos conceptos y que podemos visualizar en la Figura 41.

La clase **Material Genético Diluido** representa la disolución de las moléculas genéticas extraídas de la muestra para su posterior análisis. Esta clase está relacionada con la clase Medidor de Expresión indicando que este material se analiza utilizando medidores de expresión y con la clase Test de Expresión, representando la prueba en la que se utiliza el material genético. Esta clase está definida por varios atributos:

- ***id_material***: Contiene un identificador interno del material genético
- ***concentración***: Concentración de material genético en la muestra. Se usa para tener una estimación de cuanto material de estudio hay en la disolución acuosa que preparamos con el material genético extraído.
- ***ratio 260/280***: Es una medida de la pureza de la muestra. En la extracción es probable que, además de ARN, haya también otras moléculas como DNA, proteínas, ... Cuanto más cercano es ese ratio a 2, más pura es la muestra de ARN.
- ***V***: Volumen final de la extracción. Representa la cantidad de

muestra que tenemos para trabajar.

- ***RIN***: *RNA Integrity Number*. Es una medida de la integridad y calidad del ARN total extraído (tanto mARN, microARN, ...). Se le da un valor de N/A (sería 0 o no evaluable) a 10, siendo 10 el valor óptimo.
- ***personal***: Personal que extrajo el material genético de la muestra.

La clase **Medidor de Expresión** representa aquellos análisis de expresión realizados sobre un material genético tumoral. Se relaciona con la clase Material Genético Diluido, representando el material genético utilizado con el medidor de expresión, con la clase Análisis, donde se representan todos aquellos análisis que contienen ese medidor de expresión, y con la clase ARN, representando el ARN del que se está analizando la expresión. Está definido por dos atributos:

- ***id_medidor***: Contiene un identificador interno
- ***expresión***: Contiene la expresión de un ARN en la muestra. Se trata de un valor con 2 o 3 decimales.

Esta clase se especializa en dos subclases que representan las dos tecnologías utilizadas para medir la expresión de los ARNs:

- **Chip de Expresión**: Representa aquellas medidas de expresión realizadas con un chip comercial con los microARNs predefinidos. Está definida por dos atributos:
  - ***.dat***: Contiene la ruta del archivo que contiene los datos de expresión representados de forma gráfica.
  - ***tipo_chip***: Contiene los datos identificativos del chip utilizado.

- **PCR-Cuantitativa:** Representa aquellas medidas de expresión realizadas utilizando PCRs, técnica de fluorescencia en cuyo chip se puede incluir los microARNs relevantes para el estudio. Está definida por varios atributos:
  - ***CT - media***: Contiene la media obtenida de los 3 valores de

expresión registrados, que se representan por el tiempo que la máquina tarda en detectar la fluorescencia.

- ***CT - desviación***: Contiene la desviación típica de esos valores obtenida calculando la diferencia entre la media y la máxima y la media y la mínima.
- ***target/endogeno***: Indica si el ARN analizado se está utilizando para comprobar el correcto funcionamiento del medidor de expresión o es uno de los objetivos de la medición.

La clase **ARN** representa los ácidos ribonucléicos (ARN) de los cuales interesa conocer la expresión. Está relacionada, además de con la clase Medidor de Expresión, con la clase Análisis, donde representa el ARN estudiado en un análisis, con la clase Cromosoma, que representa el cromosoma donde se encuentra el ADN que se transcribe en el ARN, y con la clase Gen, que representa el gen origen a partir del cual se transcribe dicho ARN. Está definida por un conjunto de atributos:

- ***id_arn*** Contiene un identificador interno
- ***nombre***: Incluye el nombre del ARN
- ***posición_inicial***: Contiene la posición cromosómica de inicio del fragmento de ADN que se transcribe en el ARN.
- ***posición_final***: Es la posición cromosómica de fin del fragmento de ADN que se transcribe en el ARN.
- ***accesion_number_miRBase***: Contiene el identificador del ARN en la base de datos miRBase.
- ***versión_miRBase***: Incluye la versión de la base de datos de miRBase en la que se obtuvieron los datos.
- ***nombre_miRBase_old***: Contiene el nombre con el que se identifica al gen en versiones anteriores de la base de datos miRBase.

Esta clase se especializa en dos subclases que representan los tipos de ARN que se analizan:

- **microARN:** Representa los microARNs. Tiene una relación con la

clase Diana, representando los genes a los que se une ese microARN para alterar su transcripción y un atributo en la relación, el ***miTG-score***, que determina la fiabilidad de la predicción de unión entre ese microARN y su diana (cuanto mayor sea, mayor será la probabilidad unión entre ese gen y su diana). Esta clase está definida por un atributo:

  - ***secuencia***: contiene la secuencia del microARN.

- **mARN:** Representa el ARN mensajero.

La clase **Gen** representa los genes relacionados de manera directa o indirecta con la enfermedad. Esta clase se relaciona con la clase Proteína, representando las proteínas que se codifican a partir de ese gen, con la clase Ruta Metabólica, representando las rutas metabólicas en las que el gen está involucrado, con la clase Variación, donde representa las variaciones que tienen lugar en ese gen, con la clase Cromosoma, donde se representa el cromosoma donde se encuentra el gen, y con la clase Genética, que representa el análisis genético realizado sobre ese gen. Además, esta clase está definida por los siguientes atributos:

- ***id_gene***: Identificador del gen en la base de datos Gene de NCBI [99]
- ***nombre***: Contiene el nombre del gen
- ***símbolo***: Símbolo del gen
- ***posición_inicial***: Contiene la posición cromosómica donde empieza el gen.
- ***posición_final***: Contiene la posición cromosómica donde acaba el gen.
- ***ng_identificador***: Contiene el identificador de la secuencia del gen en RefSeq (Base de datos de secuencias de referencia de NCBI [100])
- ***hebra***: Contiene la hebra del cromosoma donde se encuentra el gen
- ***ensembl_id***: Contiene el identificador de ese gen en la base de datos de Ensembl

- ***ensembl_versión***: Indica la versión de Ensembl a la que hace referencia el id.

La clase Gen se especializa en una clase **Diana**, la cual representa los genes a los que se une un microARN para regular su expresión. Esta clase está relacionada con la clase microARN representando los microARNs que se encargan de regular la expresión de este gen.

Cromosoma es una clase que se define como una estructura organizada y única dentro del ADN donde genes, elementos reguladores y otras secuencias de nucleótidos son localizados. En este esquema, se relaciona con la clase Gen, con la clase ARN y con la clase Variación. Esta última clase conecta con la vista de Secuenciación Masiva. Además, un cromosoma tiene una serie de atributos por los cuales es identificado:

- ***nombre***: nombre e identificador del cromosoma en la fuente de datos de la que se ha extraído la secuencia.
- ***secuencia***: secuencia de referencia del cromosoma.
- ***longitud***: campo longitud que indica el número de nucleótidos que tiene la secuencia.
- ***nc_identificador***: contiene el identificador de la secuencia del cromosoma en RefSeq (Base de datos de secuencias de referencia de NCBI [100])

La clase **Ruta Metabólica** representa aquellas rutas metabólicas relacionadas con la enfermedad. Esta clase está relacionada con la clase Gen representando los genes implicados en la ruta metabólica e incluye un atributo ***p-valor*** que representa la relevancia de ese gen en esa ruta metabólica. Está definida por dos atributos:

- ***id_ruta***: Contiene un identificador interno
- ***nombre***: Contiene el nombre identificativo de la ruta metabólica.

La clase **Proteína** representa las proteínas implicadas de alguna manera en el cáncer de mama. Esta clase está relacionada con la clase Gen,

representando el gen que codifica dicha proteína, y con la clase Producción proteica, indicando en qué pruebas se ha encontrado dicha proteína. El porcentaje de expresión de esa proteína y si ha sido validada o no en la muestra correspondiente a esa prueba quedan representados por dos atributos en esta última relación, ***expresión%*** y ***validada*** respectivamente. Esta clase está definida por dos atributos:

- ***id_proteína***: Contiene un identificador interno de la proteína
- ***nombre***: Contiene el nombre de la proteína.

La clase **Análisis** representa los análisis estadísticos utilizados para estudiar la expresión de los ARNs. Esta clase está relacionada con la clase Medidor de Expresión, representando los medidores de expresión de un ARN utilizados en el análisis y, gracias al atributo ***función,*** el papel que cada uno de ellos tiene en ese análisis (caso/control). También está relacionada con la clase ARN indicando el ARN del cual se está haciendo el análisis de expresión y con la clase Estudio indicando a qué estudio pertenece dicho análisis. Esta clase queda definida por los siguientes atributos:

- ***id_analisis***: Identificador interno del análisis
- ***paramétrico/no param***: Indica si los datos del análisis realizado se ajustan o no a la realidad.
- ***p-valor***: Contiene el p-valor obtenido tras la realización del análisis, indicando la significatividad del mismo
- ***FoldChange (RQ)***: Contiene una medida del número de veces que se expresa más un ARN en una determinada muestra respecto a un valor de referencia o control
- ***corrector_multiple_comparación***: Análisis estadístico utilizado para verificar el p-valor y dar robustez al estudio.
- ***método_corrección***: Método utilizado para realizar la corrección por múltiple comparación.
- ***validado***: Indica si el análisis ha sido validado o no, comprobando que se han obtenido los mismos resultados utilizando varias técnicas.

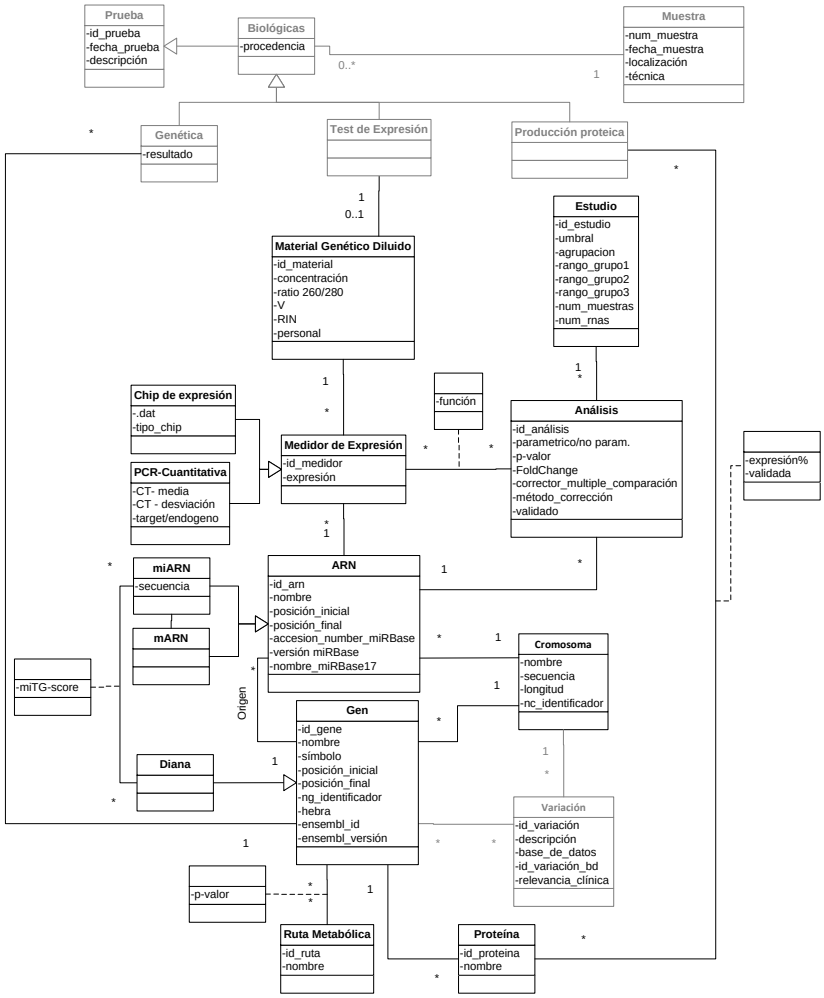

Figura 41. Perspectiva de Expresión Génica.

La clase **Estudio** representa aquellos estudios que se han realizado y que incluyen el análisis de expresión simultáneo de varios ARNs. Esta clase está relacionada con la clase Análisis representando los análisis incluidos en ese estudio. Además, está representada por los siguientes atributos:

- ***id_estudio***: Contiene el identificador interno del estudio
- ***umbral***: Contiene el umbral de selección de medidores de expresión. Para que un medidor de expresión de un ARN sea incluido en el análisis debe superar el umbral definido por el

biólogo que realiza el análisis.

- ***agrupación***: Factor por el cual se han agrupado los sujetos de este estudio (edad, tratamiento, ganglios, ...). A la hora de realizar un estudio los sujetos se agrupan en dos o tres grupos para poder hacer comparaciones entre ellos, cuyos rangos de valores se definen en los tres atributos siguientes.
- ***rango_grupo1***: Contiene el rango utilizado para formar el grupo 1 que se usará para la comparación utilizando métodos estadísticos. Un ejemplo sería un rango de edad de las pacientes menor de 35 años.
- ***rango_grupo2***: Contiene el rango utilizado para formar el grupo 2.
- ***rango_grupo3***: Contiene el rango utilizado para formar el grupo 3.
- ***num_muestras***: Contiene el número de muestras utilizadas en el estudio.
- ***num_arns***: Contiene la cantidad de ARNs que han pasado el umbral.

3.1.3. Perspectiva de Secuenciación Masiva

La última de las vistas del modelo proporciona la perspectiva de los estudios genómicos que utilizan técnicas de Secuenciación de Nueva Generación (NGS), también llamada Secuenciación Masiva. Esta tecnología se utiliza para conocer las variaciones genéticas que ocurren en el genoma del paciente y poder diagnosticar una enfermedad o saber cuál es el tratamiento más apropiado, entre otras muchas aplicaciones. Esta perspectiva incluye los conceptos genómicos que se manejan y los datos procedentes de ficheros VCF (siglas en inglés de *Variant Call Format*, fichero de salida que contiene las variaciones de un paciente obtenidas mediante la tecnología NGS) para poder almacenarlo todo relacionado y de forma estructurada. Las clases y relaciones pertenecientes a esta vista son las que se describen a continuación y que podemos ver en la Figura 42.

En la clase **Variación** se representan, como su propio nombre indica, todas las variaciones existentes en la cadena de ADN. Está relacionada con la clase Gen, que representa los genes a los que afecta la variación, y con la clase Cromosoma, que representa el cromosoma donde tiene lugar la variación. Estas dos clases conectan con la vista de Expresión Génica. También se relaciona con la clase VCF, a través de la clase Variación - VCF que representa aquellas variaciones incluidas en el fichero VCF. Los atributos de la clase Variación son:

- ***id_variación***: identificador interno de la variación.
- ***descripción***: proporciona una descripción de la variación.
- ***base_de_datos***: contiene el nombre de la base de datos de la que se ha extraído la información sobre esa variación, cuando es conocida.
- ***id_variación_bd***: identificador que proporciona la fuente de datos de la cual se ha extraído la variación.
- ***relevancia_clínica***: indica la relevancia clínica que tiene esa variación.

En la jerarquía *Descripción*, una variación puede estar especializada en dos clases: **Precisa** o **Imprecisa**, dependiendo de si se conocen datos sobre su posición.

Además, en la jerarquía *Publicación*, una variación puede estar especializada en dos clases: **Documentada** y **No Documentada**, dependiendo de si la información sobre esta variación está publicada en alguna bibliografía o si se trata de una variación encontrada en un paciente recientemente, sin información bibliográfica asociada.

La clase **Precisa**, especialización *Descripción*, representa las variaciones detectadas con posición conocida dentro del cromosoma en la secuencia de ADN. Su definición viene proporcionada por los siguientes atributos:

- ***posición***: posición en la que se encuentra la variación dentro de la secuencia del cromosoma.
- ***calidad_alineamiento***: indica la calidad del alineamiento en el que

se encontró la variación descrita.

La clase Precisa se especializa en cuatro nuevas entidades dependiendo de qué tipo de variación haya tenido lugar dentro del genoma:

- La clase **Inserción** representa variaciones que consisten en la inserción de una secuencia de nucleótidos un número de veces en la secuencia de ADN del cromosoma. Sus atributos son:
  - ***secuencia***: secuencia de nucleótidos insertados en la secuencia.
  - ***repetición***: número de veces que se repite la secuencia insertada.

- La clase **Delección** representa variaciones que consisten en el borrado de un número de nucleótidos en la secuencia de ADN del cromosoma. A pesar de ser una palabra que no existe en el diccionario de la Real Academia Española de la Lengua, se ha utilizado este nombre por ser comúnmente utilizado en el entorno biológico español para referirse a las variaciones que conllevan el borrado de algún nucleótido. El atributo que la define es:
  - ***bases***: número de nucleótidos borrados en la secuencia

- La clase **Indel** representa variaciones consistentes en inserciones y borrados a la vez en la secuencia de ADN del cromosoma. Los atributos que la definen son:
  - ***secuencia_ins***: secuencia de nucleótidos insertados en la secuencia.
  - ***repetición_ins***: número de veces que se repite la secuencia insertada.
  - ***bases_borradas***: número de nucleótidos borrados.

- La clase **Inversión** representa variaciones que invierten el orden de una secuencia de nucleótidos en la secuencia del cromosoma. Es definida por:
  - ***bases***: número de nucleótidos invertidos en la secuencia.

La clase **Imprecisa**, dentro de la jerarquía *Descripción,* representa variaciones cuya posición es desconocida dentro de la secuencia de ADN. La única información que se conoce es una descripción en lenguaje natural y es su único atributo:

- ***descripción***: descripción de la variación en lenguaje natural.

La clase **BD Bibliográfica** representa las distintas fuentes de datos de la web de las que se extraen las publicaciones científicas. La relación con la clase Referencia representa las publicaciones científicas incluidas en la base de datos bibliográfica. Tiene dos atributos:

- ***nombre***: nombre de la base de datos de la que se extraen las publicaciones científicas.
- ***url:*** dirección web de la base de datos de las que se extraen las publicaciones.

Referencia proporciona información sobre los artículos relacionados con cada una de las variaciones si se dispone de ella. Se relaciona con la clase BD Bibliográfica y con Documentada. Tiene siete atributos:

- ***id_referencia:*** identificador interno de las referencias bibliográficas.
- ***título***: título del artículo.
- ***autores***: autores que han escrito el artículo.
- ***resumen***: resumen del artículo.
- ***publicación:*** referencia resumida a la publicación.
- ***fecha_pub:*** fecha en la cual se ha publicado el artículo.
- ***id_pubmed:*** identificador que la base de datos de Pubmed proporciona al artículo.

La clase **VCF** representa un fichero VCF procedente de las técnicas de secuenciación de NGS y permite centrar la información almacenada en las variaciones genómicas más relevantes como los SNP, indels y variaciones estructurales de gran longitud encontradas en un número determinado de

muestras en base a una secuencia de referencia. Algunos de los atributos de esta clase se han mantenido con su nombre en inglés por mantener la correspondencia de la nomenclatura con el archivo VCF. Tiene varios atributos:

- ***id_vcf:*** identificador interno del fichero.
- ***file_format:*** indica la versión de VCF utilizada.
- ***file_date:*** indica la fecha en la que se generó el fichero VCF.
- ***source:*** hace referencia al software de análisis y mapeo de ficheros SAM/BAM utilizado para la obtención del fichero VCF.
- ***reference:*** indica la dirección URL del fichero FASTA correspondiente al genoma de referencia utilizado.
- ***contig:*** describe cada uno de los fragmentos formados por secuencias contiguas del genoma (contig) secuenciadas utilizados en el análisis de variaciones que representa el fichero.
- ***phasing:*** indica si los alelos se han ordenado correctamente teniendo en cuenta los dos cromosomas o si, por el contrario, no se sabe exactamente a qué cromosoma pertenecen ciertas variaciones.

Esta clase se relaciona con las clases *Variación,* a través de la clase *Variación-VCF,* y *Filtro.*

Variación-VCF proporciona información sobre cada una de las variaciones que aparecen en el fichero VCF. Es definida por:

- ***id_variación-vcf***: identificador interno de la variación.
- ***calidad:*** indica la calidad de la lectura de esa variación en ese el estudio.

Esta clase está relacionada con la clase *Variación*, que es la que define la variación, y con las clases *VCF,* que es la que representa el fichero donde se encuentra la variación, y con las clases *Filtro, Info, Variación-Muestra* y *Genómica-NGS.* Esta última clase conecta con la vista Clínica.

La clase **Filtro** permite describir los filtros creados por el usuario y aplicados por el software de análisis de variaciones, utilizado para la obtención del fichero VCF. Tiene los siguientes atributos:

- ***id_filtro:*** identificador interno del filtro.
- ***identificador:*** identificador definido por el usuario.
- ***descripción:*** contiene la descripción de la funcionalidad del filtro.

Esta clase está relacionada con la clase *VCF* y con la clase *Variación-VCF,* con la cual se asocia a través del atributo ***si/no*** que indica si esa *Variación-VCF* pasa o no un determinado filtro.

La clase **Info** permite definir variables referentes a información adicional sobre las variaciones encontradas en el estudio. Está definida por los siguientes atributos:

- ***id_info***: identificador interno de la variable.
- ***id***: identificador de la variable en el fichero origen.
- ***número***: indica el número de valores que se asocia a esta variable en cada variación. Cuando se asocien más de un valor estos irán separados mediante coma.
- ***tipo***: indica el tipo de datos de el/los valor/es asociados a la variable.
- ***descripción***: describe en lenguaje natural el significado de la variable.

Esta clase está relacionada con la clase *Variación-VCF,* con la cual se asocia a través del atributo ***valor*** que indica el valor del atributo *info* con respecto a esa variación en el estudio recogido en ese fichero VCF.

La clase **Variación-Muestra** proporciona información sobre cada una de las variaciones que aparecen en el fichero VCF con respecto a una muestra en particular analizada a través de una prueba. Viene definida por un atributo:

- ***homocigosis:*** indica si esa variación se encuentra en la muestra en homocigosis o en heterocigosis.

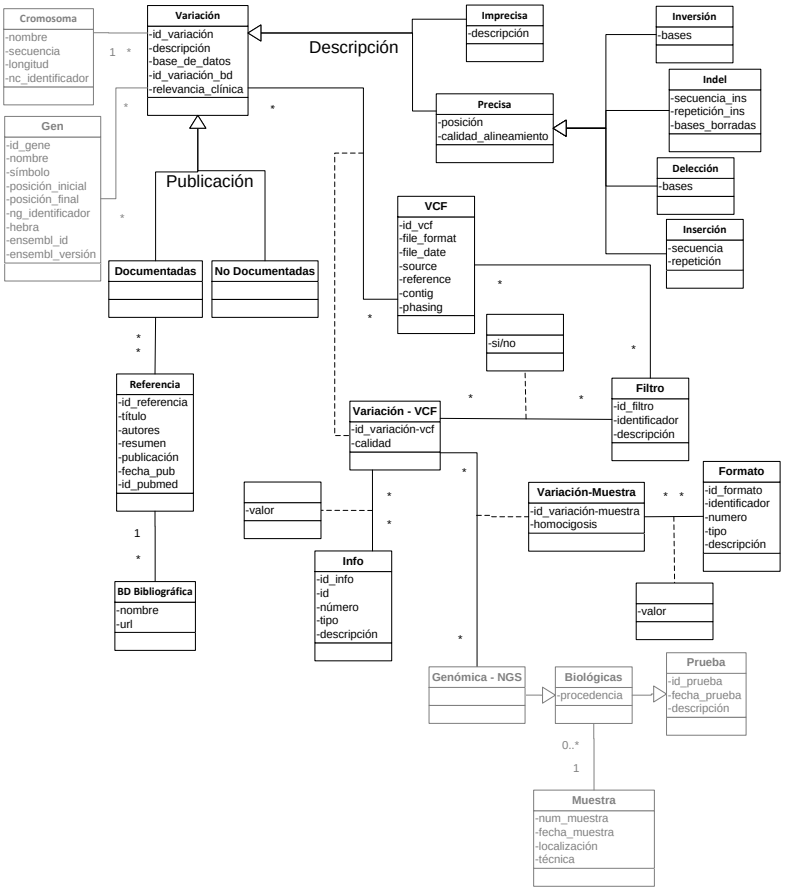

Figura 42. Perspectiva de Secuenciación Masiva

Además, esta clase está relacionada con las clases *Variación-VCF,* que indica la variación analizada, y *Genómica-NGS,* que representa la prueba genómica a la que está asociada la muestra. Esta última clase conecta con la vista Clínica. También se relaciona con la clase *Formato.*

La clase **Formato** describe información sobre los genotipos de las muestras que aparecen en el fichero VCF. Contiene información específica acerca de las muestras en la posición referente a la variación encontrada. Se define utilizando varios atributos:

- ***id_formato***: identificador interno de la variable.
- ***identificador***: identificador de la variable en el fichero origen.
- ***número***: indica el número de valores que se asocia a esta variable en cada variación. Cuando se asocien más de un valor estos irán separados mediante coma.
- ***tipo***: indica el tipo de datos del/los valor/es asociados a la variable.
- ***descripción***: describe en lenguaje natural el significado de la variable.

Esta clase está relacionada con la clase *Variación-Muestra,* con la cual se asocia a través del atributo ***valor*** que indica el valor del atributo *format* con respecto a esa variación de esa muestra en particular.

3.2. Estandarización de la información relacionada con el Cáncer de Mama utilizando el estándar ISO 13606

Como se ha comentado en el capítulo 2.3.1 de esta tesis, un arquetipo es una definición formal que documenta la estructura de información asociada a un concepto clínico (por ejemplo, el concepto de "Informe de alta", "Medida de glucosa" o "Historia familiar"). Intuitivamente se podría interpretar un arquetipo como la plantilla de un documento clínico o una parte del mismo. Una forma de caracterizar de forma precisa las primitivas propuestas por el estándar y sus relaciones es especificarlas con un Diagrama de Clases que ponga de manifiesto los conceptos básicos incluidos en la Historia Clínica Electrónica de un paciente y las relaciones que existen entre ellos. Este Diagrama de Clases es un metamodelo que define los elementos básicos incluidos en los distintos informes clínicos, estructurados en forma de arquetipos, y las relaciones existentes entre ellos.

A modo de ejemplo, cada informe clínico particular puede ser visto como una instanciación de ese metamodelo, para la que se determina cuál es la composición seleccionada, qué secciones y/o entradas la forman, y qué elementos (y agrupaciones de elementos) son usados como contenedor básico (atómico) de introducción de datos clínicos relevantes.

La utilización efectiva y eficiente de arquetipos debe ir asociada a la elaboración de un esquema conceptual que actúe como soporte ontológico-fundacional de la información que el experto clínico tiene que gestionar. De esta forma los arquetipos representan las distintas vistas que se ofrecen a los usuarios del sistema de información diseñado. Además, permiten compartir la información del sistema de información con otros sistemas compatibles con la norma ISO 13606, haciendo que sean semánticamente interoperables. Por lo tanto, la utilización conjunta de las metodologías de Modelado Conceptual y la creación de arquetipos bajo el estándar ISO 13606 supone una simbiosis entre ambas metodologías que la convierten en

una aproximación óptima para el almacenamiento, gestión y explotación de datos clínicos.

Aplicando este principio, en este trabajo de investigación vamos a proponer un conjunto básico de arquetipos diseñado a partir del esquema conceptual presentado en el capítulo 3.1 de esta tesis para la generación de varios Informes Clínicos que contengan la información definida en el Esquema Conceptual de Cáncer de Mama. De esta manera, definiremos modelos de Informes Clínicos que recojan toda la información relevante manejada durante los procesos de diagnóstico, tratamiento y estudio del cáncer de mama, definida y modelada en detalle en el capítulo 3.1. Lógicamente, estos arquetipos, definidos utilizando el estándar ISO 13606 definido en el capítulo 2.3.1 del Estado del Arte de esta tesis, son totalmente compatibles con el sistema de información que vamos a diseñar, facilitando el intercambio de información entre sistemas. Esta estrategia proporciona una metodología rigurosa de diseño y desarrollo de arquetipos basada en dos fases:

1 - Caracterización de la información relevante para el dominio analizado, que conduce a la elaboración de un esquema conceptual (definido en el capítulo 3.1.)

2 - Diseño de las distintas vistas con las que esa información va a ser gestionada, que va a traducirse en el conjunto de arquetipos que utilizará el usuario final.

Para el diseño de los arquetipos hemos utilizado la herramienta LinkEHR [66, 67], a la cual ya hicimos referencia en el capítulo 2.3.1 del Estado del Arte.

En primer lugar, y siguiendo las indicaciones de la norma ISO 13606, se ha creado el arquetipo *CEN-EN13606-FOLDER.Cancer_de _Mama.v1.adl* (ver Figura 43), considerando el "Cáncer de Mama" como una instancia de la entidad *FOLDER*, en la que integraríamos distintas *COMPOSITIONS*, que

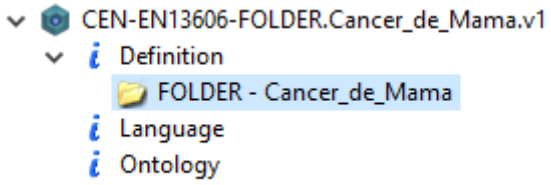

Figura 43. Arquetipo *CEN-EN13606-FOLDER.Cancer_de_Mama.v1.adl*

son los arquetipos que hemos creado haciendo referencia al Esquema Conceptual del Cáncer de Mama. Estos arquetipos representan los procedimientos de recogida de información que se llevan a cabo durante el transcurso de la enfermedad. Para facilitar la estructuración del resto de arquetipos vamos a dividir los datos en las mismas tres vistas que hemos utilizado en el capítulo 3.1 para dividir el Modelo Conceptual del Cáncer de Mama.

3.2.1. Diseño de los arquetipos relacionados con la perspectiva Clínica.

En este apartado describiremos los diseños de los arquetipos correspondientes a los datos clínicos del paciente. Los arquetipos creados son los siguientes:

- CEN-EN13606-COMPOSITION.Anamnesis.v1.adl
- CEN-EN13606-ENTRY.Vida_Reproductiva.v1.adl
- CEN-EN13606-COMPOSITION.Episodio.v2.adl
- CEN-EN13606-SECTION.Tratamientos.v1.adl
- CEN-EN13606-CLUSTER.Toxicidad.v1.adl
- CEN-EN13606-SECTION.Sintomas.v1.adl
- CEN-EN13606-SECTION.Pruebas.v1.adl
- CEN-EN13606-SECTION.Muestras.v1.adl

Como podemos ver, únicamente hay dos de ellos que son instancia de *COMPOSITION*, lo que indica que integraremos la información en dos informes. El resto de arquetipos corresponden a instancias de *SECTION,*

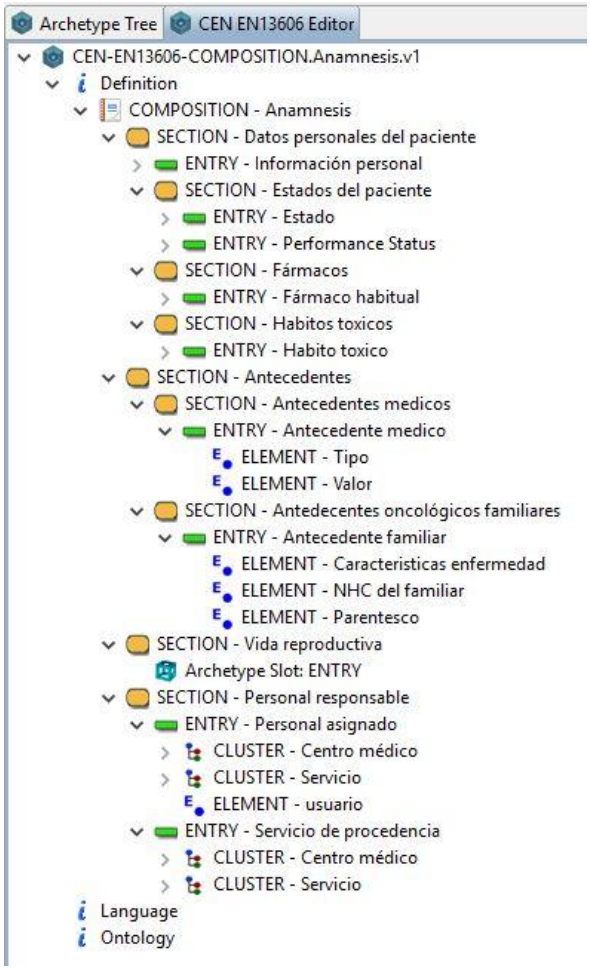

Figura 44. Arquetipo CEN-EN13606-COMPOSITION.Anamnesis.v1.adl

CLUSTER o *ENTRY*, que se integran dentro de los otros arquetipos, referenciándolos dentro de ellos, como veremos en adelante. Esta estrategia permite la reutilización de arquetipos, pudiendo definir conceptos más complejos a partir de conceptos simples ya definidos en otros arquetipos.

El arquetipo *CEN-EN13606-COMPOSITION.Anamnesis.v1*.adl podemos verlo en la Figura 44 y contiene toda la información que se recoge en la primera visita de la paciente referente a su vida previa antes de padecer la enfermedad. Esta información es la que se corresponde con las clases "Paciente", "Estado", "Performance Status", "Vida Reproductiva",

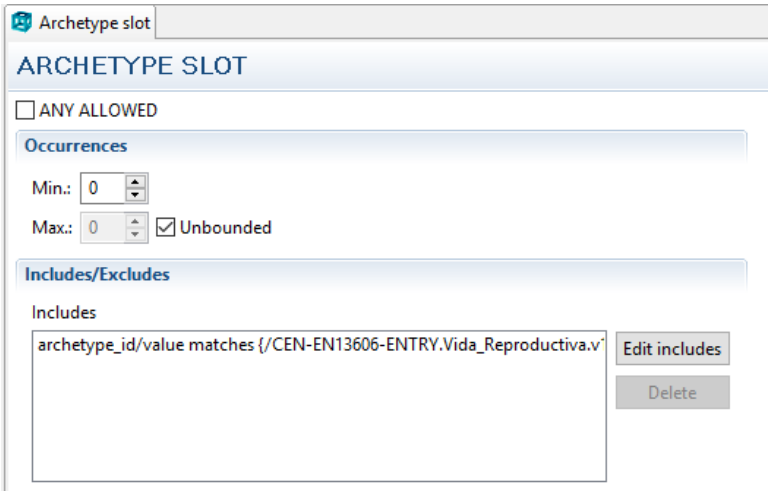

Figura 45. Detalle del *Archetype Slot* que hace referencia al arquetipo *CEN-EN13606-ENTRY.Vida _Reproductiva.v1.adl*

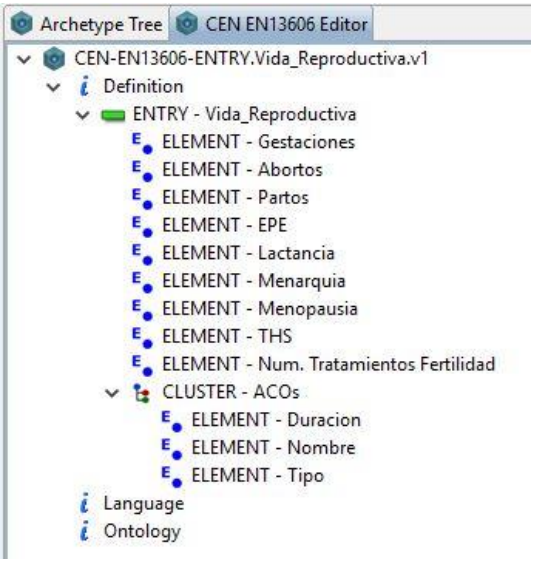

Figura 46. Arquetipo *CEN-EN13606-ENTRY.Vida_Reproductiva.v1.adl*

"Anticonceptivo", "Fármaco habitual", "Hábito Tóxico", "Antecedente Médico", "Antecedente Oncológico", "Personal", "Servicio" y "Centro Médico", y teniendo en cuenta que los datos que representamos en el arquetipo son los mismos que hemos representado en el Esquema Conceptual del capítulo 3.1, volver a explicar el significado de cada uno de

los datos sería redundante. Cabe destacar que hemos estructurado el arquetipo utilizando el tipo *SECTION*, simulando una estructura de informe clínico. En este caso, como vemos en la Figura 44, hemos creado una sección "Antecedentes" donde hemos incluido las secciones "Antecedentes médicos" y "Antecedentes oncológicos familiares". Por otro lado, dentro de la sección "Datos personales del paciente" hemos incluido la "Información personal", los "Fármacos", los "Hábitos tóxicos" y los "Estados del paciente", en los que se incluyen "Performance status" y "Estado". Además, en la sección "Personal responsable" hemos incluido dos *ENTRIES* llamadas "Personal asignado" y "Servicio de procedencia" que hacen referencia a las clases "Personal", "Servicio" y "Centro médico" y sus relaciones con la clase "Paciente". Finalmente, dentro de la sección "Vida reproductiva" hemos incluido un *Archetype Slot: ENTRY* que, como podemos ver en la Figura 45, hace referencia al arquetipo *CEN-EN13606-ENTRY.Vida _Reproductiva.v1.adl*, que podemos ver en la Figura 46. En este arquetipo quedan representadas las clases "Vida reproductiva" y "Anticonceptivo". Además, es interesante destacar que cada una de las estructuras *ELEMENT* tiene un tipo de valor de datos definido, como podemos ver en detalla en la en la Figura 47 y que no se muestran en las demás figuras de arquetipos por falta de espacio en la imagen.

A continuación, se muestra el arquetipo *CEN-EN13606-COMPOSITION .Episodio.v2.adl* en la Figura 48. Este arquetipo representa un informe clínico de un episodio de la vida el paciente en el transcurso de su enfermedad. En él podemos ver varias secciones que representan los distintos tipos de episodio, para las cuales se ha establecido una

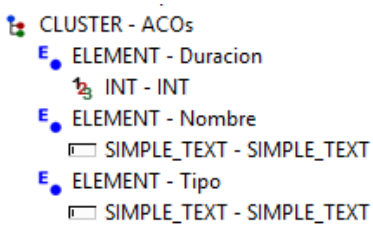

Figura 47. Detalle del arquetipo *CEN-EN13606-ENTRY.Vida _Reproductiva.v1.adl* donde se muestran los tipos de valores.

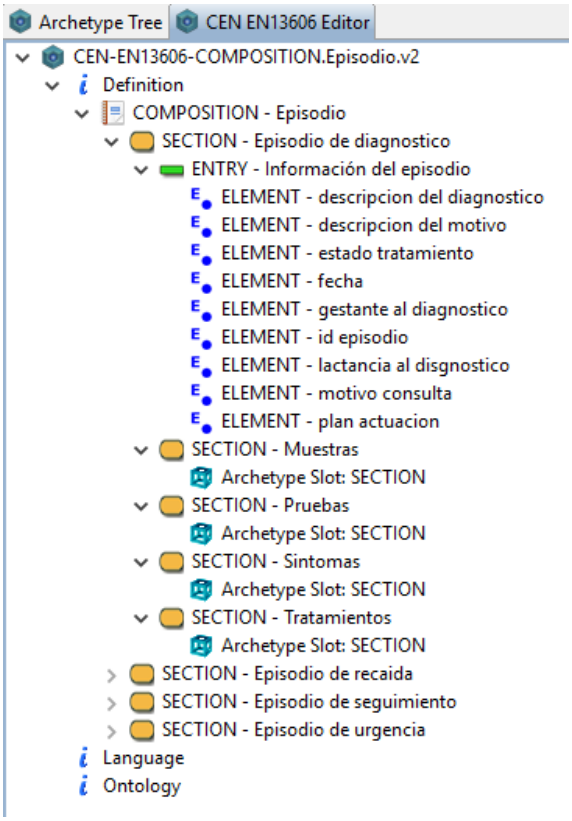

Figura 48. Arquetipo *CEN-EN13606-COMPOSITION.Episodio.v2.adl*

cardinalidad limitando que únicamente sea posible una instancia de una de las secciones por composición, controlando de esta manera que la composición represente a un único episodio del paciente. En otras palabras, si el episodio que se pretende incluir en esta composición es un "Episodio de diagnóstico" únicamente podremos incluir una instancia de esta *SECTION* en la *COMPOSITION.* Esta cardinalidad se define en las opciones internas de la *COMPOSITION,* por lo que no queda representada en la figura. En este arquetipo están representadas las clases "Episodio", "Urgencias", "Alta domiciliaria", "Ingreso", "Seguimiento" y "Recaída". En la Figura 48 podemos ver todas las *SECTION* del arquetipo, pero únicamente desplegada la sección "Episodio de diagnóstico". Esto se debe a que todos los tipos de episodio han sido diseñados de la misma manera y sería redundante explicar cada uno de ellos. En cada sección de un tipo de

episodio podemos encontrar una *ENTRY* llamada "Información del episodio" que contiene todos los atributos de la clase "Episodio" y los específicos de cada una de las especializaciones de la misma, es decir, la sección "Episodio de diagnóstico" contiene en su entrada "Información del episodio" los atributos de las clases "Episodio" y "Diagnóstico", como podemos ver en la Figura 48. Además, cada episodio contiene otras 4 secciones llamadas "Muestras", "Pruebas", "Síntomas" y "Tratamientos" que contienen cada una de ellas un *Archetype Slot* que hace referencia a los arquetipos *CEN-EN13606-SECTION.Muestras.v1.adl, CEN-EN13606-*

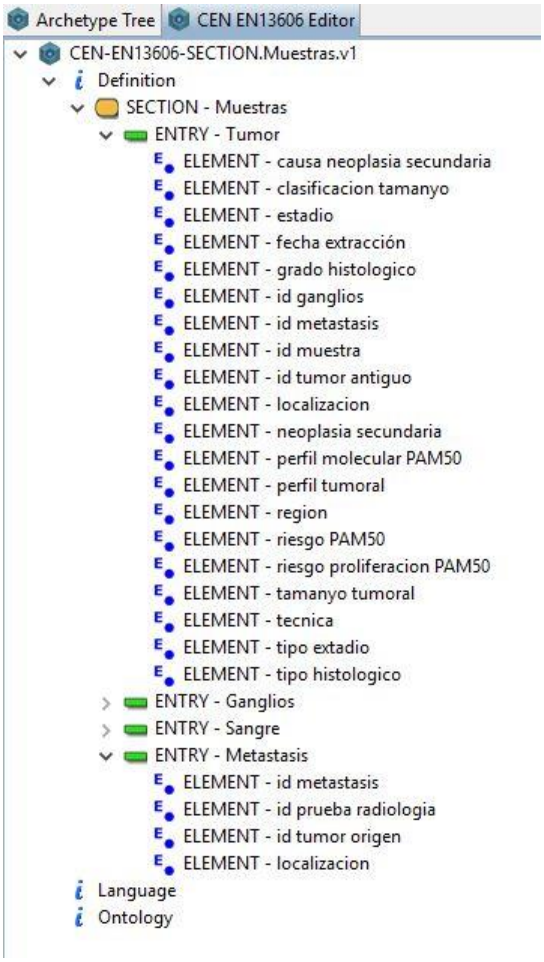

Figura 49. Arquetipo *CEN-EN13606-SECTION.Muestras.v1.adl*

SECTION.Pruebas.v1.adl, CEN-EN13606-SECTION.Sintomas.v1.adl y *CEN-EN13606-SECTION. Tratamientos.v1.adl,* respectivamente.

El arquetipo *CEN-EN13606-SECTION.Muestras.v1.adl* contiene la *SECTION* a la que se hace referencia desde los *Archetype Slot* del arquetipo *CEN-EN13606-COMPOSITION.Episodio.v2.adl.* Este arquetipo, que podemos ver en la Figura 49, representa todas las muestras y metástasis que puede tener un paciente. Aquí quedan representadas las clases "Muestra", "Tumor", "Sangre", "Ganglios" y "Metástasis". Hemos incluido las metástasis en este apartado por su relación directa con el tumor. Como

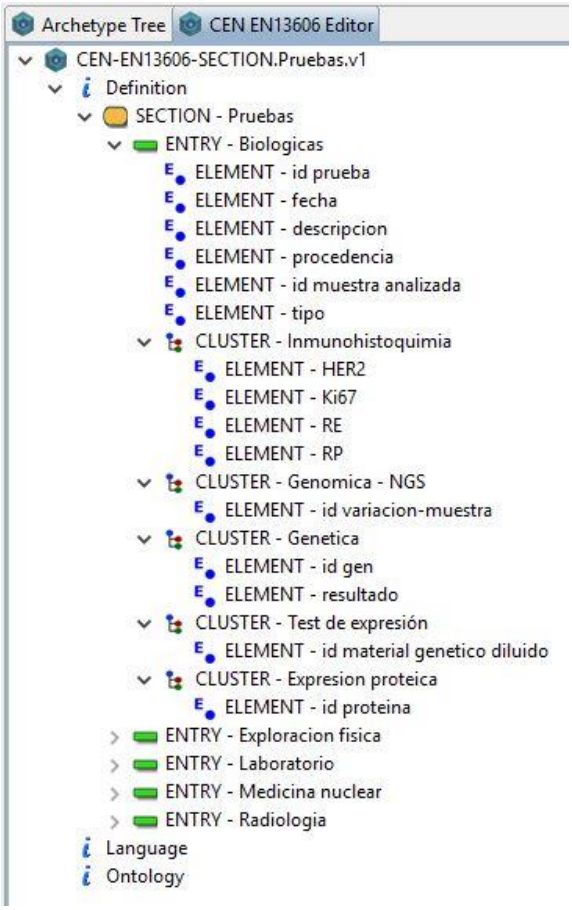

Figura 50. Arquetipo *CEN-EN13606-SECTION.Pruebas.v1.adl*

podemos ver en la Figura 49, las clases “Ganglios”, “Tumor” y “Sangre” se han convertido en estructuras *ENTRY* con *ELEMENTS* que corresponden a los atributos de la clase “Muestra” más los de la clase correspondiente. La cuarta *ENTRY* de este arquetipo corresponde a la clase “Metástasis” con todos sus atributos representados.

El siguiente arquetipo llamado *CEN-EN13606-SECTION.Pruebas.v1.adl* contiene la *SECTION* a la que se hace referencia desde los *Archetype Slot* del arquetipo *CEN-EN13606-COMPOSITION.Episodio.v2.adl.* Este arquetipo, que podemos ver en la Figura 50, representa todas las pruebas que se le han hecho al paciente. En el arquetipo se han diseñado cinco *ENTRY* que corresponden con las clases de pruebas realizadas: “Biologicas”, “Exploración física”, “Laboratorio”, “Medicina nuclear” y “Radiología”. Dentro de cada una de ellas se han incluido en forma de *ELEMENTS* los atributos de la clase prueba y los específicos de cada una de las clases hijas. De esta manera, se podrán generar tantas instancias de pruebas de cualquiera de los cinco tipos como sea necesario. Además, las especializaciones de cada tipo de prueba y los atributos de las mismas se han representado como *CLUSTERS* y *ELEMENTS* respectivamente, como podemos ver en la Figura 50, donde aparece desplegada la *ENTRY* “Biológicas”.

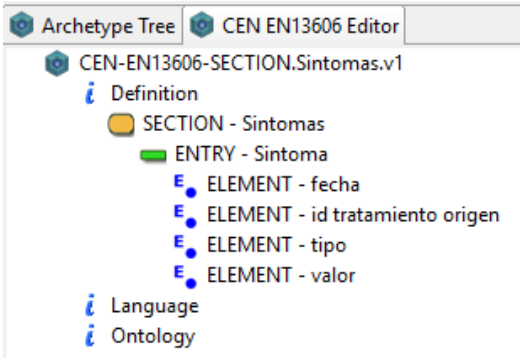

Figura 51. Arquetipo *CEN-EN13606-SECTION.Sintomas.v1.adl*

El siguiente *Archetype Slot* que aparece en el arquetipo *CEN-EN13606-COMPOSITION.Episodio.v2.adl* se encuentra en la *SECTION* "Síntomas" y hace referencia al arquetipo *CEN-EN13606-SECTION.Sintomas.v1.adl* que aparece en la Figura 51. Este arquetipo representa los posibles síntomas que puede padecer un paciente durante un episodio. Como podemos ver, contiene una *ENTRY* "Sintoma" con cuatro *ELEMENTS* que corresponden con los atributos de la clase "Síntoma" y de su relación con "Tratamiento" y "Episodio".

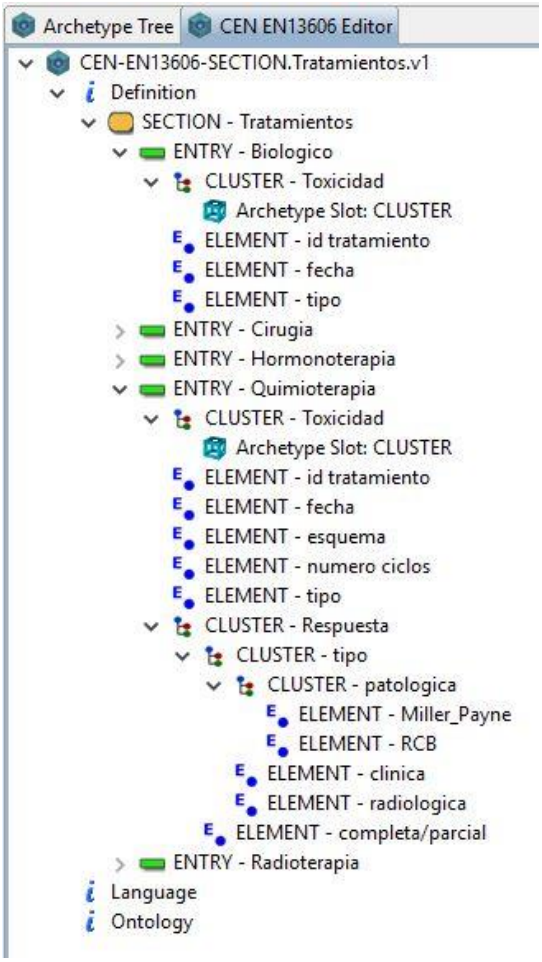

Figura 52. Arquetipo *CEN-EN13606-SECTION.Tratamientos.v1.adl*

El siguiente arquetipo de datos clínicos está enlazado con el arquetipo *CEN-EN13606-COMPOSITION.Episodio.v2.adl* a través de un *Archetype Slot* que aparece en la *SECTION* "Tratamientos" y se llama *CEN-EN13606-SECTION.Tratamientos.v1.adl.* Este arquetipo representa todos los tratamientos que se le dan a un paciente durante un episodio y está representado en la Figura 52. Este arquetipo representa las clases "Tratamiento" y todas sus especializaciones del Esquema Conceptual del Cáncer de Mama. Como podemos ver en la figura, donde están desplegadas las entradas "Biologico" y "Quimioterapia" a modo de ejemplo, las especializaciones quedan representadas por una *ENTRY*, que contiene como *ELEMENTS* los atributos de "Tratamiento" y los propios de cada especialización. En el caso de "Quimioterapia", además hay un *CLUSTER* "Respuesta" en el que, utilizando una jerarquía de *CLUSTERS* Y *ELEMENTS*, representa a la clase "Respuesta" y sus especializaciones. Además, la relación con la clase "Toxicidad" se representa mediante un *CLUSTER* que contiene un *Archetype Slot* al arquetipo *CEN-EN13606-CLUSTER.Toxicidad.v1.adl.*

El último de los arquetipos de datos clínicos es el *CEN-EN13606-CLUSTER.Toxicidad.v1.adl*, el cual está referenciado en el arquetipo *CEN-EN13606-SECTION.Tratamientos.v1.adl.* Este arquetipo representa las toxicidades que se pueden producir en referencia a un tratamiento. Representa a la clase "Toxicidad" del Esquema Conceptual del Cáncer de Mama y podemos verlo en la Figura 53.

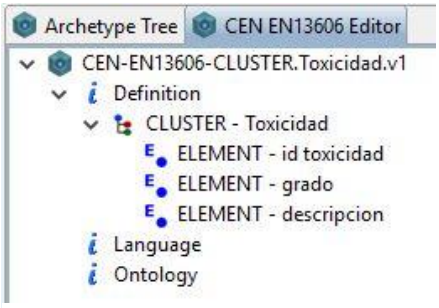

Figura 53. Arquetipo *CEN-EN13606-CLUSTER.Toxicidad.v1.adl*

3.2.2. Diseño de los arquetipos relacionados con la perspectiva de Expresión Génica

Para poder representar los conceptos relacionados con un informe de expresión génica de un paciente ha sido necesario el diseño de varios arquetipos:

- CEN-EN13606-COMPOSITION.Informe_de_expresion.v1.adl
- CEN-EN13606-CLUSTER.Lista_de_expresiones.v1.adl
- CEN-EN13606-CLUSTER.Gen.v1.adl

El arquetipo que representa la información incluida en un informe de expresión génica de un paciente y engloba los datos relacionados con las

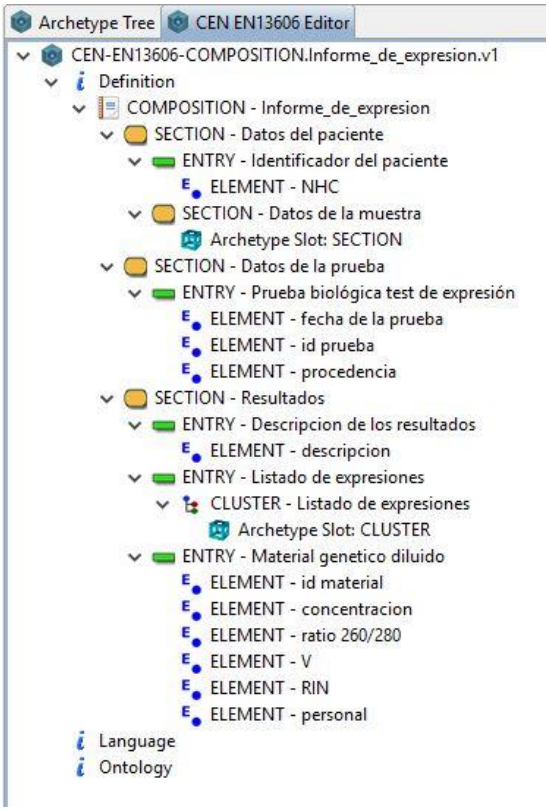

Figura 54. Arquetipo CEN-EN13606-COMPOSITION.Informe_de_expresion .v1.adl

pruebas de expresión génica es una instancia de la estructura *COMPOSITION* (como se puede ver en la Figura 54) y tiene el nombre de *CEN-EN13606-COMPOSITION.Informe_de_expresion.v1.adl.* En él podemos ver varias *SECTIONS,* entre la que se encuentra "Datos de la prueba" con la *ENTRY* "Prueba biológica test de expresión" que contiene los atributos de las clases "Prueba" y "Biológicas" del Esquema Conceptual. Sin embargo, el atributo "descripción" de la clase "Prueba" lo hemos representado con una *ENTRY* llamada "Descripcion de los resultados" en la *SECTION* "Resultados", ya que conceptualmente corresponde con la descripción de los resultados de la prueba. En este arquetipo, también hemos incluido un identificador del paciente a quien corresponde el informe (en este caso el NHC), como un *ELEMENT* de la *ENTRY* "Identificador del paciente" de la *SECTION* "Datos del paciente", en la que también incluimos la información de la muestra dentro de una *SECTION* "Datos de la muestra" que contiene un *Archetype Slot* haciendo referencia al arquetipo *CEN-EN13606-SECTION.Muestras.v1.adl* definido en el apartado 3.2.1. Además, como hemos comentado anteriormente, encontramos una *SECTION* "Resultados", que representa los resultados obtenidos en la realización del test de expresión, y en la cual, además de la *ENTRY* "Descripcion de los resultados", tenemos la *ENTRY* "Material genético diluido", que representa a la clase del Esquema Conceptual con el mismo nombre y tiene como *ELEMENTS* los atributos de la misma, y la *ENTRY* "Listado de expresiones", que contiene un *CLUSTER* con un *Archetype Slot* al arquetipo *CEN-EN13606-CLUSTER.Lista_de_expresiones.v1.adl.*

El arquetipo *CEN-EN13606-CLUSTER.Lista_de_expresiones.v1.adl,* que vemos en la Figura 55, representa aquellas expresiones génicas en forma de microARNs o mARNs medidas en la muestra de un paciente. En él podemos ver la información sobre los ARNs medidos en la muestra, estructurada en dos *CLUSTERS* principales: "Medidor de expresión" e "Información sobre el ARN". El primero de ellos representa a la clase "Medidor de Expresión" del Esquema Conceptual con sus dos especializaciones en forma de *CLUSTERS*: "Chip de expresion" y "PCR

Cuantitativa". Cada uno de ellos contiene como *ELEMENTS* los atributos de las clases a las que representan. El segundo de los *CLUSTERS* principales, llamado "Información sobre el ARN", contiene como *ELEMENTS* los atributos de la clase ARN y, además, tres *CLUSTERS* llamados "Cromosoma", que hace referencia a la clase "Cromosoma", "Gen origen" que hace referencia a la clase "Gen" mediante un *Archetype Slot* al arquetipo *CEN-EN13606-CLUSTER.Gen.v1.adl,* y "Genes diana" que hace referencia a las clases "Diana" (incluyendo como *ELEMENT* el atributo "miTG-score") y "Gen", esta última utilizando un *Archetype Slot* al arquetipo *CEN-EN13606-CLUSTER.Gen.v1.adl*.

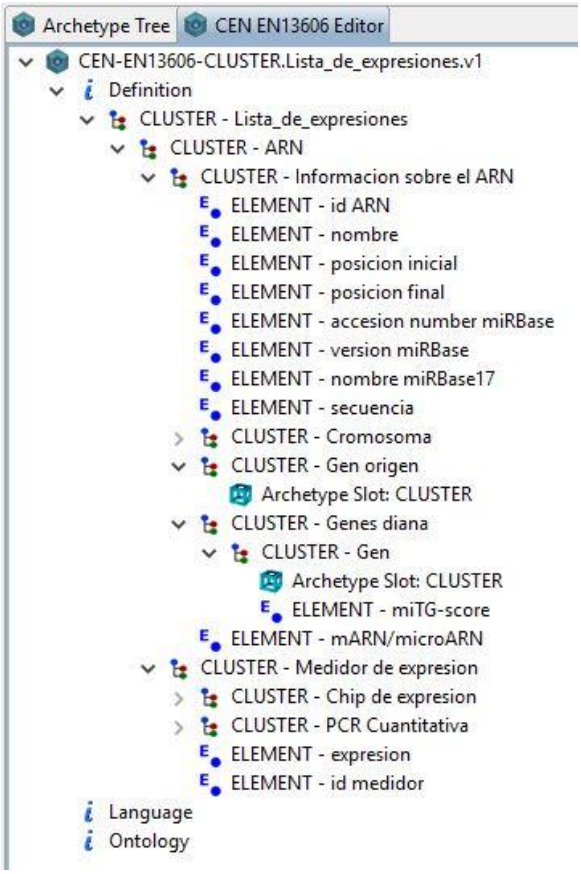

Figura 55. Arquetipo CEN-EN13606-CLUSTER.Lista_de_expresiones.v1.adl

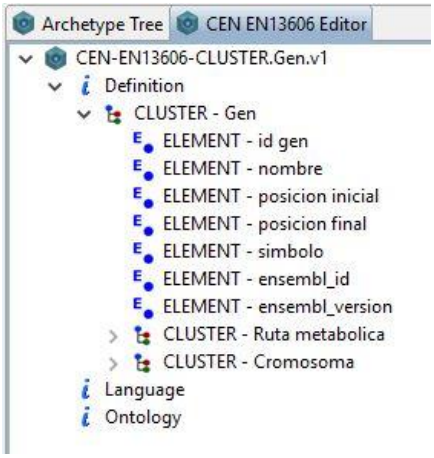

Figura 56. Arquetipo *CEN-EN13606-CLUSTER.Gen.v1.adl*

El último de los arquetipos diseñados haciendo referencia a la vista de Expresión Génica del Modelo Conceptual del Cáncer de Mama es el arquetipo *CEN-EN13606-CLUSTER.Gen.v1.adl*, que podemos ver en la Figura 56. Este arquetipo representa la clase "Gen", con todos sus atributos como *ELEMENTS,* y su relación con las clases "Ruta metabólica" y "Cromosoma", representados por dos CLUSTERS que contienen los atributos de las clases como *ELEMENTS* del *CLUSTER.*

3.2.3. Diseño de los arquetipos relacionados con la perspectiva de Secuenciación Masiva

Utilizando la información que se representa en las clases de la vista Secuenciación Masiva del Esquema Conceptual del Cáncer de Mama hemos diseñado una colección de arquetipos que representan el informe que se le ofrece al paciente tras realizarle un análisis genómico. Los arquetipos diseñados son los siguientes:

- CEN-EN13606-COMPOSITION.Informe_genomico.v1.adl
- CEN-EN13606-CLUSTER.Lista_de_variaciones.v1.adl

El arquetipo *CEN-EN13606-COMPOSITION.Informe_genomico.v1.adl* representa, como hemos comentado anteriormente, el informe genómico de un paciente. Esta *COMPOSITION*, que aparece en la Figura 57, está dividida en tres *SECTIONS* principales. La primera de ellas se llama "Datos de la prueba" y contiene una *ENTRY* "Prueba biológica genómica NGS" que representa las clases "Prueba", "Biológicas" y "Genómica – NGS" del Esquema Conceptual y contiene como *ELEMENTS* todos sus atributos.

La siguiente *SECTION* se llama "Datos del paciente" y contiene información sobre el paciente, en la *ENTRY* "Identificador del paciente" que hace referencia al atributo "NHC" de la clase "Paciente" del Esquema Conceptual, y sobre la muestra sobre la cual se ha realizado el análisis genómico, representada por una *SECTION* "Datos de la muestra" que contiene un *Archetype Slot* para hacer referencia al arquetipo *CEN-EN13606-SECTION.Muestras.v1.adl* definido en el apartado 3.2.1.

Finalmente, la última de las *SECTIONS* principales de esta *COMPOSITION* se llama "Resultados". Esta sección está dividida en tres entradas. La primera de ellas es la *ENTRY* "Descripción de los resultados" con un *ELEMENT* llamado "descripción" que corresponde con el atributo "descripción" de la clase "Prueba". La siguiente entrada recoge los datos sobre el análisis de secuenciación masiva recogidos en el fichero VCF. Esta entrada representa la clase "VCF" del Esquema Conceptual y contiene

como elementos los atributos de la misma. Además, contiene un *CLUSTER* llamado "Filtro" que representa la clase "Filtro" con sus atributos como elementos. La última *ENTRY* de esta sección se llama "Lista de variaciones" y contiene un *CLUSTER* con el mismo nombre que incluye un *Archetype Slot* que hace referencia al arquetipo *CEN-EN13606-CLUSTER.Lista_de_variaciones.v1.adl*.

El último de los arquetipos diseñados se llama *CEN-EN13606-CLUSTER.Lista_de_variaciones.v1.adl*, como podemos ver en la Figura 58. Este arquetipo tiene una estructura de tipo *CLUSTER* que representa todas

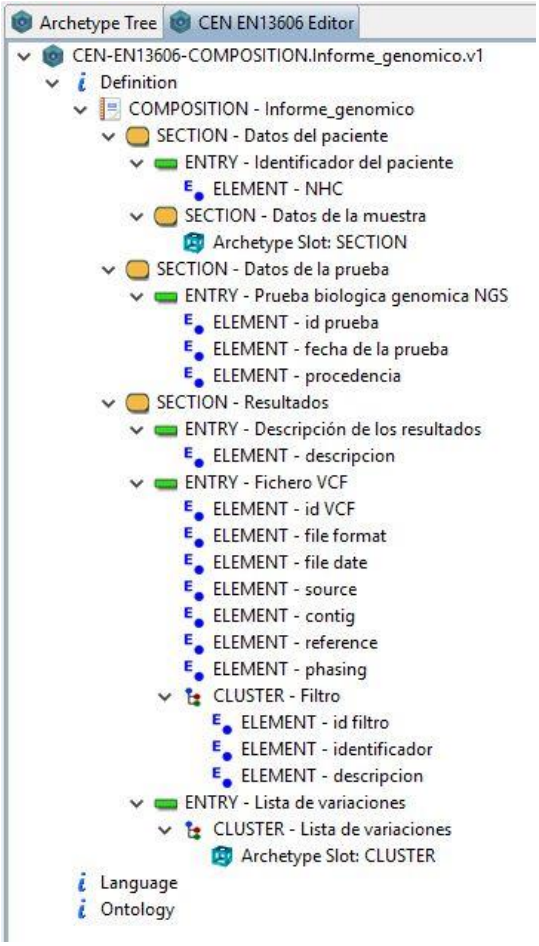

Figura 57. Arquetipo CEN-EN13606-COMPOSITION.Informe_genomico.v1.adl

las variaciones encontradas en el ADN del paciente en su análisis genómico. Para ello, tenemos un *CLUSTER* llamado "Variacion" que representa a cada una de las variaciones. Este *CLUSTER* tiene los datos estructurados, a su vez, en dos *CLUSTERS*: "Descripción" y "Listado de referencias". Además, contiene un conjunto de *ELEMENTS* que se corresponden con los atributos de la clase "Variación", representada por este *CLUSTER.* Dentro del *CLUSTER* "Descripción" encontramos otros *CLUSTERS* llamados "Cromosoma" (que hace referencia a la clase del mismo nombre y contiene como *ELEMENTS* los atributos de la misma), "Genes origen" (que a su vez contiene un *CLUSTER* "Gen" con un *Archetype Slot* que hace referencia al arquetipo *CEN-EN13606-CLUSTER .Gen.v1.adl*) y "Tipo" (que contiene una jerarquía de *CLUSTERS* y *ELEMENTS* que representan las clases de la especialización "Descripción" de la clase variación y que detallan con sus

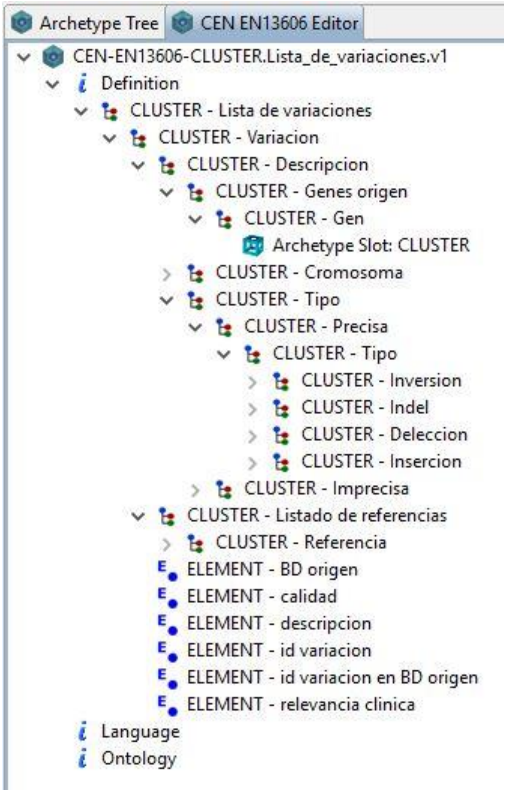

Figura 58. Arquetipo CEN-EN13606-CLUSTER.Lista_de_variaciones.v1.adl

atributos la posición y características de la variación). Por otro lado, dentro del *CLUSTER* "Listado de referencias" encontramos otro llamado "Referencia" que se corresponde con la clase en el Esquema Conceptual con el mismo nombre y contiene como *ELEMENTS* los atributos de la misma.

Cerramos esta sección resaltando la importancia de la utilización de modelos conceptuales como base para el diseño de arquetipos de historia clínica electrónica. Este planteamiento supone mejoras en la gestión de datos clínicos combinando los beneficios de ambas perspectivas, mejorando la gestión de los datos incluidos en los informes clínicos gracias al sistema de información correctamente estructurado al que hacen referencia y también la interoperabilidad semántica que proporciona la utilización de los arquetipos en sistemas de información clínica que utilicen el mismo estándar de información.

3.3. Implementación y carga de la base de datos

3.3.1. Estudio previo a la selección de un Sistema de Gestión de Bases de datos

Una vez completado el Diseño Conceptual y elaborado el Esquema Conceptual correspondiente, hay que pasar a la fase de Diseño Lógico, centrada en la gestión y explotación de los datos y en la que se trata de determinar cuál es la estrategia de almacenamiento de datos más apropiada. Este problema se ubica en el ámbito de los Sistemas de Gestión de Bases de Datos, y obliga a determina qué tipo de base de datos es la más adecuada.

Durante las últimas décadas, el aumento exponencial del volumen de datos que se generan cada día ha dado lugar a la necesidad de llevar a cabo una gestión eficiente que en primer lugar permitiera un adecuado almacenamiento de los mismos y, por otro lado, hiciera posible el acceso directo a la información requerida en cada momento.

Tradicionalmente las bases de datos se han adherido a lo que llamamos el modelo relacional, en el cual los datos se distribuyen en tablas que representan entidades y relaciones. De esta forma, cada fila representa la instancia de una entidad que lleva asociados distintos atributos, a cada uno de los cuales le corresponde una columna determinada. Así, los datos se estructuran de acuerdo con un esquema predefinido. A pesar de seguir un mismo modelo, el mercado ofrece gran cantidad de tecnologías de bases de datos relacionales con distintas características, distintas limitaciones o distintas arquitecturas hardware. Ahora bien, la mayoría de ellas, si no todas, permiten al usuario interaccionar con sus datos a través de SQL (Structured Query Language), un lenguaje declarativo de acceso que se basa en el álgebra relacional.

Durante los últimos años, como respuesta a algunas limitaciones de las bases de datos relacionales, ha surgido una serie de modelos alternativos que, si bien no sustituyen al modelo tradicional, aportan una serie de ventajas (aunque también algunos inconvenientes) que los hacen más adecuados para ciertos tipos de aplicaciones. Este tipo de bases de datos presenta mayor flexibilidad en el modelado de los datos. Pueden estar orientadas a documentos, pueden basarse en un esquema de pares clave-valor, en almacenamiento por columnas, etc. Todas ellas han sido catalogadas denominadas tecnologías NO-SQL.

Dentro de las bases de datos relacionales estudiaremos algunas de las más populares en el entorno clínico, biológico y bioinformático como son SQLite, MySQL y Microsoft Access.

Respecto a las bases de datos no relacionales, también llamadas NoSQL, examinaremos los tres modelos mencionados a través de la tecnología más popular de cada uno de ellos: MongoDB, Redis y Cassandra.

En las conclusiones, partiendo de todo lo que se haya expuesto anteriormente, trataremos de justificar la elección de una de las tecnologías presentadas en relación a los datos clínicos y biológicos que vamos a manejar y el uso que se dará a los mismos.

SQLite

SQLite es una librería software que implementa una base de datos SQL integrada, cuyas principales características son [101, 102]:

- **Autocontenida**: apenas requiere apoyo de librerías externas o del sistema operativo.
- **Sin servidores**: no requiere procesos de comunicación con servidores intermediarios, sino que accede directamente a los archivos de la base de datos que se encuentran en el disco.
- **Base de datos de archivo único:** utiliza archivos ordinarios que pueden encontrarse en cualquier parte de la jerarquía de directorios del sistema, lo que facilita las copias o los envíos por

correo electrónico (portabilidad)

- **Estabilidad multiplataforma:** una base de datos creada en una máquina puede ser copiada y utilizada en otra máquina incluso aunque la segunda tenga distinta arquitectura.
- **Sin configuración**: no requiere instalación o procesos de ajuste previos, ni se necesitan administradores que creen permisos de acceso a los usuarios.
- **Transaccional**: los cambios y consultas se llevan a cabo de manera **A**tómica (o tienen lugar todos los pasos de la transacción o no tiene lugar ninguno), **C**onsistente, (se cumplen las reglas y directrices de integridad de la base de datos, de tal forma que cualquier transacción va de un estado válido a otro estado válido) aislada (***I**solated*, una transacción no puede afectar a otras) y **D**urable (una vez se realice la transacción, esta persistirá). Estas cuatro propiedades (ACID) deben cumplirse incluso en fallos del sistema o caída del suministro eléctrico.
- **Ampliamente utilizada**: probablemente más utilizada que todos los demás motores de bases de datos juntos, pues se utiliza en Android, iOS, Dropbox, Skype, iTunes... Sin embargo, de acuerdo con el *ranking* de popularidad de DB-Engines, se encuentra en la posición número 10 [103].

Además, el código fuente de SQLite es de dominio público, lo que lo hace que se pueda copiar, modificar, publicar o distribuir libremente.

Otras características destacables son su capacidad de soportar bases de datos 140 terabytes y strings y BLOBS de unos 2 gigabytes. Por otra parte, su configuración completa ocupa menos de 500KiB.

Otras particularidades que presenta son, por ejemplo, que el tipo de variable corresponde a cada valor de forma individual, y no necesariamente a cada columna. De este modo, la longitud de cada variable dependerá también de cada valor. Así, si una columna es de tipo VARCHAR(100), no

gastaremos 100 bytes del disco para almacenar un solo carácter de un dato de esa columna, sino un solo byte.

Todo esto se consigue con un código legible y accesible para un programador medio, que añade la posibilidad de utilizar algunas funciones que normalmente no incluyen otros motores de bases de datos.

En el ámbito de la bioinformática en general, SQLite se utiliza de manera rutinaria como herramienta de almacenamiento temporal estructurado. Es habitual, por ejemplo, volcar ficheros de texto plano en una base de datos SQLite para realizar consultas sobre el mismo cuando la línea de comandos es insuficiente.

Discusión de ventajas e inconvenientes

La mayoría de ventajas e inconvenientes de SQLite gira en torno a una de sus características más distintivas: no utiliza servidores. Esto le confiere un gran punto a favor, pues hace que sea más rápida que otras bases de datos. Gracias a ello es especialmente recomendable en dispositivos integrados.

El hecho de que la base de datos se fundamente en archivos en disco la hace extremadamente portátil y se puede guardar simplemente descargando el archivo de la base de datos. Sin embargo, la falta de servidores disminuye la seguridad de la protección frente a errores en la aplicación cliente (en los sistemas cliente/servidor, problemas en aplicación cliente no causarían daños en el servidor), pero a su vez la hace más segura en el sentido de que permite que cada usuario pueda tener una base de datos completamente independiente.

Además, tener una arquitectura sin servidor limita la capacidad de almacenaje de datos (en la mayoría de casos está limitada a 2 Gb) y dificulta la concurrencia de múltiples aplicaciones escribiendo a la vez.

Se valora también como un inconveniente el hecho de que no incluya la posibilidad de gestión de la base de datos por parte del usuario, lo cual no permite, por ejemplo, crear accesos privilegiados.

Otras fortalezas son que ocupa muy poco espacio en la memoria, es autocontenida y mantiene la integridad de los datos gracias a su atomicidad, aunque en algunos casos se considera deficiente en cuanto a replicabilidad.

Ahondando un poco más en el tema de la memoria, aunque el motor SQLite ocupa poco espacio, una base de datos en SQLite ocupa más memoria que la misma base de datos con, por ejemplo, MySQL aunque después, en su utilización, necesitará menos recursos.

A pesar de la velocidad en las consultas, una gran desventaja es su lentitud en la escritura, ya que durante la misma bloquea el archivo de la base de datos por completo y el acceso por lecturas también puede ralentizarse excesivamente en tamaños de bases de datos demasiado grandes.

ACCESS

Access es un sistema de bases de datos relacional desarrollado por Microsoft que se distribuye como parte del paquete de Microsoft Office [104, 105]. Sus herramientas principales incluyen la introducción de datos, la realización de consultas o la producción de informes entre otras operaciones:

- **Introducción de datos:** Incluye una forma muy intuitiva para introducir datos de forma manual a través de formularios. Con ayuda de los formularios (aunque también se puede editar directamente las tablas), el usuario puede crear tablas y definir las columnas de las mismas, además de controlar cómo otros usuarios pueden interactuar con la base de datos. A través de la hoja de navegación se pueden seleccionar tablas, informes o consultas rápidamente a través de un clic.
- **Importación y exportación:** Las posibilidades de importación de datos se encuentran a través de un menú de "Datos externos". Una de estas herramientas incluye un sistema de ayuda a través del cual se va indicando al usuario cómo trasladar los datos de una hoja de Microsoft Excel a una base de datos Access. También permite la

exportación de datos a Microsoft Word, Excel o páginas web.

- **Informes:** Resumen y presentan los datos de las tablas. Se pueden diseñar para presentar la información de forma óptima y actualizada y se suelen utilizar formatos fácilmente imprimibles y exportables. Los informes se realizan a través de un sistema de ayuda que va solicitando los parámetros necesarios para definir el informe. Sin ellos, la capacidad de visualización de datos de Access sería mucho más limitada.
- **Consultas:** A menudo sirven de origen de registros para formularios o informes. Las consultas en Access son actualizables, es decir, podemos editar la base de datos a través de las hojas de consulta. Las consultas se distinguen en dos tipos: de selección y de acción. Mientras que las primeras simplemente recuperan los datos, las segundas llevan a cabo una tarea con ellos (creación de nuevas tablas, eliminación de datos...). Estas se pueden llevar a cabo mediante lenguaje SQL aunque también se ofrece una interfaz mucho más intuitiva y amigable.
- **Macros:** Se consideran un lenguaje de programación simplificado que aumenta la funcionalidad de la base de datos. Los macros contienen acciones que realizan tareas de forma automática, lo que supone un gran ahorro de tiempo.
- **Módulos:** También se ocupan de aumentar la funcionalidad de la base de datos. Se trata de una serie de declaraciones, instrucciones y procedimientos que se almacenan como una unidad y se escriben en el lenguaje de programación Visual Basic para aplicaciones.

Al igual que SQLite, Access trabaja con toda la base de datos en un único archivo, aunque también permite el acceso a servidores SQL. De acuerdo con Microsoft, el tamaño máximo de una base de datos Access es de 2 GB. En los últimos 16 años se han desarrollado 7 versiones de Access (2000, 2002, 2003, 2007, 2010, 2013 y 2016) que han ido tratando de mejorar la fiabilidad, la funcionalidad y la rapidez.

Discusión de ventajas e inconvenientes

Respecto a Microsoft Access encontramos puntos de vista en ambos extremos. Mientras algunos usuarios la consideran una base de datos "de juguete", interesante para abordar problemas de una magnitud simple, pero inapropiada para usos más exigentes, en el lado contrario encontramos usuarios que lo defienden principalmente por su gran facilidad de uso en gran cantidad de situaciones. Lo cierto es que de acuerdo con el *ranking* de popularidad de DB-Engines, se sitúa en el número 8, por encima de otras como SQLite o Redis [103].

La principal ventaja de Access es que proporciona un entorno muy familiar para la inmensa mayoría de usuarios. Para cualquier persona acostumbrada a Office le resultará sencillo trabajar con esta base de datos. Además, es considerablemente más barata que otras bases de datos (Access 2016 cuesta 135€ [106]), aunque esto es una desventaja si lo comparamos con los sistemas de bases de datos de dominio público.

Tanto en la creación de una nueva base de datos como en la apertura de una ya existente existen dos vistas posibles: la de datos y la de diseño, lo que facilita su manejo. Además, su instalación es sencilla, en cualquier paso el usuario cuenta con un sistema de ayuda y con métodos que permiten implementar sistemas para asegurar la correcta introducción de datos.

Access es un estándar en la industria para usos "de escritorio" y el motor de la base de datos es bastante potente. El reconocimiento de voz y la posibilidad de personalizarlo hace que pueda ajustarse a las necesidades específicas de cada usuario y, puesto que no requiere programación compleja, prácticamente cualquier persona puede crear bases de datos avanzadas.

Por otra parte, presenta dificultades a la hora de trabajar con bases de datos mayores de 2GB y por seguridad se suele limitar su tamaño a 1GB. Además, el hecho de que sea una base de datos de archivo único dificulta la escalabilidad. A medida que aumentamos la base de datos, Access se

ralentiza (si superamos las decenas de miles de filas en una tabla, las búsquedas pasan de ser de segundos a minutos).

MySQL

MySQL es un sistema de gestión de bases de datos relacional [107, 108]. Fue desarrollado por *MySQL AB*, que en 2008 fue absorbida por *Sun Microsystems* y esta a su vez por *Oracle Corporation* en 2009. Las principales características que presenta son:

- **Arquitectura cliente/servidor:** se trata de una base de datos servidor (MySQL) y un gran número de clientes (programas de aplicaciones) que comunican con él. No es necesario que las aplicaciones cliente y el servidor se encuentren en la misma máquina.
- **Compatibilidad SQL:** MySQL se adhiere al estándar SQL. Ajustando la configuración de modo SQL podemos hacer que el servidor sea compatible con varios sistemas de bases de datos como IBM/DB 2 y Oracle.
- **SubSELECTs:** desde la versión 4.1 se pueden realizar subconsultas.
- **Vistas:** una consulta con SQL puede dar lugar a una vista particular que hace que veamos una base de datos distinta.
- **Procesos almacenados:** son procesos automáticos que simplifican ciertos pasos como la inserción o el borrado de ciertos datos. Son útiles en la administración de bases de datos muy grandes, incrementando la eficiencia.
- ***Triggers*:** son comandos que se ejecutan de forma automática por el servidor en ciertas operaciones (INSERT, UPDATE y DELETE).
- **Unicode:** MySQL soporta casi cualquier conjunto de caracteres: *latin1, german, big5, ujis*... Caracteres escandinavos como "å", "ä" y "ö" pueden introducirse en nombre de tablas o columnas.
- **Interfaz:** presenta numerosas interfaces adecuadas para la administración del servidor.
- **Búsqueda en texto completo:** esto simplifica y acelera la búsqueda

de palabras que se encuentran en un campo de texto. Esto hace mucho más eficiente las búsquedas si usamos MySQL para almacenamiento de textos (como por ejemplo un foro de discusión).

- **Replicación:** la replicación permite que el contenido de la base de datos se replique en otros ordenadores. Esto permite, por un lado, mejorar la protección del sistema frente a fallos y, por otro lado, mejorar la velocidad de las consultas.
- **Transacción:** es la ejecución de varias tareas en bloque, de manera que se llevan a cabo o todas de manera correcta o ninguna. Esto ocurre también incluso aunque en medio de una transacción el sistema falle. Gracias a InnoDB, MySLQ permite realizar transacciones ACID.
- **Restricciones de clave ajena:** esto asegura que en tablas ligadas no existen referencias que llevan a ninguna parte.
- **GIS** (***Geographic Information System)*:** Soporta el almacenamiento y procesado de datos geográficos bidimensionales.
- **Lenguajes de programación:** Existe un gran número de APIs y librerías para el desarrollo de aplicaciones de MySQL. Para programación del cliente se puede usar C, C++, Java, Perl, Python, PHP... entre otros. Soporta también la interfaz ODBC.
- **Independencia de sistemas operativos:** el propio servidor puede ejecutarse en Apple Macintosh OS X, Linux, Microsoft Windows y todas las variantes de Unix.
- **Software libre:** MySQL se distribuye con licencias de código abierto, lo que te permite poder modificar la fuente del programa sin restricciones de licencia y de manera gratuita.

Son varias las bases de datos relevantes en bioinformática que utilizan MySQL como sistema de gestión como, por ejemplo, Ensembl [22], HGMD [109], COSMIC [37] o dbSNP [18].

Discusión de ventajas e inconvenientes

MySQL es una de las bases de datos relacionales más populares del mundo. Aunque desde hace un par de años parece perdiendo apoyos y algunos usuarios aseguran que estamos viendo poco a poco la caída del sistema, ocupa el puesto número 2 en el *ranking* de popularidad de DB-Engines [103]. Estas afirmaciones parecen olvidar algunas ventajas claras que MySQL puede aportar a las organizaciones.

Una de ellas es su facilidad de uso. Siempre y cuando el usuario entienda el lenguaje SQL, no debería tener problemas a la hora de trabajar con el sistema. Además, resulta fácil de instalar y gracias a una serie de herramientas adicionales la configuración e implementación se convierten en tareas sencillas.

La popularidad de MySQL ha dado lugar a una gran comunidad de desarrolladores. No hay escasez de expertos, lo que permite que el usuario pueda disponer de ayuda prácticamente cuando la necesite. Se podría considerar un estándar en la industria, en parte gracias a que es compatible con prácticamente cualquier sistema operativo y soporta un gran número de interfaces de desarrollo de software.

A pesar de la compra de MySQL (a través de *Sun Microsystems*) por parte de *Oracle*, todavía se considera software libre y está accesible libremente en su web.

Una de las mayores ventajas de MySQL radica en sus sólidos sistemas de seguridad que protegen ciertos datos frente a intrusos y otorga privilegios a algunos usuarios.

También se considera un punto a favor su velocidad a pesar de que no es tan rápido como SQLite.

En cuanto a escalabilidad, puede manejar hasta 50 millones de filas y aunque por defecto el tamaño de los ficheros es de 4GB, teóricamente esto se podría ampliar hasta 8TB de datos. Por otra parte, una misma base de datos ocupa una magnitud similar con MySQL que con SQLite. También

ha demostrado ser un sistema adecuado para prevenir problemas de pérdidas en la memoria.

Algunos expertos lo consideran mejor que SQLite en cuanto a herramientas de gestión y creación de informes y también más fiable a la hora de enfrentarse a fallos del sistema o caídas del suministro eléctrico [110].

Una debilidad que se le confiere es el hecho que desde que *Oracle* se hizo con MySQL, la comunidad de expertos en MySQL ha ido perdiendo influencia y muchos desarrolladores ven cada vez más difícil el diálogo con *Oracle* acerca de la gestión del sistema.

Por otro lado, a diferencia de SQLite que es autocontenida, la funcionalidad de MySQL depende en gran medida de complementos externos para características tales como la búsqueda de texto o las características ACID (InnoDB) aunque esto lo hace altamente personalizable.

Otro inconveniente añadido es que en general consume más recursos que SQLite, ya que requiere establecer una conexión cliente/servidor, mantenerla abierta, estar enviando y recibiendo datos...

En conclusión, MySQL resulta adecuado en situaciones en las que se requiera alta seguridad en el control de acceso a los datos, aplicaciones y páginas web, además de en la búsqueda de soluciones muy personalizadas (gracias a la variedad de posibles configuraciones). Sin embargo, no es tan recomendable en caso de requerir una gran adherencia al estándar SQL (por ejemplo en el caso de necesitar conectar con otras bases de datos) o cuando se esperen altos niveles de concurrencia con un ratio escritura/lectura elevado.

Redis

Redis se define como un servidor de estructura de datos [111]. No es una base de datos relacional, sino que utiliza un esquema “clave-valor”. Esto significa que los datos no están ordenados en tablas, sino en instancias de valores, cada una de ellas identificadas por una clave. Todas las instancias

no contendrán el mismo tipo de información, al igual que ocurría en las distintas filas de una misma tabla, sino que cada una, de forma independiente, llevará asociado un conjunto de pares "clave-valor" (el primero el nombre de la variable y el segundo su valor). Una diferencia importante respecto a otras bases de datos "clave-valor" es que puede contener tipos de datos como *strings, hashes, lists, sets, sorted sets, bitmaps* y *hiperlogs*.

Por otro lado, mientras que la mayoría de sistemas de gestión de datos almacena los mismos en la memoria secundaria (cosa que ralentiza mucho las operaciones), Redis almacena todo en la memoria primaria, que es mucho más rápida a la hora de leer y escribir datos.

Presenta una arquitectura cliente/servidor. El servidor se encarga de almacenar los datos en memoria y gestiona la arquitectura de datos. El cliente puede ser una consola de Redis o cualquier otra API con un lenguaje de programación admisible. Cliente y servidor no es necesario que se encuentren en el mismo ordenador.

Redis almacena todo en la memoria primaria, pero esta es volátil y los datos se perderían cada vez que reiniciáramos el servidor, por lo que se requiere algún sistema de persistencia de datos.

Sin embargo, no incluye mecanismos de copia de seguridad o recuperación de datos, para lo cual se requieren servidores de terceros. Aun así, si se utiliza Redis en entornos replicados facilita la realización de copias de seguridad.

Discusión de ventajas e inconvenientes

Una de las principales ventajas es su enorme velocidad: puede llevar a cabo más de 100.000 SETs (definiciones de variables) por segundo y más de 80.000 GETs (recuperaciones de variables). Además, realiza una gestión eficiente de la concurrencia. Redis es software abierto y su instalación es sencilla.

En cuanto a los principales inconvenientes, Redis es un sistema de almacenamiento de datos en memoria primaria y esto limita enormemente la cantidad de datos disponibles. Mientras que en el resto de bases de datos se puede trabajar con más capacidad de la que se tenga en la memoria primaria, esto no ocurre con Redis. Además, para conseguir mayor velocidad también se compromete parte de la memoria, de forma que para ciertas tareas la diferencia entre el número de bytes de datos y la memoria que utiliza puede ser diez veces superior.

Por otro lado, al ser un servidor de estructuras de datos, no existe el lenguaje de consulta ni es compatible con el álgebra relacional. Todas las posibles formas de acceso a los datos deben ser anticipadas por el desarrollador para poder diseñar una vía adecuada de acceso a los datos. Esto supone una gran pérdida de flexibilidad.

Haciendo referencia al *ranking* de DB-Engine [103], nos encontramos este sistema de gestión de bases de datos en el puesto 9, por debajo de todos los sistemas vistos hasta ahora, a excepción de SQLite.

Cassandra

Cassandra es un software abierto para la gestión de bases de datos distribuidas que fue originalmente desarrollado en *Facebook*, se hizo público en 2008 y pasó a ser parte del proyecto *Apache* en 2010 [112, 113]. Algunas de sus características son:

- **Masivamente escalable:** una arquitectura sin un nodo maestro en el que todos los nodos son igual de importantes (arquitectura de anillo) proporciona simplicidad operacional y facilita la escalabilidad.
- **Disponibilidad continua:** la redundancia tanto a nivel de datos como de nodos elimina posibles fallos y proporciona una alta disponibilidad. Esto le confiere también un alto nivel de seguridad ante la pérdida de datos.
- **Modelos de datos flexibles y dinámicos:** soporta tipos de datos

modernos que permiten lectura y escritura rápida. Estos son: numéricos (int, biging, decimal...), caracteres (*ascii, varchar...*), fechas (*timestamp*), desestructurados (*blob*) y especializados (*set, list, map...*)

- **Compresión de datos:** se puede alcanzar un nivel de compresión de hasta el 80% sin sobrecarga en el rendimiento.
- **CQL:** lenguaje similar a SQL que facilita en gran medida la migración desde bases de datos relacionales.
- **Escritura:** los datos se graban primero en disco y después en una estructura en memoria llamada *memtable*. Cuando una *memtable* supera un umbral, los datos se escriben en un fichero inmutable en disco llamado *SSTable*.
- **Lectura:** Cassandra consulta una estructura de datos llamada filtro *Bloom* que indica la probabilidad de que una *SSTable* contenga los datos requeridos, y en función de la respuesta se realizará la consulta en el fichero o pasará a otro.
- **Distribución de datos:** A lo largo de la estructura de anillo los datos se distribuyen de forma automática.
- **Replicación:** la replicación se define a nivel de "espacio de claves" o "*keyspace*", lo que permite que distintos espacios de claves tengan distintos modelos de replicación.
- **Centro multi-datos y soporte en la nube:** los datos pueden estar repartidos y replicados geográficamente.
- **Modelo de datos:** Presenta diferencias significativas con respecto al modelo *entidad-relación*. Por ejemplo, el sistema *entidad-relación-atributo* no tiene sentido en Cassandra. Sus objetos son *espacios de claves, tablas*, *claves primarias* e *índices*.
- **Transacciones:** son atómicas, aisladas y durables.
- **Seguridad:** Permite gestionar el acceso a los datos en función de los usuarios e incluye sistemas de criptografía en la comunicación.

Discusión de ventajas e inconvenientes

La gran ventaja que presenta Cassandra se encuentra en su arquitectura de anillo en la que no hay nodos más importantes que otros, lo que hace que no disponga de un punto único de fallo. Además, podemos añadir al *cluster* todos los nodos que queramos o quitarlos, lo que lo convierte en altamente escalable en ambas direcciones sin necesidad de reiniciar las aplicaciones o el *cluster* y sin sufrir retrasos en su funcionamiento.

Por otra parte, dado su sistema de replicación, incluso si un nodo falla se podrá requerir a otro para la recuperación de datos, lo que convierte a Cassandra en un sistema altamente resistente al fallo y con altas competencias en lo que a recuperación de datos se refiere.

También algunos usuarios ven como punto a favor el hecho de que Cassandra sea una base de datos sin esquema o de esquema opcional. A diferencia de las bases de datos relacionales, no hay necesidad de mostrar todas las columnas que la aplicación necesita ya que no todas las filas tendrán el mismo conjunto de columnas.

El hecho de que sea código abierto, además del ahorro que supone, ha permitido desarrollar una gran comunidad de usuarios que comparten sus puntos de vista, consultas y sugerencias en torno al desarrollo de aplicaciones *Big Data*.

Otros usuarios ven como desventajas la falta de integridad referencial y están de acuerdo con que las consultas son limitadas: pensamos cuáles serán las búsquedas más comunes y estructuramos la base de datos en función de esto. También echan de menos algún sistema para ordenar la aparición de los datos en las búsquedas, ya que aparecen según el orden predefinido en el diseño.

Todas estas características hacen de Cassandra un sistema gestor de bases de datos muy recomendable en aplicaciones de internet de las cosas, monitorización y seguimiento de usuarios, servicios de mensajería, análisis de redes sociales o para aplicaciones basadas en series temporales.

MongoDB

MongoDB es un sistema gestor de bases de datos de propósito general y código abierto que fue lanzado en 2009 [114-116]. Presenta las siguientes características principales:

- **Base de datos orientada a documentos:** Una base de datos MongoDB almacena una serie de colecciones, cada una de las cuales, a su vez, contiene documentos. Un documento JSON (JavaScript Object Notation) es un conjunto de pares clave-valor con esquema libre.
- **Alto rendimiento:** Las relaciones entre documentos se pueden llevar a cabo mediante referencias o integrando documentos.
- **Consultas:** MongoDB proporciona una serie de operadores que giran en torno a *find()* y actúan por medio de documentos de especificación de consultas. A pesar de no ser lenguaje SQL, tanto la lectura como la escritura deben ser sencillas y rápidas.
- **Arquitecturas de despliegue:** Aunque MongoDB soporta su despliegue en un único servidor, por defecto en entornos de producción se asume un despliegue distribuido. Este puede tratarse de un despliegue mediante conjuntos de replicación con tolerancia a fallos automática o con *clusters* que hacen posible la partición de datos en distintas máquinas de manera transparente para el usuario.
- **Potencia:** Proporciona muchas de las características de las bases de datos relacionales tales como índices secundarios, consultas dinámicas, ordenación y agrupación, etc. con la flexibilidad y escalabilidad que permiten los modelos de bases de datos no relacionales.
- **Autoescalable:** La autofragmentación permite aumentar un *cluster* de manera lineal añadiendo más ordenadores.
- **Configuración automática:** Aunque se pueden personalizar para la propia aplicación, MongoDB proporciona una configuración previa que le permite funcionar inmediatamente después de su

instalación.

- **Sistema operativo:** Funciona en Linux, Windows y OS X, tanto en 64 como 32 bits, aunque en este último caso las bases de datos se ven limitadas a 2GB.
- **Procesado y MMS:** Incluye herramientas de cálculo y procesamiento de datos. Existe además el *MongoDB Management System*, una aplicación web que permite monitorizar las bases de datos y los ordenadores, además de hacer copias de seguridad de los datos.
- **Seguridad:** Permite utilizar controles de acceso a los datos en función de los usuarios e incluye sistemas de criptografía en las comunicaciones.

Discusión de ventajas e inconvenientes

MongoDB ocupa el puesto número 4 en el *ranking* de popularidad de DB-Engines [103] y es la primera si hablamos de bases de datos NoSQL.

Gran parte de sus ventajas se deben principalmente a tratarse de una base de datos orientada a documentos, en los cuales muchos ven un punto a favor por su claridad estructural. De esta forma, cuando llevamos a cabo consultas, conseguimos un documento completo, identificado por una clave. Así, las búsquedas en MongoDB son "clave-valor", no relacionales, lo que hace que sean mucho más rápidas en algunos casos. Esta ventaja solo es real cuando tus datos son auténticamente documentos y no tratan de imitar las tablas relacionales (por ejemplo, al realizar consultas que necesiten combinar varias claves), lo que enlentecería mucho las búsquedas.

Su segunda gran ventaja es la escalabilidad, pues resulta muy sencillo escalar horizontalmente añadiendo nodos gracias a los mecanismos automáticos de fragmentación de datos y tolerancia frente a fallos. Cuando un determinado nodo supera un umbral de volumen, los datos se reorganizan para distribuirse uniformemente.

Otro punto positivo es la libertad de esquema en el documento y la facilidad de comunicación con la base de datos mediante JavaScript, PHP o Python.

Tampoco debemos olvidar que se trata de un software abierto, lo que abarata en gran medida los costes de nuestra aplicación.

Sin embargo, algunos usuarios expertos ven en el diseño de documentos integrados un problema, ya que al contener en un documento toda la información que necesitas no existe la operación JOIN. Esto conduce a que en ocasiones debamos realizar distintas búsquedas y unirlas de forma manual, lo cual complica el código y reduce la flexibilidad cuando la estructura cambia.

Otra desventaja que presenta es el elevado uso de memoria, ya que en cada documento debe almacenar cada clave debido a las diferencias en las estructuras de distintos documentos. Además, debido a que no hay posibilidad de realizar JOIN, nos vemos obligados a duplicar los datos, o bien a utilizar el sistema de referencias (en estos casos tal vez sería más interesante utilizar una base de datos relacional).

Por último, también presenta algunas desventajas en cuanto a concurrencia, ya que cada vez que se lleva a cabo una operación de escritura la base de datos entera se bloquea durante la operación, impidiendo otras escrituras o lecturas.

Conclusiones

Tras el profundo análisis de las tecnologías de gestión de bases de datos realizada, nos decantaremos por utilizar una base de datos relacional para la implementación de nuestro sistema de información. Los motivos que nos han llevado a tomar esta decisión abarcar distintos frentes.

Una de las ventajas es que asegura la integridad y consistencia de los datos. Esto es muy importante en un entorno como el sanitario, ya que los datos deben cumplir unos requisitos de calidad, sin posibilidad de errores, que quedan reforzados con este tipo de sistemas. Además, la atomicidad de las

operaciones en la base de datos asegura que las acciones dirigidas a la base de datos se realicen en su totalidad utilizando la técnica del *rollback* en el caso en el que ocurra algún error en el procedimiento evitando, de esta manera, que los datos queden perdidos o incompletos. Estas dos características son esenciales en el entorno de los datos médicos y la mayoría de las bases de datos NoSQL carecen de ellas, soportando lo que se llama "consistencia eventual".

Por otra parte, debido al largo tiempo que llevan en el mercado, los sistemas de gestión de bases de datos relacionales tienen un mayor soporte y mejores "suites" de productos y "add-ons" para gestionar estas bases de datos, lo que facilita el trabajo con ellas. Las NoSQL suelen tener herramientas de gestión más complicadas o de acceso por consola, lo que supone un mayor tiempo de aprendizaje y adaptación.

Además, la tecnología SQL permite combinar de forma eficiente diferentes tablas para extraer información relacionada, mientras que NoSQL no lo permite o lo hace de forma bastante limitada. SQL permite gestionar los datos junto con las relaciones existentes entre ellos; en NoSQL no existe este tipo de utilidades. En un entorno como el de la investigación traslacional en el cáncer de mama es muy importante tener correctamente almacenados y relacionados todos los datos de forma correcta en una base de datos para poder realizar consultas e inferir nueva información que pueda mejorar el diagnóstico y tratamiento de los pacientes.

Una vez definido el tipo de tecnología a utilizar, debemos seleccionar cuál de los sistemas de gestión de bases de datos relacionales vamos a utilizar y, en este caso, nos decantamos a utilizar la tecnología de MySQL para implementar nuestra base de datos. Esta selección se ha realizado por varios factores entre los que destacan su facilidad de uso, instalación y manejo gracias al kit de herramientas que ofrece de forma gratuita y que permite realizar las tareas de diseño y configuración de la base de datos de una forma sencilla para el usuario. Además, esta tecnología viene implícita con unos sólidos sistemas de seguridad que permiten configurar una alta protección de los datos clínicos y genómicos de los pacientes frente a

cualquier tipo de ataque o amenaza externa y, además, la gestión de usuarios y sus privilegios. También cabe destacar su popularidad, ya que ocupa el puesto número 2 en el *ranking* de popularidad de DB-Engines [103], lo que le permite tener una amplia comunidad detrás que facilita el aprendizaje y la resolución de posibles dudas y errores.

Para complementar esta tecnología y ofrecer un acceso y consulta de datos sencillo a los usuarios, se ha planteado la posibilidad de utilizar la tecnología de Access y su generador de formularios. Se ha seleccionado esta tecnología, principalmente por la familiaridad de su entorno con nuestros usuarios finales, los clínicos. En el entorno clínico es muy común el uso del paquete de Microsoft Office para múltiples tareas, lo que les ayuda a ver la herramienta más familiar, accesible y sencilla de utilizar. Otra característica que ha decantado la elección de esta tecnología son sus amplias posibilidades de conectividad (tablas de Excel, servidores SQL o SharePoint...) y la capacidad de generar informes que faciliten la visión y el análisis de datos procedentes de estas fuentes.

La combinación de la potencia y seguridad de MySQL y la familiaridad y facilidad de uso de los formularios de Access nos permitirán crear una herramienta adecuada para la gestión de datos clínicos y genómicos que se plantea como uno de los objetivos de esta tesis.

3.3.2. Inferencia del Esquema de base de datos e implementación de la base de datos en MySQL

Esta base de datos ha sido implementada a partir de las vistas Clínica, de Expresión Génica y de Secuenciación Masiva del Esquema Conceptual del Cáncer de Mama diseñadas y detalladas en los capítulos 3.1.1, 3.1.2 y 3.1.3 de esta tesis. Este esquema conceptual ha servido de base para generar el modelo de base de datos necesario para implementar esta base de datos,

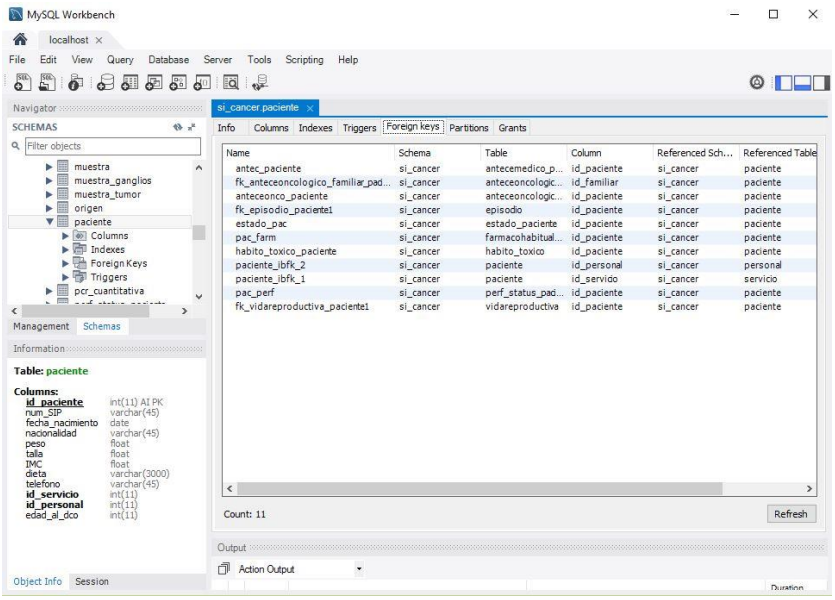

Figura 59. MySQL Workbench 6.3 CE. Hoja de propiedades de la tabla "Paciente"

mejorando la calidad de los datos y las posibilidades de explotación fructífera en diversos contextos.

Debido a que los conceptos que se manejan en el esquema de base de datos son los mismos que en el esquema conceptual descrito en capítulos anteriores de esta tesis, no se vuelven a explicar ya que sería innecesariamente redundante.

Tras las conclusiones obtenidas en el estudio planteado en el capítulo 3.3.1, la base de datos se ha implementado utilizando la tecnología del Sistema de Gestión de Base de Datos MySQL. Se ha usado la potente herramienta visual del MySQL Workbench 6.3 CE (ver Figura 59) para diseñar cada una de las tablas de la base de datos, junto con las claves primarias, ajenas y restricciones de integridad inferidas del esquema conceptual.

3.3.3. Integración de la información pública relevante para el Cáncer de Mama en la Base de Datos

Para poder realizar consultas a la base de datos y poder extraer la información que se necesita para realizar análisis genómicos en muestras de Cáncer de Mama, se ha realizado la carga de la información de las bases de datos más consultadas por los expertos a la hora de realizar análisis de muestras de esta enfermedad. Podemos dividir estas bases de datos en varios grupos, teniendo en cuenta en tipo de información que contienen.

Las bases de datos de las cuales hemos extraído información sobre las variaciones referentes al cáncer de mama son dbSNP, perteneciente al grupo de bases de datos de NCBI (*National Center for Biotechnology Information* de EEUU) y BIC (*Breast Cancer Information Core*, del *National Human Genome Research Institut* de EEUU). Como hemos detallado en el capítulo del estado del arte dedicado a las bases de datos con información genómica, dbSNP contiene información sobre múltiples especies y enfermedades, mientras BIC se centra en variaciones de los genes BRCA1 y BRCA2 relacionadas con el Cáncer de Mama en humanos específicamente. Hemos tomado estas dos bases de datos por dos motivos: por ser las dos más utilizadas por los expertos a la hora de realizar análisis genómicos sobre la enfermedad y por ser una de carácter genérico y otra de carácter más específico de la enfermedad.

En cuanto a información sobre microARNs hemos seleccionado la base de datos miRBase. Nos centramos en este tipo de micromoléculas porque son de las más estables en cualquier tipo de tejido a nivel biológico, su papel modulador de la expresión génica y su análisis es bastante sencillo por la cantidad limitada de microARNs a analizar. Como comentábamos en el capítulo 2.2.2 del estado del arte de esta tesis, esta es la base de datos de referencia cuando nos referimos a información sobre los microARNs y la que establece la nomenclatura que se utiliza en el entorno biológico para estas moléculas.

Para relacionar los datos de microARNs cargados desde miRBase con los genes afectados por la sobreexpresión de ciertos microARNs hemos extraído los datos de la base de datos de DIANA microT-CDS. Esta base de datos nos permite hacer una predicción sobre qué genes quedan afectados por alteraciones en la expresión de ciertos microARNs. Nos hemos basado en la opinión y práctica de los expertos para hacer la selección de esta fuente de información.

Finalmente, hemos cargado la información de las bases de datos KEGG, Reactome y Gene Ontology disponibles en la plataforma EnrichR para relacionar los genes con los *pathways* en los que están involucrados. Esta selección también ha sido dirigida por biólogos profesionales que utilizan estas herramientas en su trabajo diario.

Al tratarse de bases de datos distintas, con datos distintos y organismos de origen distintos, han de tratarse cada una de ellas de forma totalmente individualizada para extraer de ellas únicamente la información que consideramos relevante para nuestra base de datos. Es importante destacar de nuevo que la posibilidad de realizar una gestión unificada a través del uso de modelado conceptual de una información tan diversa y compleja como la analizada, conforma el eje del valor del trabajo presentado en esta Tesis Doctoral.

Carga de los datos de variaciones genómicas relacionados con el Cáncer de Mama: dbSNP y BIC.

dbSNP

De esta base de datos de SNPs genérica vamos a extraer las variaciones sobre Cáncer de Mama que nos interesan. Tras varios intentos fallidos de carga, debido al gran tamaño que tienen los ficheros que dbSNP ofrece en su FTP con toda la información que se almacena de sus variaciones, se decidió estudiar otras opciones, entre ellas: dividir los ficheros en trozos más pequeños, optimizar el código disminuyendo el número de consultas a la base de datos, extraer la información utilizando la herramienta BioMart de Ensembl o estudiar un nuevo fichero, existente también en el FTP de dbSNP, que contiene todas las variaciones de la base de datos pero filtradas siguiendo el formato VCF.

Esta última opción nos parecía la más conveniente, pero primero decidimos hablar con un experto bioinformático del Hospital Clínico de Valencia, para cerciorarnos de que a partir de los datos proporcionados por el fichero podíamos obtener a toda la información que necesitamos. Este fichero contiene por cada variación información sobre el cromosoma en el que se encuentra, su posición (cromosómica), las secuencias de nucleótidos insertados y borrados, además del identificador que dbSNP proporciona a cada una de ellas (idenfiticador rs).

Los siete pasos que se han llevado a cabo para realizar el proceso de carga representado en la Figura 60 han sido los siguientes:

- **Paso 1: Crear un fichero de texto con los identificadores (NC) de los cromosomas y el nombre**. El primer paso para resolver esta tarea ha sido la extracción de la información que nos interesa de cada cromosoma de la base de datos:
    - Identificador NC
    - Nombre del cromosoma

Se tomó la decisión de realizar este fichero de forma manual debido a que trabajamos con una versión del genoma concreta y los cromosomas no varían hasta que el responsable de la base de datos no cambie de versión (lo que ocurre una vez cada cuatro años, aproximadamente). Es por esto que este fichero no hace falta que se genere cada vez que se actualicen las variaciones de dbSNP.

- **Paso 2: Crear un fichero de texto con los identificadores (NG) de los genes que nos interesan.** En este caso hemos descargado los genes relacionados con el cáncer de mama que nos han indicado los expertos y la información relacionada con los mismos. Para

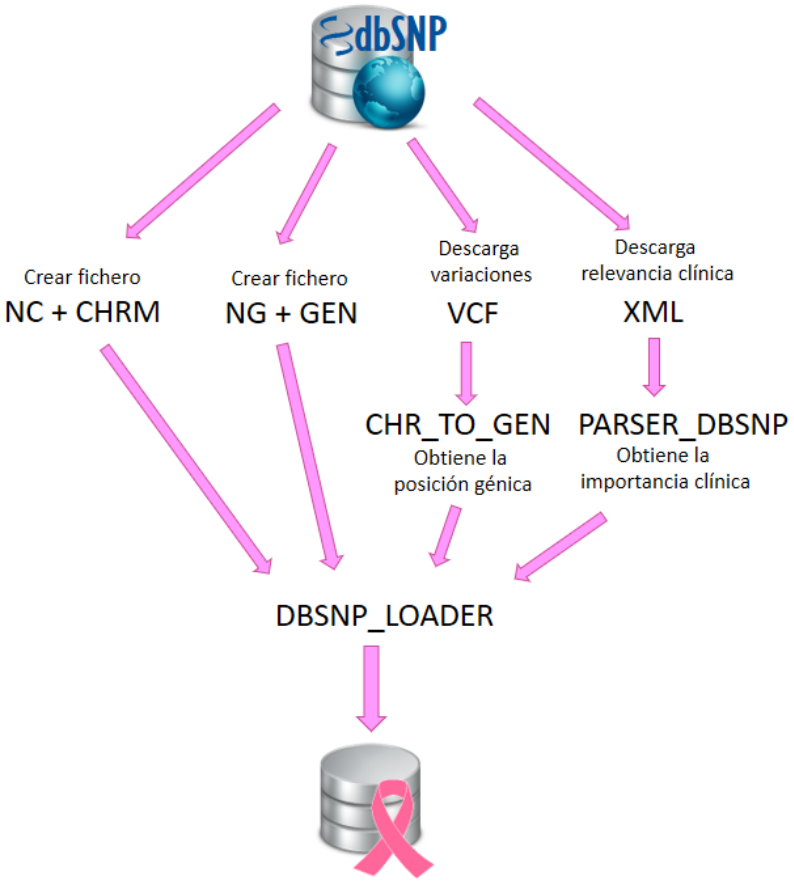

Figura 60. Representación gráfica de la carga de dbSNP

abordar esta tarea se ha desarrollado un script en Python. Este script genera un fichero tabulado con la siguiente información para cada gen:

- Identificador symbol
- Identificador NG
- Gene ID
- Posición de inicio cromosómica
- Posición de inicio génica
- Hebra en la que se encuentra
- Posición final cromosómica
- Cromosoma en el que se encuentra

- **Paso 3: Descargar el fichero VCF del FTP de dbSNP y extraer la información sobre las variaciones de los genes seleccionados en el fichero anterior**. De este fichero VCF con las variaciones nos interesa obtener de cada variación:
  - Identificador rs
  - Cromosoma en el que se encuentra
  - Posición
  - Tipo de variación
  - Numero de nucleótidos borrados
  - Número de nucleótidos insertados
  - GENEINFO (donde indica el símbolo del gen y el identificador del mismo en NCBI, por ejemplo "BRCA2:675")
  - Versión de creación de la variación en dbSNP

  Para llevar a cabo esta tarea debemos de tener en cuenta algunas diferencias en la nomenclatura de este fichero respecto a la de nuestra base de datos.

  La posición que el fichero ofrece para cada variación en la columna 2 del fichero (columna "POS"), no es correcta. Para obtener la posición correcta de la variación debemos obtener el dato del atributo "rsPos" de la columna "INFO".

En cuanto al tipo de variaciones y sus nucleótidos asociados, existen 4 tipos:

- **Indel simple**: Se trata de la sustitución de un nucleótido por otro. En el VCF nos la encontraríamos representada con una línea de este tipo:

CH	POS	ID	REF	ALT	INFO
1	234232	rs201432159	G	A	rsPos:234232

  La significancia de los valores, tal como indica el formato VCF [117, 118] y la web de dbSNP [119] son los siguientes:
  - CH (o CHROM): Indica el cromosoma donde se ha producido la variación.
  - POS: Indica la posición cromosómica del primer nucleótido de la referencia indicada en REF
  - ID: Indica el identificador de dbSNP
  - REF: Contiene el/los nucleótido/s que aparecen en la posición indicada en POS de la secuencia de referencia del cromosoma indicado en CHROM
  - ALT: Contiene la alteración de los nucleótidos producida en la variación con respecto a los nucleótidos de REF
  - INFO rsPos: Contiene la posición cromosómica de la variación reportada en dbSNP.

  En este caso, los nucleótidos y posiciones no necesitan transformación para adaptarse a la base de datos.

- **Inserción:** Se trata de la inserción de un nucleótido en una posición determinada de la secuencia de referencia. En el VCF nos la encontraríamos representada de esta manera:

CH	POS	ID	REF	ALT	INFO
1	10433	rs56289060	A	AC	rsPos:10434

En este caso, para hacer la adaptación debemos quitar el nucleótido de relleno (en este caso una A). Como vemos, en los nucleótidos de referencia hay una A y en la alteración vemos AC. En este caso estamos hablando de una Inserción de una C en la posición 10434.

- **Delección:** Se llama así a las variaciones en las que se elimina el nucleótido de la posición indicada. En el VCF nos la encontraríamos representada así:

CH	POS	ID	REF	ALT	INFO
1	10329	rs150969722	AC	A	rsPos:10330

En este caso, para hacer la adaptación debemos quitar el nucleótido de relleno (en este caso una A). Como podemos ver, en los nucleótidos de referencia tenemos AC y en la alteración una A únicamente. Por lo tanto, estamos ante un borrado de una C en la posición 10330.

- **Indel múltiple**: Se trata de una sustitución de un nucleótido por otro, pero en este caso hay varias opciones de sustitución. En el fichero VCF la vemos representada así:

CH	POS	ID	REF	ALT	INFO
1	91536	rs77418980	G	A,C,T	rsPos:91536

Cuando en la columna ALT los nucleótidos están separados por comas, significa que en esa posición hay más de una variación, es decir, que la variación puede tomar uno de esos valores separados por comas.

- **Delección e Indel**: Este caso es un caso especial en el que, como en el indel múltiple, nos encontramos varias opciones de

sustitución, pero una de ellas es un borrado. La variación en el fichero VCF quedaría representada de la siguiente manera:

CH	POS	ID	REF	ALT	INFO
1	54788	rs59861892	CC	C,CCT	rsPos:54789

Para tratar esta situación, en primer lugar, dividimos las variaciones separadas por comas, tal y como se ha comentado en la variación anterior. Nos quedarían las siguientes variaciones que cargaríamos en nuestra base de datos como dos variaciones totalmente independientes:

- 1 54788 rs59861892 CC C rsPos:54789: En esta variación, dado que en la referencia aparece CC y en la alterada aparece una única C, estamos ante un borrado de una C en la posición 54789.

- 1 54788 rs59861892 CC CCT rsPos:54789: En este caso estamos ante un tipo especial. Debemos tener en cuenta que el nucleótido de relleno será 1 siempre (no más) por lo que, en este caso, si quitamos los nucleótidos sobrantes nos quedaríamos con una variación que en referencia tendría una C y en alterada tendría CT, por lo que en este caso tendríamos una sustitución de C por CT en la posición 54789.

Otro dato que también se extrae del fichero VCF es la versión en la que se creó la variación dentro de dbSNP. Este dato se almacena con el objetivo de que el biólogo al mirar la fecha de creación sea capaz de identificar una mayor validez en la variación. Es decir, si la variación lleva creada desde la versión 80 y sigue vigente y revisada en la nueva versión 127, puede significar que dicha variación es válida.

- **Paso 4: Generar un fichero tabulado a partir de los ficheros descargados de dbSNP para poder asociar cada rs con su importancia clínica.** Este paso se realiza a través de un script en Python. Para ello, se deben descargar los dos ficheros xml de dbSNP que contienen todas las variaciones divididas por cromosomas. La dirección de descarga es la siguiente: ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/XML/. Estos ficheros se descargarán en una carpeta. El script creado se llama *parser_dbsnp.py*. Este programa recibe como entrada un fichero de texto que contiene los nombres de los ficheros XML con la información de las variaciones registradas en dbSNP para cada cromosoma. Cada línea debe contener un solo nombre de fichero. El script analiza cada uno de los ficheros XML y devuelve un fichero tabulado con el identificador rs y la relevancia clínica, solo de aquellas variaciones que tengan relevancia clínica.

- **Paso 5: Definir un método para calcular a partir de las posiciones cromosómicas las posiciones génicas con respecto a un NG.** En este caso se ha optado por el uso de una HashMap [120] en el script *chr_to_gen.java*. Esta HashMap tiene como índice el nombre del gen y como valor le pasaremos un objeto tipo Gene.

- **Paso 6: Definir una estructura para el fichero que permitirá realizar la carga directa en la base de datos.** El fichero de carga realizará las siguientes instrucciones para cada una de las variaciones:
  - Inserción en Variacion: *db_version_id* ("dbSNP" + la última versión), *id_variacion_bd* (rs) y *relevancia_clinica* (leído del TAB).
  - Inserción en Precisa: *tipo* (del VCF), *secuencia_ins* (del VCF), *bases* (del VCF) y *posicion* (del VCF).
  - Relacionar Variacion con el Gen: *ng_identificador* (del fichero).

- **Paso 7: Implementar el paquete de carga y ejecutarla.** Finalmente,

se implementó en Java un script llamado *dbnsp_loader.java* que genera los ficheros LOAD necesarios para cargar en la base de datos de manera directa. Este script tiene como ficheros de entrada los ficheros que hemos creado en los pasos anteriores:

- Fichero de NCs y cromosomas.
- Fichero de NGs y posiciones génicas.
- Fichero de importancia clínica.
- VCF con las variaciones

BIC

En este caso, la base de datos BIC contiene únicamente información asociada a los genes BRCA1 y BRCA2. Para explicar detalladamente la carga realizada de la información de esta base de datos en la base de datos creada en esta tesis y que podemos ver en la Figura 61, dividiremos el trabajo en 4 pasos:

- **Paso 1: Descarga de los ficheros de variaciones de la página web de BIC.** Para acceder a los datos de BIC hay que darse de alta como usuario. Los ficheros están situados en los siguientes links: http://research.nhgri.nih.gov/projects/bic/Member/brca1_mutation_database.shtml y http://research.nhgri.nih.gov/projects/bic/Member/brca2_mutation_database.shtml.
- **Paso 2: Lectura de los ficheros y selección de la información relevante.** Una vez descargados los ficheros debemos extraer los siguientes datos para cada variación de las que se encuentran en los ficheros:
  - Identificador de la variación (*Accesion Number*)
  - Definición HGVS [121] en base al cDNA (*HGVS cDNA*)
  - Importancia clínica (*Clinically Important*)
  - Referencias bibliográficas (*Reference*).

  La limpieza de estos ficheros la realiza un script llamado

bic_parser.py implementado en Phyton, el cual además de quitar los campos innecesarios, elimina las variaciones repetidas y junta todas las referencias bibliográficas.

- **Paso 3: Transformación de los datos para adaptarlos al formato de la Base de Datos del Cáncer de Mama.** La descripción de la variación viene descrita en el campo *HGVS cDNA*. Esta descripción proporciona los siguientes datos:
  - Posición de la variación
  - Tipo
  - Secuencia insertada
  - Secuencia borrada.

  Para poder obtener estos datos a partir de la descripción HGVS [121], se ha implementado un script en Phyton, *HGVS_parser.py*, que permite obtener toda la información por separado pasándole como parámetro la notación entera.

  Sin embargo, hay que tener en cuenta que las variaciones están descritas en base a dos secuencias de referencia pertenecientes a dos cDNAs distintos dependiendo del gen al que pertenezca la variación (BRCA1 y BRCA2). Dado que nuestro modelo conceptual (y por consiguiente nuestra base de datos) se centra en el cromosoma (y no en la secuencia de ADN codificante, como es este caso), necesitamos traducir dichas posiciones de su posición en el ADN codificante a su posición cromosómica. Para realizar la transformación de dichas posiciones, se necesitan conocer los identificadores de dichas secuencias de referencia. Este dato se extrae de manera manual de la web y lo escribimos en un fichero externo para no tener que hacer el cambio en el código HTML.
  - Para BRCA1: http://www.ncbi.nlm.nih.gov/nuccore/555931
  - Para BRCA2: http://www.ncbi.nlm.nih.gov/nuccore/1161383

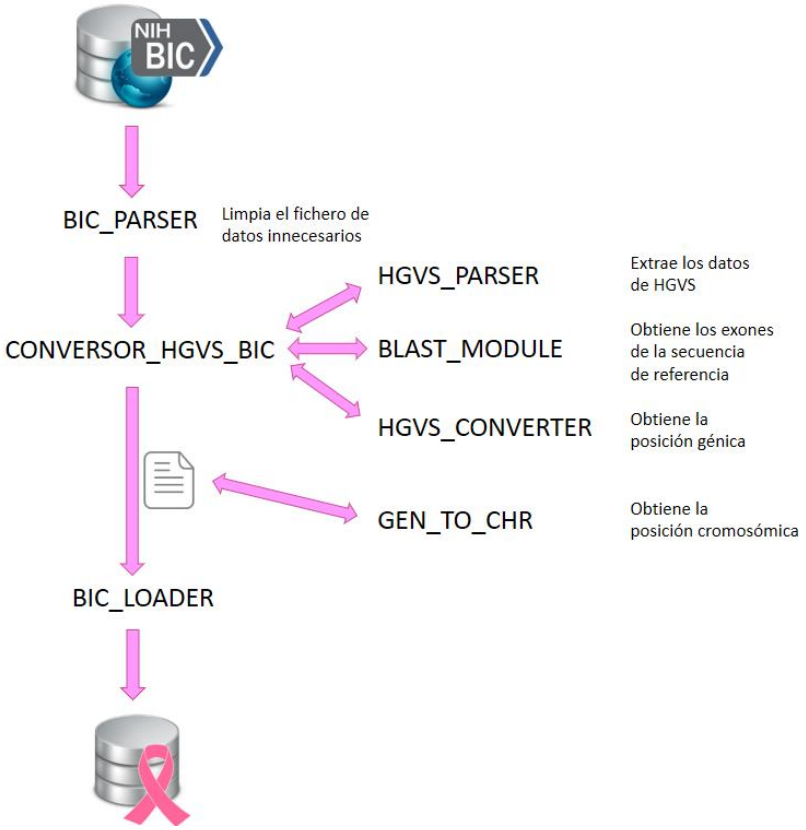

Figura 61. Proceso de extracción, transformación y carga de BIC.

En el caso de BIC, los identificadores de las secuencias no son de GenBank [21], por lo que no podemos acudir a la base de datos a obtener directamente las secuencias de los exones. Para solucionar este problema se implementó un módulo en Phyton llamado *conversor_hgvs_bic.py* que, haciendo uso de BLAST [122, 123], obtiene los exones de la secuencia codificante y los posiciona en la nueva secuencia del gen almacenada en nuestra base de datos. Haciendo uso de dichos exones que hemos obtenido, somos capaces de, utilizando el script *HGVS_converter.py*, transformar la posición codificante a posición génica. Este script se ha implementado en Phyton y se llama *BLAST_module.py*. Los pasos que tiene que seguir este script, *BLAST_module.py*, son los siguientes:

- Llamar a BLAST pasándole como parámetros el identificador de la secuencia de referencia de BIC y el identificador de la secuencia de referencia del gen que hay en nuestra base de datos (BRCA1 o BRCA2). La secuencia de referencia del gen está almacenada en la base de datos en la tabla Gene con el atributo NG_identifier.
- Limpiar el fichero resultante de la siguiente manera:
  - Eliminar la información innecesaria y dejar solo con lo que vamos a trabajar, es decir, los exones.
  - Ordenar exones por porcentaje de acierto.
  - Reconstruir la secuencia con los exones con mayor porcentaje de acierto.
  - Seleccionar los exones que se van a unir.
  - Identificar cuáles de esos exones tienen un *gap* (eliminan algún nucleótido) y en cuál de las dos secuencias.
  - Si tienen *gap*, buscar el exón anterior o posterior (dependiendo de donde esté el *gap*) y comprobar si el *gap* se puede eliminar cogiendo los nucleótidos del otro exón.
  - Eliminar los nucleótidos que se solapen por delante o por detrás.
  - Si aun así sigue habiendo *gaps* en la secuencia sumar o restar el número de *gaps* al exón, dependiendo de la secuencia en la que se eliminen o inserten.
  - Coger para cada uno de los exones de la secuencia reconstruida, su posición de inicio y fin y escribirlo en un fichero con formato: "principio1Exon...fin1Exon, principio2Exon...fin2Exon, principio3Exon...fin3exon, ..." y así hasta el final.

Con este script, tal y como se ha comentado antes, se obtienen los exones y con este resultado, se llama al script *HGVS_converter.py* para obtener la posición génica. El resultado del script Phyton es un fichero de texto tabulado con los siguientes campos (y sus correspondencias con la base de datos):

- AccesionNumber (*id_variacion_db*)
- NGidentifier (ng_identificador)
- StartPos (*posicion*)
- InsertedSeq (*secuencia_ins*)
- NumDeletedBases (*bases*)
- VariationType (*tipo*)
- ClinicallyImportant (*relevancia_clinica*)
- References (*id_pubmed*)

El script *conversor_hgvs_bic.py,* implementado en Phyton, es el módulo encargado de llamar a los scripts *BLAST_module.py, HGVS_parser.py* y *HGVS_converter.py.*

Una vez tenemos la posición génica, se ejecuta el método *gen_to_chr.java*, implementado en Java, que se encarga de transformar las posiciones génicas a cromosómicas.

- **Paso 4: Generar los ficheros de carga de la base de datos y proceder con la carga de la información de BIC.** Con toda esta información, y la posición transformada, se llama a un script generado en java, llamado *bic_loader.java*, que escribirá los ficheros necesarios para ejecutar el LOADER y cargar la información en nuestra base de datos.

Carga de datos sobre microARNs: miRBase.

Para poder trabajar con información sobre microARNs es imprescindible disponer de la información contenida en la base de datos de referencia "miRBase".

Para ello, accedemos a la página de acceso al FPT de descarga de la base de datos http://www.mirbase.org/ftp.shtml donde hay accesibles varios ficheros descargables. El que a nosotros nos interesa es el que contiene el listado de microARNs de la especie homo sapiens. Este fichero está

accesible bajo el nombre de hsa.gff3 (ver Figura 62) en el siguiente link del FTP público de la base de datos

ftp://mirbase.org/pub/mirbase/CURRENT/genomes/hsa.gff3.

Además, desde este FTP tenemos acceso también a versiones anteriores de la base de datos. En nuestro caso, vamos a cargar tres versiones: la 21 (la actual), la 18 y la 17. De estas dos últimas versiones cargaremos únicamente los nombres de los microARNs relacionados con los de la versión actual por un motivo principal: los ficheros de expresión de microARNs procedentes de los microarrays de Affymetrix que utilizaremos más adelante para validar esta tesis hacen referencia a los nombres de estas dos últimas versiones. También descargaremos el fichero *mature.fa*, que contiene las secuencias de referencia de los microARNs maduros.

Estos ficheros vienen en un formato de texto tabulado. Para facilitar el acceso a los datos desde Java, se ha optado por abrir el fichero utilizando la aplicación de Microsoft Excel y guardarlo en formato .xls para tener los datos organizados en columnas y poder acceder a ellos desde Java utilizando una librería JXL para la manipulación de ficheros Excel.

Este fichero tiene una estructura muy sencilla: una cabecera (filas de la 1 a la 13) que contiene información genérica sobre el contenido del fichero como la versión de la base de datos y la secuencia de referencia y el contenido del fichero (filas de la 14 en adelante) donde están anotados los microARNs. Tenemos clasificados dos tipos de microARNs: los precursores

```
##gff-version 3
##date 2014-6-22
#
# Chromosomal coordinates of Homo sapiens microRNAs
# microRNAs:                miRBase v21
# genome-build-id:          GRCh38
# genome-build-accession:   NCBI_Assembly:GCA_000001405.15
#
# Hairpin precursor sequences have type "miRNA_primary_transcript".
# Note, these sequences do not represent the full primary transcript,
# rather a predicted stem-loop portion that includes the precursor
# miRNA. Mature sequences have type "miRNA".
#
chr1    .       miRNA_primary_transcript        17369   17436   .       -       .       ID=MI0022705;Alias=MI0022705;Name=hs
chr1    .       miRNA   17409   17431   .       -       .       ID=MIMAT0027618;Alias=MIMAT0027618;Name=hsa-miR-6859-5p;Deri
chr1    .       miRNA   17369   17391   .       -       .       ID=MIMAT0027619;Alias=MIMAT0027619;Name=hsa-miR-6859-3p;Deri
chr1    .       miRNA_primary_transcript        30366   30503   .       +       .       ID=MI0006363;Alias=MI0006363;Name=hs
chr1    .       miRNA   30438   30458   .       +       .       ID=MIMAT0005890;Alias=MIMAT0005890;Name=hsa-miR-1302;Derives
chr1    .       miRNA_primary_transcript        187891  187958  .       -       .       ID=MI0026420;Alias=MI0026420;Name=hs
chr1    .       miRNA   187931  187953  .       -       .       ID=MIMAT0027618_1;Alias=MIMAT0027618;Name=hsa-miR-6859-5p;De
chr1    .       miRNA   187891  187913  .       -       .       ID=MIMAT0027619_1;Alias=MIMAT0027619;Name=hsa-miR-6859-3p;De
chr1    .       miRNA_primary_transcript        632325  632413  .       -       .       ID=MI0022558;Alias=MI0022558;Name=hs
chr1    .       miRNA   632382  632403  .       -       .       ID=MIMAT0025855;Alias=MIMAT0025855;Name=hsa-miR-6723-5p;Deri
chr1    .       miRNA_primary_transcript        1167104 1167198 .       +       .       ID=MI0000342;Alias=MI0000342;Name=hs
chr1    .       miRNA   1167124 1167145 .       +       .       ID=MIMAT0004571;Alias=MIMAT0004571;Name=hsa-miR-200b-5p;Deri
chr1    .       miRNA   1167160 1167181 .       +       .       ID=MIMAT0000318;Alias=MIMAT0000318;Name=hsa-miR-200b-3p;Deri
chr1    .       miRNA_primary_transcript        1167863 1167952 .       +       .       ID=MI0000737;Alias=MI0000737;Name=hs
chr1    .       miRNA   1167878 1167899 .       +       .       ID=MIMAT0001620;Alias=MIMAT0001620;Name=hsa-miR-200a-5p;Deri
chr1    .       miRNA   1167916 1167937 .       +       .       ID=MIMAT0000682;Alias=MIMAT0000682;Name=hsa-miR-200a-3p;Deri
chr1    .       miRNA_primary_transcript        1169005 1169087 .       +       .       ID=MI0001641;Alias=MI0001641;Name=hs
chr1    .       miRNA   1169055 1169076 .       +       .       ID=MIMAT0001536;Alias=MIMAT0001536;Name=hsa-miR-429;Derives_
chr1    .       miRNA_primary_transcript        1296110 1296170 .       -       .       ID=MI0022571;Alias=MI0022571;Name=hs
chr1    .       miRNA   1296145 1296165 .       -       .       ID=MIMAT0027353;Alias=MIMAT0027353;Name=hsa-miR-6726-5p;Deri
chr1    .       miRNA   1296110 1296129 .       -       .       ID=MIMAT0027354;Alias=MIMAT0027354;Name=hsa-miR-6726-3p;Deri
chr1    .       miRNA_primary_transcript        1312502 1312566 .       -       .       ID=MI0022572;Alias=MI0022572;Name=hs
chr1    .       miRNA   1312539 1312561 .       -       .       ID=MIMAT0027355;Alias=MIMAT0027355;Name=hsa-miR-6727-5p;Deri
chr1    .       miRNA   1312502 1312521 .       -       .       ID=MIMAT0027356;Alias=MIMAT0027356;Name=hsa-miR-6727-3p;Deri
chr1    .       miRNA_primary_transcript        1339650 1339708 .       -       .       ID=MI0022653;Alias=MI0022653;Name=hs
chr1    .       miRNA   1339682 1339703 .       -       .       ID=MIMAT0027516;Alias=MIMAT0027516;Name=hsa-miR-6808-5p;Deri
chr1    .       miRNA   1339650 1339670 .       -       .       ID=MIMAT0027517;Alias=MIMAT0027517;Name=hsa-miR-6808-3p;Deri
```

Figura 62. Detalle del fichero "hsa.gff3" de miRBase con los miRNAs

y los maduros. En nuestro caso, nos vamos a centrar en los microARNs maduros, que son las secuencias finales.

Para llevar a cabo la carga de los datos del fichero hsa.gff3 ejecutaremos el script *CargaMiRBase.java*. El primer paso a realizar es recorrer las filas que contienen microARNs una a una leyendo únicamente los que están marcados con la palabra "miRNA" en la tercera columna, descartando de esta manera los precursores. Una vez seleccionado el microARN que vamos a cargar, nos aseguramos de que no esté ya cargado en el base de datos, para evitar duplicados, y almacenamos el contenido de la fila en las variables. Una vez hecho esto, buscaremos en el fichero de secuencias la que haga referencia al microARN en cuestión y la guardaremos. Con todos estos datos generamos la secuencia INSERT y cargamos el microARN con todos sus datos en la base de datos.

Para los ficheros de las versiones 18 y 17, recorreremos las filas comparando el identificador del microARN, y al encontrar coincidencia almacenaremos el valor del nombre como nombre alternativo.

Carga de datos que relacionan los microARNs con los genes afectados: DIANA microT-CDS

Para llevar a cabo la carga de la base de datos microT-CDS de la colección de datos de "DIANA tools", primero hay que descargar el fichero que contiene la información sobre la relación entre microARNs y los genes alterados por la expresión. Para ello debemos acceder a la página web de "DIANA tools", introducir el usuario y contraseña que podemos obtener de forma gratuita y hacer clic en "Downloads" donde tenemos varios conjuntos de datos disponibles para la descarga. El que nos interesa es el fichero "microT-CDS data" que viene comprimido en un archivo "microT-CDS_data.tar. gz". Seleccionamos este fichero guiados por la opinión de los expertos, que nos comentaron que las bases de datos con dianas predictivas, como microT-CDS contienen más información que las validadas y les interesa poder tener toda la información posible (aunque no esté validada) y validar los resultados a posteriori asegurándose, de esta manera, de que no se dejan información interesante por valorar.

Si descomprimimos el archivo nos encontramos con un fichero tabulado enorme de 2,09GB (ver Figura 63), el cual dividiremos en 677 ficheros tabulados de menor tamaño para poder procesarlos más fácilmente utilizando la aplicación *Text File Cleaver* de *BartDart.com* (http://www.bartdart.com). El número de ficheros en el que se va a dividir el fichero inicial de las variaciones se calcula automáticamente con la herramienta, que te permite indicar el número de filas que quieres que tenga cada uno de los ficheros resultantes.

Como se puede ver en la Figura 63, este fichero está formado por una primera línea, que es la cabecera del archivo e informa del contenido del resto del documento, y el cuerpo del documento, que contiene una relación microARN-gen en cada fila. Cada fila contiene el identificador del tránscrito, el gen, el microARN y el miTG-score de la relación entre el gen y el microARN.

```
TranscriptId,GeneId(name),Mirna-Name(miRBase-version),miTG-score
F52F10.2,F52F10.2(F52F10.2),cel-miR-62(18),0.488
UTR3,V:1550484-1550512,0.00994712053496221
F11A1.3a,F11A1.3(daf-12),cel-miR-62(18),0.702
UTR3,X:10665625-10665653,0.0247451317964074
T11G6.5a,T11G6.5(T11G6.5),cel-miR-62(18),0.511
CDS,IV:10841175-10841203,0.022290972286146
CDS,IV:10840760-10840788,0.0110658552017645
F28H6.4,F28H6.4(F28H6.4),cel-miR-62(18),0.494
UTR3,X:14133330-14133358,0.00425015346910806
CDS,X:14128414-14128442,0.0140299826282741
CDS,X:14128723-14128751,0.00484069798891656
CC8.2a,CC8.2(CC8.2),cel-miR-62(18),0.513
UTR3,IV:3645324-3645352,0.0115627544056086
C09D4.4c,C09D4.4(C09D4.4),cel-miR-62(18),0.471
UTR3,I:5479878-5479906,0.00461240045896611
CDS,I:5484466-5484494,0.0132503417774017
```

Figura 63. Detalle del fichero "microT-CDS_data"

El script de carga *CargaMicroTCDS.java* recorre los 677 documentos fila por fila, recogiendo la información únicamente de aquellos cuyo microARN contiene las letras "hsa", lo cual indica que es de la especie homo sapiens. Antes de insertar la relación en la base de datos, comprueba si existe el microARN para poder asociarlo y, si no existe lo inserta. Realiza el mismo proceso con el gen: comprueba si ya existe en la base de datos, y si no existe lo inserta. Para evitar duplicados, hace la misma comprobación para la relación microARN-gen: comprueba si ya existe y si no existe la inserta en la base de datos.

Carga de datos de rutas metabólicas: KEGG, Reactome y Gene Ontology

La carga de estas tres bases de datos la vamos a realizar con el mismo script. Esto es debido a que los ficheros de datos los descargamos de EnrichR [124], una herramienta online gratuita que permite hacer análisis de genes para conocer con qué rutas metabólicas están relacionados.

Esta herramienta, también tiene disponible de un conjunto de librerías descargables con la información utilizada para realizar los análisis. Debido a la fiabilidad de la herramienta EnrichR, la facilidad de acceso a estos datos (KEGG [43] no permite la descarga de sus datos desde su página web) y a la sencillez con la que los presentan en formato descargable (los ficheros de Reactome [125] y Gene Ontology [54] son mucho más complejos) se decidió utilizar las librerías de esta herramienta para cargar los datos de las

```
positive regulation of insulin secretion (GO:0035774)          GPLD1   GPR68   BAD     SRI     GCG     PLA2G6
retinoid metabolic process (GO:0001523)        RARRES2 APOA2   APOA4   CYP26A1 RBP3    RALDH2  RBP4    APOA1


CDP-diacylglycerol metabolic process (GO:0046341)              CDS2    AGPAT6  CDS1    AGPAT9  TAMM41  AGPAT1
cellular response to heat (GO:0034605)         LYN     ST8SIA1 STAC    HSPA6   SCARA5  MYOF    SLC52A3 TFEC

dendrite development (GO:0016358)              SS18L1  TRAPPC4 MCF2    BDNF    CIT     BAIAP2  BMP7    KLHL1

negative regulation of lymphocyte mediated immunity (GO:0002707)               NDFIP1  CR1     XCL1    SH2D1B
regulation of lymphocyte mediated immunity (GO:0002706)        TGFB1   IL10    LAG3    RIPK3   XCL1    CRTAM

positive regulation of leukocyte mediated immunity (GO:0002705)        TGFB1   LAG3    XCL1    CRTAM   CADM1

negative regulation of leukocyte mediated immunity (GO:0002704)        CCR2    NDFIP1  CR1     CD300A  XCL1
regulation of leukocyte mediated immunity (GO:0002703)         TGFB1   UNC13D  LAG3    CD300A  CRTAM   CADM1

positive regulation of production of molecular mediator of immune response (GO:0002702)        TGFB1   CD74
negative regulation of production of molecular mediator of immune response (GO:0002701)        TGFB1   NDFIP1
regulation of production of molecular mediator of immune response (GO:0002700)         TGFB1   TGFB3   CD74

regulation of steroid hormone secretion (GO:2000831)           CRHR1   TMF1    RUNX1   PTPN11  RETN    GDF9


establishment of organelle localization (GO:0051656)           TOR1A   MREG    SPICE1  MYO7A   SCRIB   MOS
1       PCLO    CLN3    KIF13A  OPHN1   AP3S1   CEP152  NUSAP1  AP3S2   SPDL1
maintenance of location in cell (GO:0051651)           CIZ1    PEX14   SUPT7L  PML     CD4     NIN     TLN2

mammary gland alveolus development (GO:0060749)        DDR1    EGF     PHB2    PRLR    TPH1    TNFRSF11A
regulation of T cell mediated immunity (GO:0002709)            RIPK3   XCL1    RSAD2   HLA-C   LILRB1  MR1
positive regulation of lymphocyte mediated immunity (GO:0002708)                TGFB1   LAG3    XCL1    CRTAM
```

Figura 64. Detalle del fichero "GO_Biological_Process_2015.sdx"

bases de datos KEGG; Gene Ontology y Reactome, seleccionadas por los biólogos del INCLIVA como las más relevantes para el estudio de las relaciones entre genes y *pathways*. Los ficheros descargados son *GO_Biological_Process_2015.sdx*, *KEGG_2016.sdx* y *Reactome_2016.sdx* y se trata de ficheros tabulados donde encontramos en cada fila el nombre de la ruta metabólica, dos tabulaciones y los símbolos de los genes relacionados con la ruta metabólica a continuación separados por tabuladores. En el caso del fichero de GO, además encontramos junto al nombre de la ruta metabólica el identificador de GO correspondiente entre paréntesis, como podemos ver en la Figura 64.

El script *CargaPathways.java* se encarga de recorrer los 3 ficheros descargados de EnrichR y cargarlos en la base de datos. Para ello recorre cada fila del fichero, recogiendo en primer lugar el nombre de la ruta metabólica y el identificador si es el fichero de GO. En el caso de los otros dos ficheros el identificador lo crea utilizando el nombre de la base de datos seguido de dos puntos y el número de la fila donde se encuentra, por ejemplo "KEGG: 132". El siguiente paso es insertar la ruta metabólica (o *pathway*) en la base de datos, con el nombre y el identificador correspondientes.

El proceso de carga continúa leyendo cada uno de los símbolos de gen que se encuentra en la misma fila de la ruta metabólica. Comprueba si ese gen está en la base de datos y obtiene su identificador para poder relacionarlo con la ruta metabólica. Para finalizar, hacemos la inserción de la relación gen-ruta metabólica en la base de datos, siempre comprobando antes que no existe, para evitar duplicados.

3.4. Diseño e implementación de una herramienta software para la gestión de datos clínicos y biológicos del Cáncer de Mama

Siguiendo la necesidad de demostrar las premisas que se plantean en esta tesis, se inició una colaboración con un Grupo de Investigación en Biología Molecular de la Fundación de Investigación INCLIVA. Este grupo trabaja codo con codo con los oncólogos del Hospital Clínico de Valencia en el estudio molecular del Cáncer de Mama. Pudiendo estudiar sus datos clínicos y biológicos se ha llegado a realizar el diseño e implementación de este prototipo de herramienta, que permite gestionar estos datos de una forma cómoda y ágil para los profesionales sanitarios.

La herramienta ha sido diseñada para que los oncólogos puedan introducir los datos de los pacientes durante la realización de la consulta, de forma intuitiva y ágil, permitiendo además que esos datos queden almacenados en la base de datos de forma automática. Estos datos se irán completando con distintos episodios, tratamientos y pruebas a medida que avanza el tiempo y la evolución del paciente y servirán para su utilización posterior en trabajos de investigación realizados a partir de las muestras de dichas pacientes almacenadas en los biobancos.

Debido a la disponibilidad de licencias del laboratorio donde vamos a realizar las pruebas posteriores de la herramienta y el conocimiento que ya tienen, tanto biólogos como médicos, en el manejo de las herramientas de Microsoft Office, se ha decidido realizar la implementación de la parte visual de la herramienta utilizando formularios de Microsoft Access como ya se ha explicado anteriormente. Así mismo, para asegurar la calidad de los datos y la integridad de los mismos, se ha decidido llevar a cabo la implementación de la base de datos haciendo uso del SGBD MySQL. Estas dos tecnologías, junto a un conjunto de scripts implementados en R (necesario para disponer de la funcionalidad requerida a la hora de realizar

análisis estadísticos de la parte biológica) y en Java (para realizar la carga de datos de microARNs) han sido las escogidas para la implementación de esta herramienta que servirá para demostrar las hipótesis en las que se basa esta tesis.

Para facilitar la entrada de datos clínicos y biológicos, hemos implementado un conjunto de formularios utilizando la herramienta de generación de formularios disponible en Access. Estos formularios están implementados en forma de ventanas y están conectados entre sí para permitir una navegación ágil, sencilla e intuitiva.

3.4.1. Ventana de inicio

Nada más iniciar la herramienta nos encontramos con una ventana de inicio a la aplicación muy sencilla, con cinco botones organizados en dos bloques que nos dan acceso a otros formularios. En el primer bloque tenemos los botones de acceso a la funcionalidad relacionada principalmente con los datos clínicos. El segundo de ellos contiene los botones relacionados con los análisis de microARNs. En la Figura 65 podemos ver la apariencia de este formulario.

Figura 65. Formulario de entrada a la aplicación

3.4.2. Lista de pacientes

El botón "Ver lista de pacientes" de la "Ventana de inicio" nos lleva a otra ventana (ver Figura 66) donde podremos visualizar una tabla con todos los pacientes, seleccionar pacientes introducidos previamente, modificar sus datos o incluir nuevos episodios, tratamientos, muestras o pruebas.

Esta lista permite seleccionar un paciente para poder, posteriormente, visualizar o modificar sus datos (haciendo clic en el botón "Editar") o bien generar y descargar un informe con toda la información relativa al paciente seleccionado en formato PDF, que permite la visualización de la historia clínica del paciente en un único documento (haciendo clic en el botón "Generar informe").

3.4.3. Paciente

En esta ventana (que podemos ver en la Figura 67) puedes introducir datos de los pacientes o modificar los que ya se almacenaron previamente.

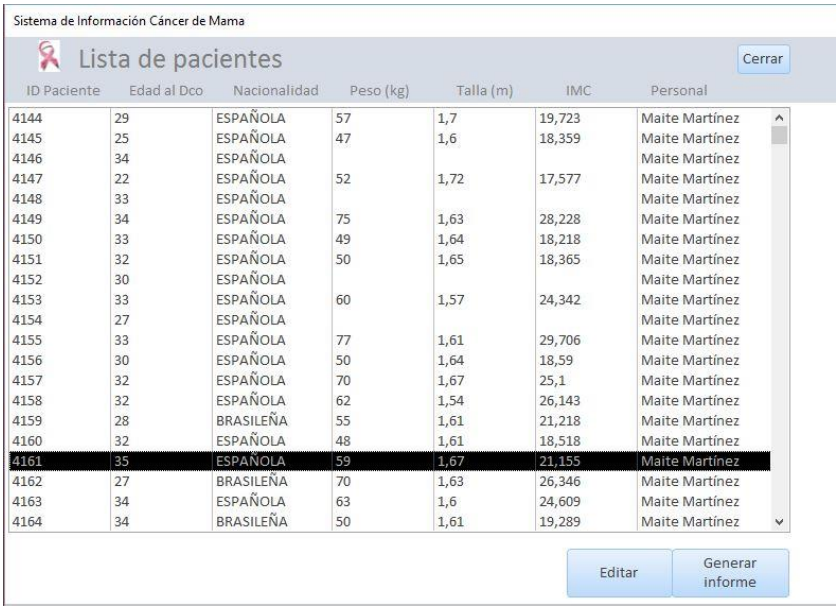

Figura 66. Ventana "Lista de pacientes"

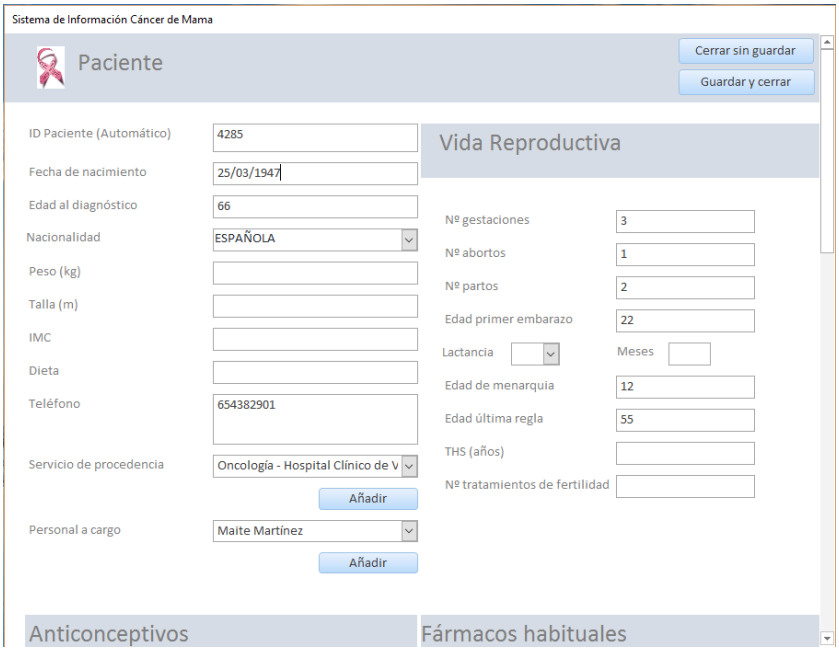

Figura 67. Ventana "Paciente"

En este caso, tenemos acceso a esta ventana por dos vías: El botón "Crear nuevo paciente" desde la "Ventana de inicio", el cual nos dará acceso a esta ventana, pero sin ningún contenido para poder añadir un nuevo paciente a la base de datos, y el botón "Editar" desde la ventana de "Lista de pacientes", gracias al cual podremos visualizar y modificar los datos del paciente seleccionado en la lista.

Este formulario está diseñado para que, desde él, el oncólogo pueda incluir los datos del paciente, entre los que se encuentran sus datos personales, hospital de procedencia y el médico asignado.

Además, encontramos varios apartados dentro del mismo formulario. El primer apartado que nos encontramos a la derecha de los datos personales del paciente es el de "Vida reproductiva". En este apartado podemos incluir información sobre la vida reproductiva del paciente antes del diagnóstico de cáncer de mama.

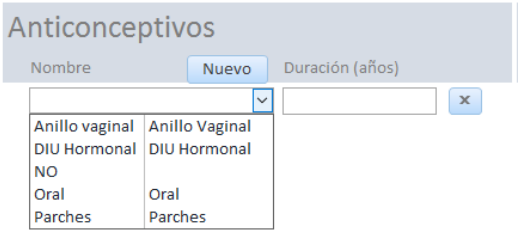

Figura 68. Detalle de la ventana "Paciente" del apartado "Anticonceptivos"

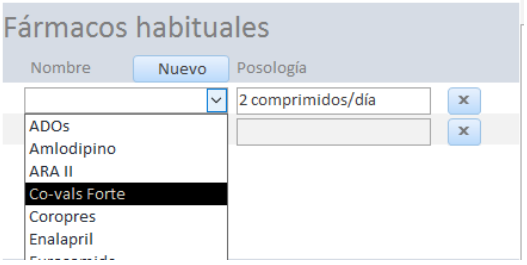

Figura 69. Detalle de la ventana "Paciente" del apartado "Fármacos habituales"

Otro de los aparados es el de "Anticonceptivos", que podemos ver en la Figura 68. En él podemos incluir los anticonceptivos que ha tomado el paciente en su vida (si es que ha tomado alguno) y el tiempo que los estuvo tomando. Además, si el anticonceptivo no se encuentra en la lista, tenemos la opción de añadirlo haciendo clic en el botón "Nuevo" e incluyéndolo en la lista que aparece.

El apartado "Fármacos habituales" de esta ventana (ver Figura 69) contiene información sobre los fármacos que consume el paciente en su vida diaria con asiduidad. El funcionamiento de este apartado es bastante similar al de "Anticonceptivos" con un desplegable con los fármacos y un campo para indicar la posología. Además, también incluye el botón "Nuevo" para poder incluir en la lista aquellos fármacos que no aparezcan.

El apartado "Estados", el cual se muestra en la Figura 70, contiene información referente a los distintos estados de salud por los que pasa el paciente.

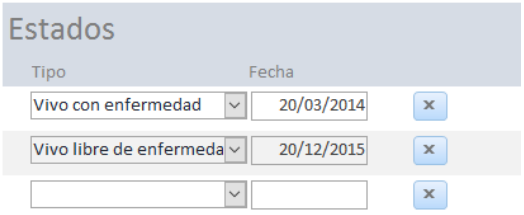

Figura 70. Detalle de la ventana "Paciente" del apartado "Estados"

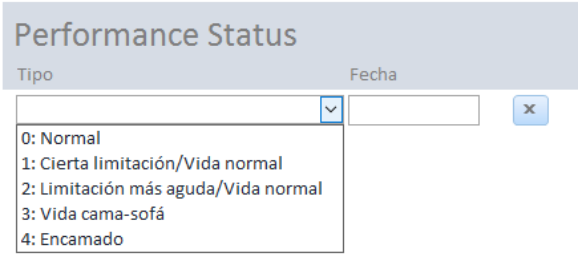

Figura 71. Detalle de la ventana "Paciente" del apartado "Performance status"

De forma similar, como podemos ver en la Figura 71, el apartado "Performance status" alberga información sobre la calidad de vida del paciente.

En el apartado "Antecedentes médicos" encontraremos información del paciente sobre posibles enfermedades o alergias que puede sufrir el paciente, tal y como vemos en la Figura 72 .

"Hábitos tóxicos", como se puede observar en la Figura 73, es otro de los apartados que nos encontramos en la ventana paciente, en este caso con

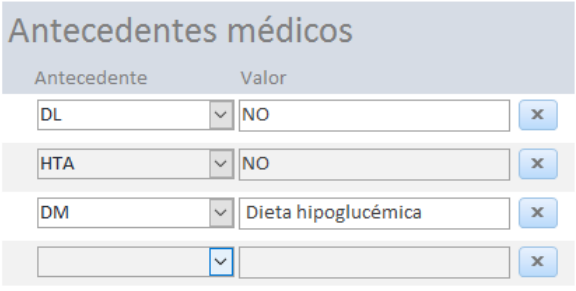

Figura 72. Detalle de la ventana "Paciente" del apartado "Antecedentes médicos"

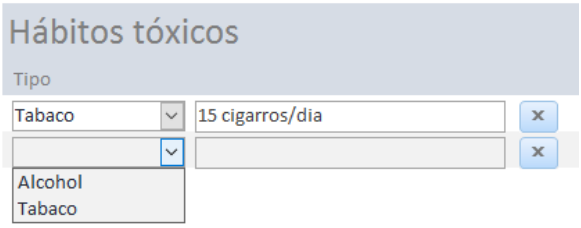

Figura 73. Detalle de la ventana "Paciente" del apartado "Hábitos tóxicos"

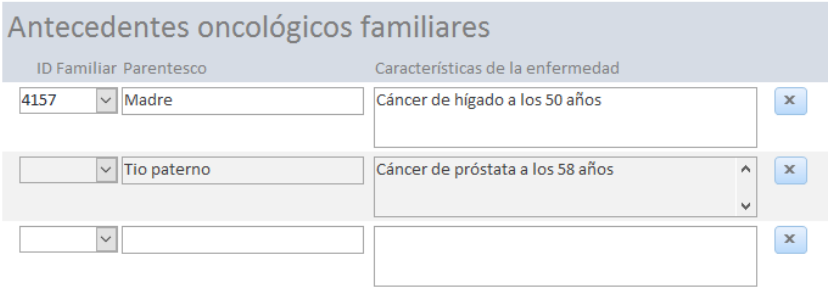

Figura 74. Detalle de la ventana "Paciente" del apartado "Antecedentes oncológicos familiares"

información sobre las adicciones tóxicas que puede tener el paciente, como el alcohol o el tabaco. En este caso no disponemos de un botón "Nuevo" para añadir un campo más a la lista desplegable, pero podemos escribir en el campo "Tipo" y se almacena directamente en lista el valor introducido.

El siguiente apartado a tener en cuenta es el de "Antecedentes oncológicos familiares" en el que, como su nombre indica y podemos verificar en la Figura 74, contiene información sobre los casos de cáncer sufridos en la familia del paciente. Además del parentesco y de las características de la enfermedad sufrida, podemos incluir un número de paciente, que nos permita relacionar pacientes que sean familia entre ellas.

El último apartado de la ventana "Paciente" podemos verlo en la Figura 75, se titula "Episodios" y es un listado donde aparecen todos los episodios de salud por los que pasa el paciente e información relacionada. De esta manera, echando un vistazo rápido podemos observar la cantidad de tratamientos, pruebas, muestras, metástasis o síntomas que ha tenido el paciente durante ese episodio. Además, desde el botón "Nuevo episodio" el

Figura 75. Detalle de la ventana "Paciente" del apartado "Episodios"

usuario puede introducir un nuevo episodio a la historia clínica del paciente accediendo al formulario "Episodio", al que también se puede acceder desde el botón detalles de un episodio ya creado que aparezca en la lista para poder visualizar los detalles del episodio. De la misma manera, si queremos visualizar los tratamientos, pruebas muestras, metástasis o síntomas, simplemente tenemos que hacer clic en el botón "Ver" situado al lado del número en cuestión para que nos abra un listado con dicha información.

3.4.4. Episodio

El formulario "Episodio", como vemos en la Figura 76, nos permite añadir un episodio a la historia clínica del paciente o modificar uno ya creado,

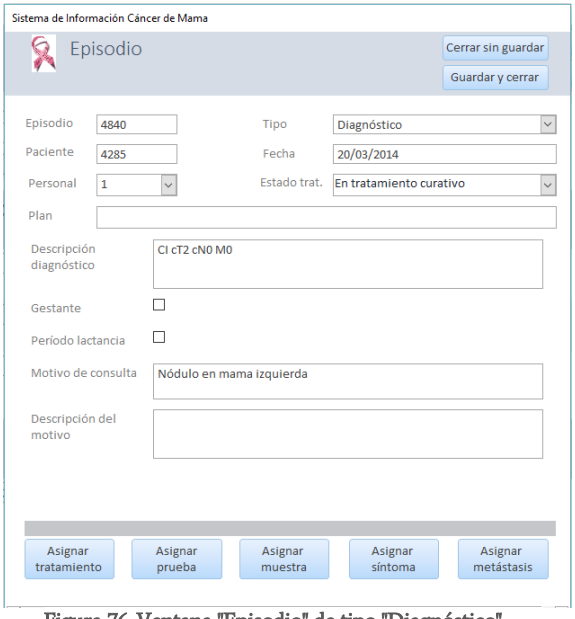

Figura 76. Ventana "Episodio" de tipo "Diagnóstico"

dependiendo del botón seleccionado en la ventana "Paciente", como hemos comentado en el punto anterior.

En esta ventana, podemos seleccionar el tipo de episodio que queremos crear, en función del cual variarán los atributos contenidos en la ventana. Además, desde esta misma ventana podemos proceder a añadir nuevos tratamientos, pruebas, muestras metástasis o síntomas haciendo clic en cada uno de los botones destinados a tal fin.

3.4.5. Lista de tratamientos, pruebas, muestras, metástasis o síntomas

En este caso, tenemos varias ventanas que por tener un funcionamiento similar vamos a describir de forma conjunta: "Lista de tratamientos", "Lista de pruebas", "Lista de muestras", "Lista de metástasis" y "Lista de síntomas".

Poniendo como ejemplo la "Lista de tratamientos" (ver Figura 77), en la ventana nos encontramos un listado que incluye todos los tratamientos incluidos en el episodio seleccionado del paciente que estamos manipulando con un pequeño resumen de la información relacionada.

Además, esta ventana tiene la funcionalidad de poder eliminar tratamientos haciendo clic en el botón marcado con una "X", o visualizar o modificar los datos relacionados con ese tratamiento, haciendo clic en el botón “Detalles” del tratamiento que deseamos visualizar. De esta manera accedemos al formulario "Tratamiento" donde podremos manipular los datos del

Figura 77. Ventana "Lista de tratamientos"

tratamiento. Para acceder al formulario "Tratamiento", también podemos hacer clic en el botón "Añadir nuevo tratamiento" que nos lleva a dicho formulario vacío de información para poder asignar un nuevo tratamiento al episodio de un determinado paciente.

3.4.6. Tratamiento

Este formulario, como hemos ido comentando en apartados anteriores, tiene tres puntos de acceso: el botón "Asignar tratamiento" del formulario "Episodio", y los botones "Detalles" y "Añadir tratamiento" del formulario "Lista de tratamientos". Su funcionalidad es la de insertar, modificar o visualizar datos sobre un tratamiento que ha sido aplicado al paciente.

Además, dependiendo del tipo de tratamiento que queramos insertar aparecerán unos campos de información u otros referentes a ese tipo de

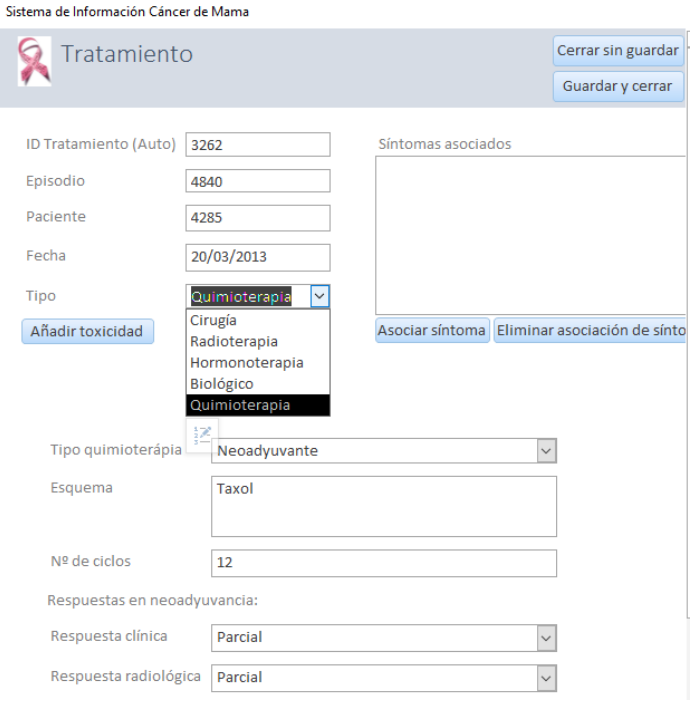

Figura 78. Ventana "Tratamiento"

tratamiento. Por ejemplo, en el caso de la Figura 78 hemos seleccionado como "Tipo" = "Quimioterapia" y a continuación nos han aparecido algunos parámetros como "Tipo quimioterapia" donde nosotros hemos seleccionado "Neoadyuvante" e instantáneamente nos han aparecido campos como los tipos de respuesta.

También cabe resaltar que desde esta ventana es posible añadir una toxicidad al tratamiento desde el botón "Añadir toxicidad" o asociar un síntoma desde el botón "Asociar síntoma", el cual nos aparecerá en la lista de encima del botón junto al resto de síntomas asociados si los hubiese.

Finalmente, en el caso en el que el tipo de tratamiento sea "Cirugía" es posible asociar el tratamiento con la muestra extraída durante la cirugía, bien seleccionando el identificador de la muestra en el menú desplegable, o bien añadiendo una muestra nueva desde el botón "Añadir muestra".

3.4.7. Prueba

Este formulario sirve para visualizar, modificar o insertar una nueva prueba del paciente en un episodio. Podemos acceder a él por dos vías: El botón "Añadir prueba" de la ventana "Lista de pruebas" o el botón "Asignar prueba" de la ventana "Episodio".

Tal y como ocurre en el formulario "Tratamiento", este formulario es un formulario dinámico, es decir, sus campos varían dependiendo del tipo de prueba seleccionada. En el caso de la Figura 79, nos encontramos con una muestra de un formulario de una prueba de tipo "Biológicas" y de subtipo "Inmunohistoquímica". En este caso, en el del resto de pruebas de tipo "Biológicas", las de tipo "Laboratorio" y en el caso de las pruebas de tipo "Radiología" y subtipo "BAG/PAAF", además de los parámetros particulares de la prueba, tenemos un campo desplegable donde podemos escoger la muestra a la que hace referencia dicha prueba. Igual que ocurría en el formulario de "Cirugía". En el caso en el que la muestra no se encuentre en el desplegable, podemos hacer clic en el botón "Añadir muestra" donde

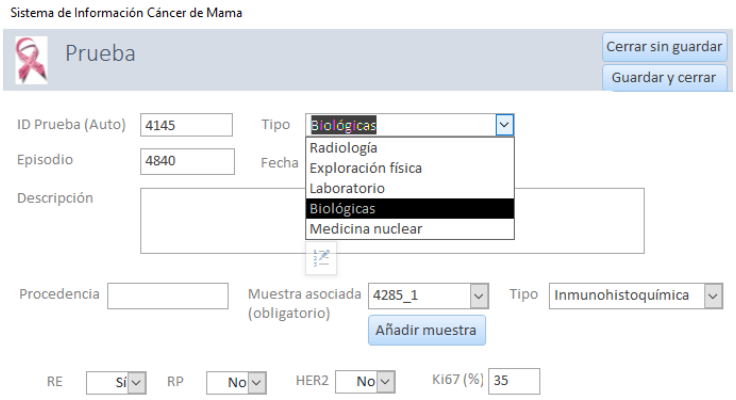

Figura 79. Formulario "Prueba" de tipo "Biológicas" y subtipo "Inmunohistoquímica"

accederemos directamente al formulario "Muestra" y podremos incluir la muestra a la que hace referencia la prueba.

3.4.8. Muestra

La funcionalidad de este formulario es la de visualizar, modificar o añadir muestras a un episodio de un paciente. Como hemos ido comentando a lo largo del capítulo, se puede acceder a él desde sendos botones distribuidos en varias ventanas: la ventana "Episodio", la ventana "Prueba" (en alguno de sus tipos), la ventana "Tratamiento" de tipo "Cirugía", y la ventana "Lista de muestras".

Como algunas de las ventanas explicadas anteriormente, esta ventana que vemos en la Figura 80 es de contenido dinámico, variando sus parámetros dependiendo del tipo seleccionado. Seleccionando el tipo "Tumor", aparecen un conjunto de campos relacionados con las características del tumor, como podemos ver en la Figura 80. Además, es posible asociar una muestra de tipo "Ganglios" a este tumor principal, seleccionando el identificador de la muestra de ganglios en el campo desplegable, como podemos ver en la Figura 81, o bien incluyendo una nueva muestra de tipo "Ganglios" indicando la cantidad de ganglios extraídos y afectos por este

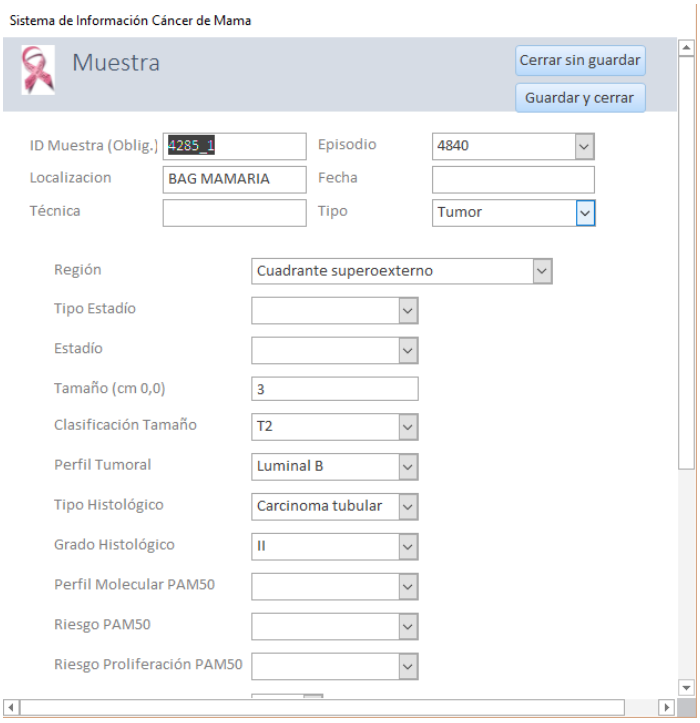

Figura 80. Ventana "Muestra" de tipo "Tumor"

Figura 81. Ventana "Muestra" de tipo "Ganglios"

primer tumor. Acceder a la ventana de "Muestra" de tipo "Ganglios" es posible desde el botón "Añadir Ganglios" de esta ventana, o simplemente cambiando el tipo de muestra en el desplegable del inicio de la página.

Figura 82. Detalle de la ventana "Muestra" de tipo "Tumor"

También es posible asociar una metástasis a un tumor, mediante el campo desplegable correspondiente o utilizando el botón "Añadir metástasis" (en el caso en el que el identificador de la metástasis que queremos asociar no se encuentre en el listado), ambos disponibles en la ventana "Muestra" de tipo "Tumor", como aparece en el detalle de la Figura 82.

Finalmente, otra de las opciones disponibles para las muestras de tipo "Tumor" es la posibilidad de relacionar dicha muestra con otra muestra tumoral introducida previamente en la base de datos. Para ello, necesitaríamos simplemente hacer clic en el botón "Añadir muestra antigua" y seleccionar la muestra antigua en la lista desplegable que aparece en el formulario para tal fin.

3.4.9. Metástasis

El formulario "Metástasis" tiene la funcionalidad de visualizar, modificar o añadir metástasis en el episodio de un paciente. Tiene acceso desde la

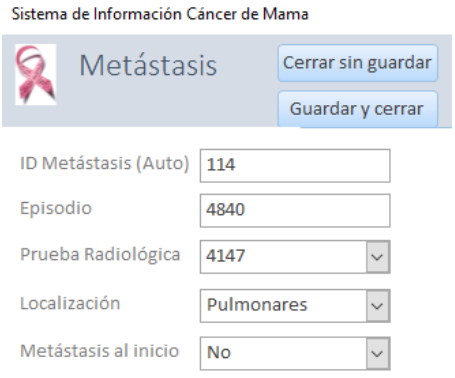

Figura 83. Ventana "Metástasis"

ventana "Episodio", la ventana "Lista de metástasis" y la ventana "Muestra" de tipo "Tumor", como hemos ido comentando a lo largo de este capítulo.

En esta ventana que se muestra en la Figura 83, podemos encontrar información sobre la metástasis del paciente, como su localización o si la tuvo desde el inicio. Además, podemos relacionarla con la prueba radiológica por la que se detectó (si la tiene) utilizando para ello el desplegable diseñado con ese propósito.

3.4.10. Síntoma

Esta ventana que vemos en la Figura 84 nos permite incluir aquellos síntomas que puede padecer el paciente durante cualquier episodio de su enfermedad. Puede estar asociado a un tratamiento, por lo que podemos seleccionarlo desde el desplegable diseñado para seleccionar el tratamiento responsable.

Además, el acceso a esta ventana puede realizarse desde el formulario "Tratamiento" y su botón "Asociar síntoma", desde la ventana "Episodio" o desde la ventana "Lista de síntomas", como se ha comentado anteriormente.

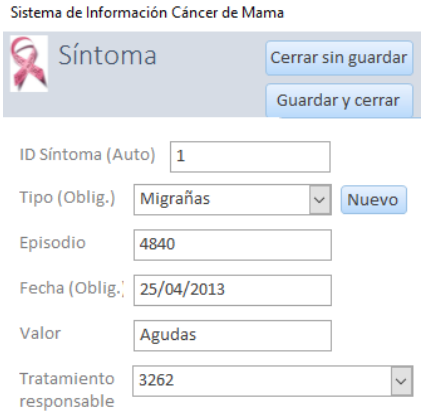

Figura 84. Ventana "Síntoma"

3.4.11. Filtrado de pacientes

Este formulario está diseñado para realizar consultas multiparámetro a la base de datos. Seleccionando o definiendo los parámetros en la parte superior de la pantalla que aparece en la Figura 85, gracias al formulario de consulta, y haciendo clic en el botón "Aplicar filtro" se ejecuta una consulta a la base de datos con los parámetros establecidos por el usuario. Esta consulta devuelve un listado de pacientes que cumplen los parámetros designados y el número de pacientes de ese listado.

Los parámetros seleccionados para formar parte de este formulario fueron sugeridos por oncólogos especialistas del INCLIVA, que consideraron que eran los más relevantes para las investigaciones realizadas hoy en día en su grupo de investigación.

Además, es posible realizar una exportación de los datos de la tabla de resultados a un fichero Excel para que el usuario pueda disponer de esos datos fuera de la herramienta.

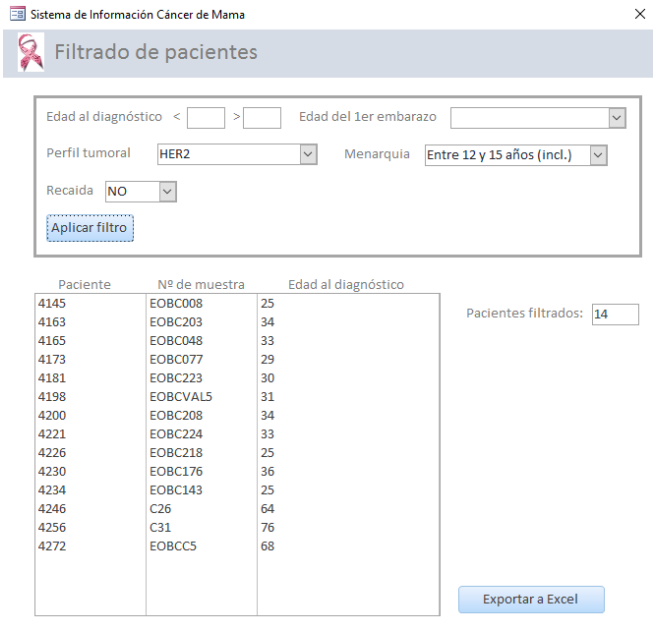

Figura 85. Ventana "Filtrado de pacientes"

3.4.12. Carga de datos de microARNs

Esta ventana que podemos ver en la Figura 86, se trata de la primera relacionada la vista de Expresión Génica. En este caso, siguiendo las recomendaciones de los biólogos expertos, hemos tomado como ejemplo el análisis de expresión de microARNs, porque estas micromoléculas son de las más estables en cualquier tipo de tejido a nivel biológico, por su función moduladora y su análisis es bastante sencillo por la cantidad limitada de microARNs a analizar.

Para acceder a esta ventana hay que hacer clic en el primer botón del segundo bloque de botones de la "Ventana de Inicio", el botón "Cargar fichero miRNAs". [3]

La funcionalidad de esta ventana consiste en permitir al usuario seleccionar el fichero que contiene los valores de expresión de microARNs obtenidos mediante la técnica del microarray de Affymetrix [126] para poder cargarlos en la base de datos. Para ello existe en la ventana un botón "Examinar" que permite seleccionar la ruta del fichero con las medidas de las expresiones.

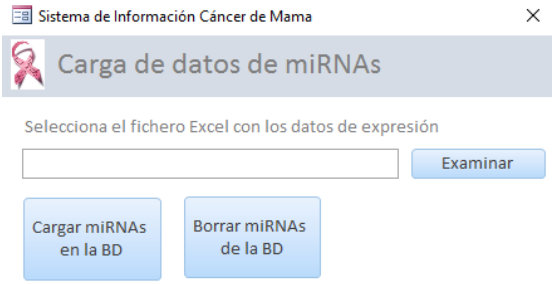

Figura 86. Ventana "Carga de datos de miRNAs"

[3] En la herramienta se ha utilizado la abreviatura “miRNA” (nomenclatura inglesa) en lugar de “microARN” por ser la abreviatura utilizada comúnmente por los biólogos en el laboratorio para referirse a estas moléculas.

El siguiente paso será la ejecución del script de carga llamado *Carga.java* haciendo clic en el botón "Cargar miRNAs en la BD". Para ello, desde la ventana de Access ejecutaremos un fichero *ejecuta_Carga.bat* que se encargará de ejecutar las librerías necesarias para que el script funcione correctamente y de ejecutar el script pasándole como argumento la ruta de archivo seleccionada.

Finalmente, también tenemos disponible un botón llamado "Borrar miRNAs de la base de datos" por si consideramos conveniente limpiar la base de datos de los valores de expresión introducidos previamente.

Funcionamiento del script Carga.java

Este script de carga ha sido desarrollado en Java y se encarga de leer el fichero Excel de la ruta seleccionada para cargar cada uno de los valores de expresión contenidos en el fichero.

Para poder realizar la carga, primero debemos establecer las correspondencias entre el fichero Excel y la base de datos. Estas correspondencias han quedado reflejadas en la Tabla 3 y la Tabla 4, las cuales especifican, además, la estructura que debe tener el fichero para una correcta lectura.

En primer lugar, selecciona entre todos los microARNs analizados por el microarray aquellos que nos interesan, descartando los que no sean humanos y los que sean precursores (nos quedaremos únicamente con los microARNs maduros). Una vez seleccionados, comprueba que estén en la base de datos, insertándolos en el caso contrario (no hay ningún caso de este tipo ya que se han cargado los identificadores de varias versiones de miRBase para evitar este error).

CORRESPONDENCIAS PESTAÑA "EXTRACCIÓN RNA"			
COLUMNA	TABLA	ATRIBUTO	DATOS PREDEFINIDOS
id	muestra	id_muestra	
ng/ul	material_genetico	concentracion	

260/280	material_genetico	ratio260-280	
V	material_genetico	volumen_final	
NHC	paciente	NHC	
fecha	muestra	fecha_muestra	Dato en formato fecha. Si encontramos únicamente el año cargaremos como día y mes el 1 de enero.
edad	paciente	edad_al_dco	
muestra	prueba	fecha_prueba	Dato en formato fecha.

Tabla 3. Correspondencias del fichero de datos de expresión de microARNs, en la pestaña "Extracción RNA" con las tablas de la base de datos.

CORRESPONDENCIAS PESTAÑA "DATOS_CHIP_ENTERO_NORMALIZADO"			
COLUMNA	TABLA	ATRIBUTO	DATOS PREDEFINIDOS
probeset_id	arn	id_arn (cada una de las filas es un id_arn)	
id_muestra	muestra medidor_ expresion	id_muestra (Las cabeceras de las siguientes columnas corresponden a los id_muestra) expresion (Cada una de las filas corresponde a la expresión del microARN de esa fila en la muestra de la columna correspondiente)	

Tabla 4. Correspondencias del fichero de datos de expresión de microARNs, en la pestaña "datos_chip_entero_normalizado" con las tablas de la base de datos.

A continuación, relaciona los pacientes de la base de datos con las muestras del fichero, buscando correspondencias entre los identificadores del fichero y los identificadores de las muestras que tenemos precargados en la base de datos. Si no encuentra el identificador de la muestra en la base de datos es porque el paciente no estaba precargado, por lo que generaremos un nuevo paciente, un nuevo episodio asociado a ese paciente y una nueva muestra de tipo tumor con ese identificador de muestra.

El siguiente paso es insertar una nueva *Prueba Biológica* de tipo *Test de Expresión* para cada muestra del fichero Excel y asociarle un nuevo *Material Genético Diluido*.

A partir de este momento, leeremos fila por fila la columna correspondiente a la muestra y cargaremos todas las medidas de expresión de microARNs relacionadas con dicho *Material Genético Diluido* y con el correspondiente *ARN* precargado en la base de datos. Una vez haya finalizado este proceso con cada una de las columnas del Excel habrá finalizado la carga de datos.

3.4.13. Análisis de expresión de microARNs

La ventana que aparece en la Figura 87, accesible únicamente a través del botón "Analizar miRNAs" de la "Ventana de inicio", permite al usuario la realización de análisis de los datos de expresión disponibles en la base de datos a partir de una serie de parámetros configurables, como son el umbral a partir del cual se considera que el análisis es fiable y los rangos de edad mayor y menos para establecer los dos grupos de pacientes a comparar. También ofrece un listado con los análisis realizados previamente y la consulta de los resultados de esos análisis de tres formas distintas: por microARNs, por genes y por *pathways.*

Para llevar a cabo un análisis de los datos de expresión, el primer paso será completar los campos que aparecen en la sección "Realizar nuevo análisis" introduciendo un valor umbral, así como la edad máxima del grupo de pacientes jóvenes y la edad mínima del grupo de pacientes mayores para establecer los dos grupos que queremos comparar. Una vez hecho esto,

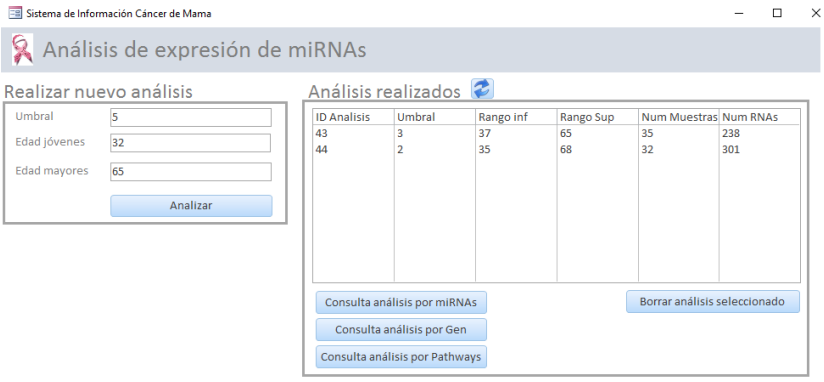

Figura 87. Ventana de "Análisis de expresión de miRNAs"

hacemos clic en el botón "Analizar" para ejecutar el script *analisisPorEdades.R* que realizará el análisis utilizando los paquetes estadísticos de las librerías de R.

Una vez finalizada la ejecución del script *analisisPorEdades.R* el análisis aparecerá en la lista de "Análisis realizados", a la derecha de la pantalla, mostrando los parámetros definidos por el usuario al realizar el análisis y los microARNs que pasaron el filtro del umbral y cuya expresión se utilizó para la realización de los cálculos estadísticos.

Seleccionando uno de los análisis de la lista "Análisis realizados" podemos llevar a cabo varias acciones: borrarlo presionando el botón "Borrar análisis seleccionado" o consultar los resultados del análisis.

La consulta de los resultados del análisis se puede realizar de tres maneras distintas: visualizando los microARNs sobreexpresados, visualizando los genes afectados, o visualizando los *pathways* alterados. Cada una de ellas tiene su botón correspondiente que re redirige a la ventana diseñada para cada tipo de consulta.

Funcionamiento del script analisisPorEdades.R

Este script se encarga comparar la expresión de los microARNs en dos grupos de mujeres agrupadas por la edad.

En primer lugar, carga las librerías RMySQL (para poder establecer la conexión con la base de datos) y gtools (librería de R con algunas funciones de manipulación de datos).

Seguidamente, obtiene los datos de expresión de la base de datos y selecciona únicamente aquellos microARNs en los cuales haya más de tres valores de expresión que superen el umbral definido por el usuario. Esta selección se realiza para evitar seleccionar microARNs que hayan podido tener valores de expresión mayores que el umbral en una o dos muestras, lo que en este caso se consideraría un error de medida.

El siguiente paso es generar una matriz que contenga todos los valores de expresión de los microARNs seleccionados en el paso anterior en las muestras de las mujeres cuya edad sea igual o menor a la edad establecida por el usuario como parámetro del análisis.

Continuaremos realizando el mismo proceso del paso anterior, pero en este caso con los datos de las mujeres mayores, introduciendo en la matriz los valores de expresión de los microARNs seleccionados en el paso inicial en las muestras de las mujeres cuya edad sea igual o mayor a la edad establecida por el usuario como parámetro del análisis.

Con todos los datos seleccionados seguiremos realizando los análisis estadísticos necesarios para poder comparar la expresión de esos microARNs en ambos grupos. Para ello utilizaremos el análisis estadístico T-test, del cual obtendremos los p-valor ajustados (utilizando el método *Benjamini & Hochberg*) para cada microARN y realizaremos un *Fold_Change* para cada microARN con las medianas de las muestras de jóvenes y la mediana de las muestras de las mayores para valorar cuan diferentes son los valores de expresión entre los dos grupos de pacientes.

Para acabar, el script inserta en la base de datos los valores obtenidos tras la realización de todos los cálculos estadísticos.

3.4.14. Consulta de análisis por microARNs

Esta ventana que aparece en la Figura 88, accesible desde el botón "Consulta análisis por miRNAs" en la ventana "Análisis de expresión de miRNAs", tiene la funcionalidad de mostrar los resultados del análisis seleccionado en la ventana "Análisis de expresión de miRNAs" ordenados según el *p-valor* de cada uno de los microARNs analizados. En ella, tenemos varias secciones con distinta información en cada una de ellas.

En la sección "Detalles del análisis", como su nombre indica, podemos visualizar los parámetros seleccionados por el usuario para la realización del análisis, así como el número de ARNs y de muestras incluidos en el análisis, cuyo valor depende de los parámetros anteriores.

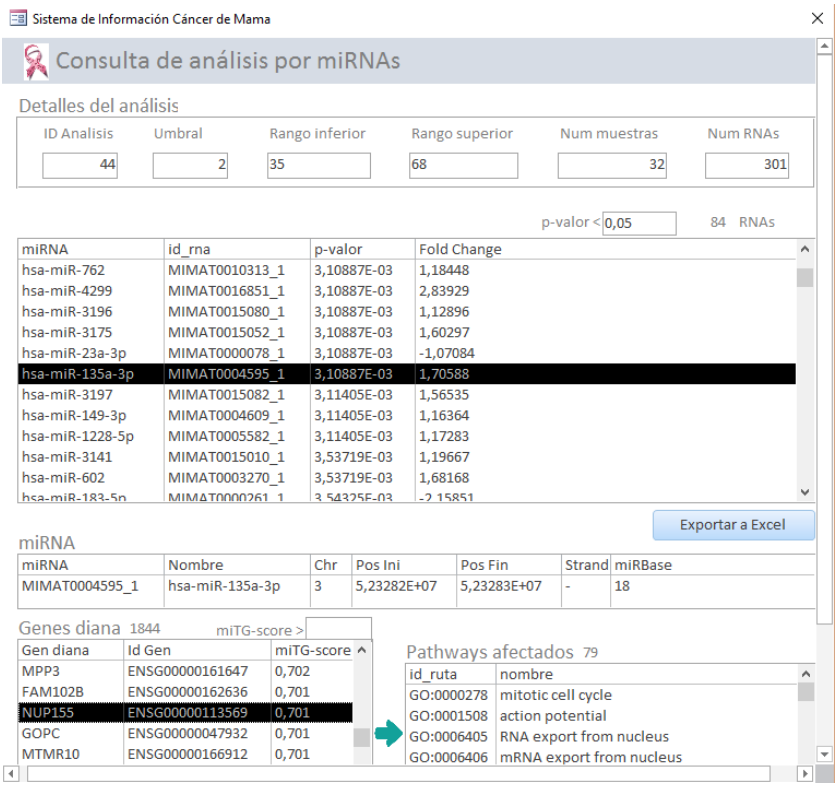

Figura 88. Ventana "Consulta de análisis por miRNAs"

En la siguiente sección nos encontramos una tabla con los microARNs ordenados por *p-valor* de menor a mayor y el *Fold Change* calculado de cada uno de ellos, pudiendo visualizar en ella los microARNs más significativos arriba. Además, es posible filtrar los microARNs que aparecen por el valor de su *p-valor* gracias al filtro que aparece en la tabla y exportar todos los resultados de la tabla a un fichero Excel utilizando el botón "Exportar a Excel".

Si seleccionamos uno de los microARNs de la tabla de resultados, se actualiza la información de la tabla "miRNA". En ella aparece información propia del microARN, procedente de la base de datos miRBase.

También al seleccionar uno de los microARNs de la lista de resultados se actualiza la tabla "Genes diana". En ella te muestra los genes afectados por ese microARN y el *miTG-score* asociado a cada uno de ellos, por el cual también se puede filtrar.

Finalmente, seleccionando uno de estos genes diana se actualiza la lista de *pathways*, donde se muestran los *pathways* alterados cuando el gen seleccionado está afectado por una sobreexpresión de un microARN.

3.4.15. Consulta de análisis por genes afectos

Esta ventana es accesible desde el botón "Consulta análisis por gen" de la ventana "Análisis de expresión de miRNAs" y tiene la funcionalidad de mostrar el listado de genes que han quedado afectados por la sobreexpresión de los microARNs analizados, tal como vemos en la Figura 89.

La sección "Detalles del análisis" de esta ventana es igual que la de la ventana de "Consulta de análisis por miRNAs".

Sin embargo, la característica principal de esta ventana es que, al seleccionar uno de los genes afectos, actualiza el contenido de la tabla principal de la ventana, incluyendo información sobre los microARNs que interfieren en las funciones de ese gen y *el p-valor* obtenido en el análisis

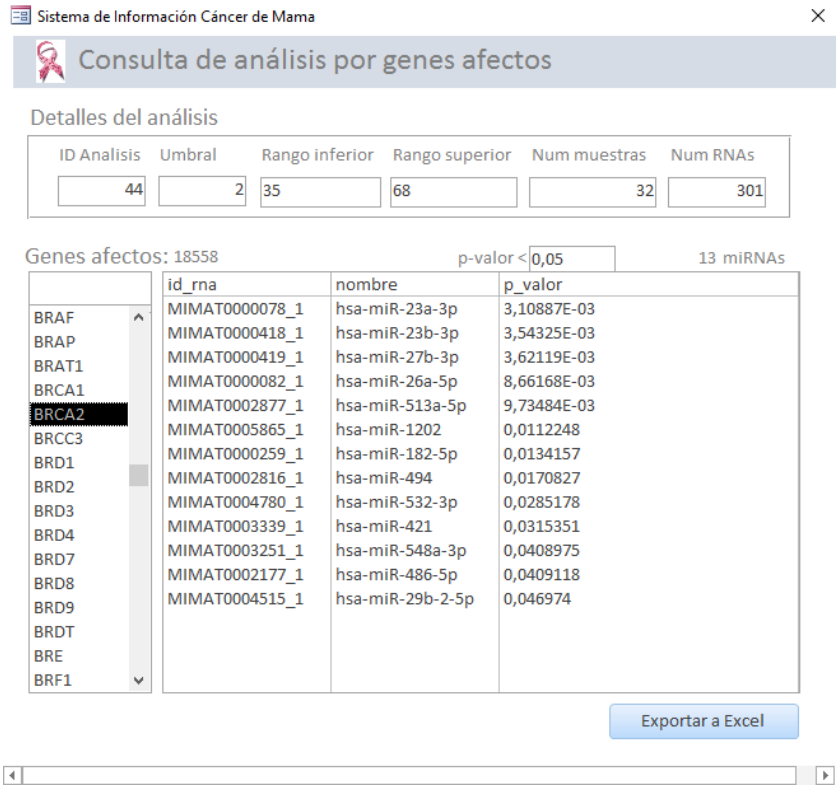

Figura 89. Ventana "Consulta de análisis por genes afectos"

que hemos realizado, indicando la significatividad de ese microARN en el análisis. Como en la tabla de la otra ventana, también podemos filtrar los resultados por *p-valor* menor que el introducido por el usuario y tenemos la funcionalidad de exportar la tabla a un fichero Excel desde el botón "Exportar a Excel" situado bajo la tabla.

3.4.16. Consulta de análisis por *pathways* alterados.

La funcionalidad de esta ventana es visualizar los *pathways* que han sido alterados por la sobreexpresión de los microARNs analizados. Para acceder a la ventana que vemos en la Figura 90 debemos hacer clic sobre el botón

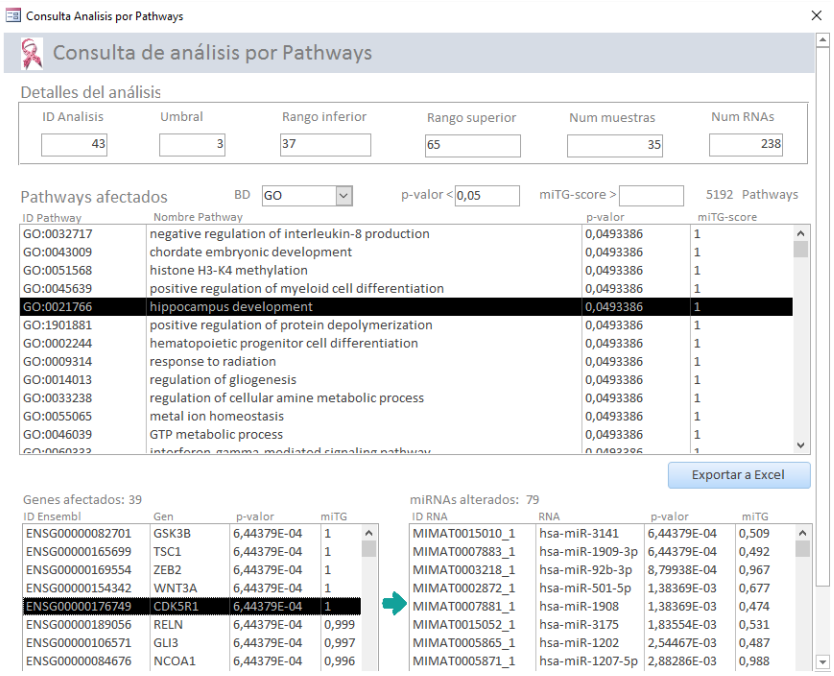

Figura 90. Ventana "Consulta de análisis por Pathways"

"Consulta análisis por pathway" en la ventana "Análisis de expresión de miRNAs".

La sección "Detalles del análisis" de esta ventana es igual que la de la ventana de "Consulta de análisis por miRNAs".

La siguiente sección de la ventana corresponde al listado de *pathways* afectados por la sobreexpresión de los microARNs. Este listado aparece ordenado de menor a mayor *p-valor*, siendo el *p-valor* que aparece en cada fila el *p-valor* más alto de todos los microARNs relacionados con el *pathway*, y de mayor a menor *miTG-score*. Además, en esta tabla se pueden aplicar filtros para mostrar únicamente los *pathways* que nos interesan. Existen filtros por *p-valor* menor que el valor introducido, por *miTG-score* mayor que el introducido y por base de datos origen de la información. Actualmente hay información sobre la relación sobre *pathway*-gen procedente de tres bases de datos distintas: Gene Ontology (GO) [54], Reactome [125] y KEGG [43]. Gracias al carácter "holístico" del diseño de

este sistema de información, se pueden ofrecer resultados de estas bases de datos integradas, pudiéndose ampliar el catálogo de bases de datos en un futuro de la misma forma que se ha hecho con las bases de datos actuales. Los resultados mostrados en esta tabla se pueden exportar a un fichero Excel haciendo clic en el botón "Exportar a Excel"

La tabla "Genes afectados" se actualiza cuando seleccionamos uno de los *pathways* de la lista "Pathways afectados". En ella incluye los genes afectados por la sobreexpresión de microARNs en ese análisis que están relacionados con el *pathway* seleccionado. Esta tabla se encuentra ordenada por los mismos parámetros que la tabla "Pathways afectados".

Finalmente, cuando se selecciona un gen de la tabla "Genes alterados" se actualiza la información de la tabla "miRNAs alterados". En ella aparece información sobre los microARNs que aparecen sobreexpresados en el análisis, en mayor o menos medida, ordenados por *p-valor* y relacionados con el gen seleccionado en la tabla anterior.

4. Validación: Implantación de la solución en el Proyecto Cáncer de Mama en Mujeres Jóvenes

Con el fin de probar el funcionamiento de la tecnología diseñada y desarrollada en esta tesis, se han llevado a cabo una serie de pruebas dentro del marco de un proyecto nacional coordinado junto a un Grupo de Investigación perteneciente a la Fundación de Investigación INCLIVA y titulado "Análisis de la desregulación transcriptómica en tejido tumoral de mujeres jóvenes con cáncer de mama e implicaciones funcionales ".

4.1. Objetivo

El objetivo principal de estas pruebas de validación es verificar que la utilización de sistemas de información basados en modelos conceptuales en entornos clínicos y de investigación médica, como el del diagnóstico, tratamiento e investigación en el cáncer de mama en mujeres jóvenes, mejora la gestión y análisis de los datos para llevar a cabo los procedimientos de actuación habituales de una forma mucho más sencilla, ágil y eficiente que con los métodos tradicionales.

El experimento llevado a cabo ha seguido la metodología de *Technical Action Research (TAR)* propuesta por Roel Wieringa [11, 12]. Según Wieringa, TAR está relacionado con el uso de un artefacto experimental para ayudar a un cliente y aprender sobre sus efectos en la práctica. En un proceso de validación con TAR, el investigador usa un artefacto (en nuestro caso, la herramienta descrita en el capítulo 3.4) en un proyecto del mundo real para ayudar al cliente, o le da el artefacto a otros para que puedan usarlo de manera asistida por el investigador. Teniendo en cuenta que el experimento se realiza en un entorno real, el cual es un área de trabajo bastante cerrada, hay pocos expertos que trabajan con estos datos, por lo que tenemos un grupo de sujetos reducido que nos permite utilizar esta metodología. Además, esto nos ha permitido trabajar de forma cercana con los sujetos, utilizar herramientas como entrevistas personales y la realización de un *focus-group,* que no habría sido posible teniendo una cohorte mayor en el experimento. Sin embargo, las herramientas seleccionadas para medir la validez de la herramienta nos permiten recoger datos cualitativos, por lo que no podemos hacer análisis estadísticos sobre ellos.

El prototipo descrito en el capítulo 3.4 se ha diseñado específicamente para explotar el Esquema Conceptual del Cáncer de Mama descrito en el capítulo 3.1 y poder detectar las posibles mejoras e inconvenientes que supone la utilización de sistemas de información basados en modelos conceptuales en entornos como el que aquí planteamos.

Para lograr este objetivo principal, planteamos dos subobjetivos. El primero de ellos es comparar el uso de la herramienta planteada con el procedimiento tradicional utilizado actualmente en el INCLIVA para almacenar los datos de los pacientes y realizar análisis de los datos almacenados. Nos centraremos, principalmente, en las diferencias temporales a la hora de realizar el mismo ejercicio utilizando ambos procedimientos y en las diferencias en cuanto a precisión y satisfacción del usuario.

El segundo subobjetivo es la realización de un *focus-group* para conseguir algunos datos cualitativos sobre la herramienta donde se pretende obtener pros y contras del método tradicional y de la herramienta propuesta en esta tesis, así como posibles mejoras o potencial futuro de la herramienta planteada y del Esquema Conceptual que la sustenta.

4.2. Caso de estudio

4.2.1. Qué es el INCLIVA

La Fundación para la Investigación del Hospital Clínico de la Comunidad Valenciana, INCLIVA, se constituyó en el año 2000 como fundación privada y sin ánimo de lucro bajo el protectorado de la Generalitat Valenciana, siendo la primera fundación de la Comunidad Valenciana adscrita a un hospital público. 2011 marca un hito en su desarrollo al obtener la acreditación como Instituto de Investigación Sanitaria del Instituto de Salud Carlos III. Desde su creación, tiene como objetivo genérico impulsar, promover y favorecer la investigación científica y técnica en el seno del Departamento de Salud de Valencia Clínico-Malvarrosa.

Las tradicionales relaciones entre el Hospital Clínico Universitario de Valencia y la Facultad de Medicina de la Universidad de Valencia, se materializan en el convenio de colaboración existente entre la Fundación INCLIVA y la Universidad mediante el cual se adscriben a la Fundación un conjunto de grupos de investigación de excelencia de ambas instituciones y el Instituto Universitario Valenciano de Infertilidad, IUVI.

En los últimos años, la institución ha experimentado un crecimiento de su actividad que se ha materializado en el desarrollo de proyectos europeos, de ensayos clínicos internacionales y en la participación de Redes de Investigación en Red como el CAIBER o el BIOBANCO. Especial mención merece la Unidad de Ensayos Clínicos Fase I Oncológicos, dado que el Hospital Clínico de Valencia es uno de los pocos centros que realiza este tipo de estudios en España.

La investigación de la Fundación del Hospital Clínico Universitario de Valencia - Instituto de Investigación Sanitaria INCLIVA está organizada en torno a cuatro líneas de investigación priorizadas, resultantes de un profundo análisis interno en busca de excelencia científica y basadas en las

necesidades de salud de la población, del sistema de I+D+i y de los investigadores de la institución:

- Línea de Investigación Cardiovascular.
- Línea de Investigación en Metabolismo y Daño Orgánico.
- Línea de Investigación en Medicina Reproductiva.
- Línea de Investigación en Oncología.

Es en esta última línea de investigación donde se integra el *Grupo de Investigación en Biología en Cáncer de Mama* junto al cual se ha creado el proyecto nacional coordinado "Análisis de la desregulación transcriptómica en tejido tumoral de mujeres jóvenes con cáncer de mama e implicaciones funcionales " (PI13/02247).

4.2.2. Qué es el proyecto de Cáncer de Mama en Mujeres Jóvenes

Este proyecto se construye con el objetivo de detectar diferencias significativas en expresión de microARNs, metilación global y transcriptómica de muestras tumorales de pacientes de cáncer de mama menores de 35 años y, a su vez, proporcionar una plataforma robusta y compatible que permita la gestión y almacenaje correcto de toda esta información genómica junto a la información clínica de las pacientes, así como una buena plataforma de consultas y generación de datos más elaborados para extraer conclusiones relevantes de los mismos de una forma sencilla y eficiente.

Dada la amplitud del estudio, se requiere un enfoque multidisciplinar que aconseja la elaboración de un proyecto coordinado en el que participan dos grupos con experiencia en diferentes áreas que abarca el proyecto. Aunque hay numerosos puntos de conexión temática entre los dos grupos, el subproyecto 1, liderado por la Dra. Gloria Ribas, con personal clínico y biólogos moleculares abordan las facetas clínicas y de experimentación básica molecular mientras que el subproyecto 2, cuyo investigador principal

es el Dr. Óscar Pastor, incide en la parte diseño de un Sistema de Información correctamente estructurado y diseñado específicamente para gestionar de manera efectiva y eficiente toda la información genómica y transcriptómica generada, además de integrarlo con datos clínicos.

Este proyecto ha sido el núcleo de unión de los dos grupos de disciplinas tan distintas y de los líderes de los mismos, provocando la co-dirección de la presente tesis doctoral con un marcado carácter multidisciplinar.

4.2.3. Necesidades del grupo de investigación

Bajo el entorno de este proyecto, *Grupo de Investigación en Biología en Cáncer de Mama* necesita un sistema de información que gestione los datos clínicos y biológicos de los pacientes para conseguir abordar los objetivos planteados en el proyecto y detectar adecuadamente marcadores genéticos específicos del cáncer de mama en mujeres jóvenes que posibiliten avanzar hacia la medicina personalizada. El grupo empezó análisis ómicos mediante una plataforma de estudios de expresión de microARNs. Centrándonos en el trabajo desarrollado en esta tesis, vamos a abordar las necesidades de este proyecto relacionadas con los datos clínicos y de expresión de microARNs.

Habiendo estudiado en detalle el procedimiento llevado a cabo por los clínicos y biólogos para recolectar los datos clínicos de los pacientes, estudiar los datos biológicos y evaluar los resultados en conjunto (véase capítulo 2.1 de esta tesis) podemos detectar las dificultades, en relación a la gestión y análisis de datos, con las que se encuentran estos profesionales a la hora de llevar a cabo su trabajo. Estas dificultades podríamos definirlas en los siguientes puntos:

- Introducción y gestión ineficiente de datos clínicos y biológicos.
- Realización manual y costosa de búsquedas de información entre sus propios datos para la realización de análisis y estudios estadísticos.
- Procesos largos y tediosos en el análisis de los datos de expresión.
- Realización de múltiples procedimientos y utilización de varias

herramientas para conseguir relacionar los datos de expresión con información sobre genes y *pathways* afectados.

Teniendo en cuenta estas restricciones hemos diseñado la herramienta definida en el capítulo 3.4 de esta tesis, pretendiendo mejorar los procedimientos de trabajo del equipo, haciéndolos más ágiles y eficientes.

4.2.4. Sujetos experimentales

La investigadora principal del *Grupo de Investigación en Biología en Cáncer de Mama* es Ana Lluch Hernández, Jefa del Servicio de Hematología y Oncología del Hospital Clínico de Valencia. Es a este Grupo de Investigación al que pertenecen la Dra. Gloria Ribas Despuig, bióloga, investigadora principal de este proyecto coordinado y co-directora de esta tesis doctoral, y las investigadoras Dra. María Peña (doctora en Biología) y Dra. Maite Martínez (doctora en Medicina). Son estas tres últimas personas las que participarán como sujetos de nuestro experimento. Todas ellas son expertas en el manejo de datos clínicos y biológicos y llevan años trabajando con datos sobre cáncer de mama, lo que las convierte en los sujetos ideales para nuestro experimento, ya que conocen perfectamente el tipo de datos que vamos a manejar. Además, se incluye en el experimento Anna Heredia, una Ingeniera Biomédica con experiencia en la manipulación de este tipo de datos, que aportará una visión externa al experimento.

4.2.5. Preparación del sistema: Carga de la base de datos con los datos clínicos

Antes de iniciar los ejercicios de validación es esencial llevar a cabo la carga de datos clínicos reales que permitan verificar el correcto funcionamiento del sistema. Para ello, el *Grupo de Investigación en Biología en Cáncer de Mama* nos ha proporcionado su base de datos de pacientes, totalmente anonimizada. En este caso, los datos de los pacientes se encuentran

almacenados en ficheros de Microsoft Excel de forma totalmente heterogénea y desestructurada. Los scripts de carga de datos se han realizado en Java, utilizando librerías de lectura de ficheros Excel y de conexión con la base de datos de MySQL.

Antes de iniciar la carga, es necesario entender perfectamente la nomenclatura utilizada por los médicos a la hora de almacenar los datos y establecer un conjunto de correspondencias que determinen dónde y cómo debe almacenarse cada uno de los datos que nos proporcionan a través de dichos ficheros. En este caso, tenemos dos ficheros de datos de pacientes, uno con los datos de mujeres jóvenes y otro con los datos de mujeres de edad avanzada, con la información almacenada estructurada de forma distinta. Dicha correspondencia de datos queda especificada en la Tabla 5 y la Tabla 6 que aparecen a continuación, con los datos de pacientes jóvenes y mayores respectivamente.

CORRESPONDENCIA FICHERO DE DATOS CLÍNICOS DE MUJERES JÓVENES			
COLUMNA	TABLA	ATRIBUTO	DATOS PREDEFINIDOS
EOBC	muestra	id_muestra	
Edad al Dco	paciente	edad_al_dco	Edad en años
Perfil	muestra _tumor	Perfil_tumoral (Si aparece CD IN SITU o CDI el atributo donde debe incluirse es Tipo_hist)	• Luminal A • Luminal B • Her2 • Her2 Luminal • Triple negativo • Luminal/ Triple negativo • Her2/Luminal B
Tipo Histológico	muestra _tumor	Tipo_hist	• CDI (incluir "CD sin tipo especial") • Carcinoma tubular • Medular atípico (incluir "Medular") • Lobulillar

			• Cribiforme • CI con patrón mixto tubular y cribiforme
RE	prueba _biologica _inmunohisto quimia	RE (si el valor es 0 poner NEGATIVO, sino POSITIVO)	• POSITIVO • NEGATIVO
RP	prueba _biologica _inmunohisto quimia	RP (si el valor es 0 poner NEGATIVO, sino POSITIVO)	• POSITIVO • NEGATIVO
KI67	prueba _biologica _inmunohisto quimia	KI67	El valor es un %
HER2	prueba _biologica _inmunohisto quimia	HER2 Si pone "AMPLIFICADO" o "FISH AMPLIFICADO" significa POSITIVO Si pone "C-erb.2+" y no hay más datos es NEGATIVO Si pone "C-erb2, Fish no amplificado" es NEGATIVO Si pone "HERB-2 POSITIVA" es POSITIVO	• POSITIVO • NEGATIVO • NH (cuando no hay valor)
Grado	muestra _tumor	grado_histologico	• I • II • III

Tamaño tumoral	muestra _tumor	Tamano o clasif_tamano (si aparece uno de los valores Tis, T1a, T1b, T1c, T1, T2, T3 o T4) Si es INFLAMATORIO poner en clasif_tamano = T4 Si es MULTIFOCAL y están los focos poner el tamaño del más grande. Si no están los focos no poner nada, dejarlo vacío.	Valor en centímetros
N	muestra _ganglios	Extirpados = 1 Positivos = 1(si es POSITIVO o N+) 0(si es NEGATIVO)	
Neoadyuvancia	tratamiento _quimioterapia	tipo (si es SI es NEOADYUVANTE, si es NO no ponemos nada) Añadimos el resto de información en "esquema"	• ADYUVANTE • NEOADYUVANTE
Esquema de tratamiento neoadyuvante	tratamiento_ quimiotera-pia	esquema Tipo = NEOADYUVANTE	
Tipo de cirugía	tratamiento _cirugia	Tipo (si aparece "Tumorectomía"= "Conservadora", "MRM"="Mastectomí	• Conservadora • Mastectomía • Ganglio Centinela • Linfadenectomía

		a", "con vaciamiento axilar"="Linfadenectomía (nivel= completa)")	
Metastásicas de inicio	metastasis (relacionada con el episodio de diagnóstico)	Creamos una instancia de la clase metástasis	
TNM tras neoadyuvancia/ cirugia	muestra_tumor muestra_ganglios evolucion_tumor metastasis tratamiento_quimio_neo tratamiento_quimioterapia	Suele contener un dato del tipo: CDI yp T2(5 cm) Yp N2/14 M1 "Descripción del tratamiento neoadyuvante y/o adyuvante llevado a cabo" Donde: El primer valor corresponde al atributo "tipo_hist" de "muestra_tumor". El segundo valor "y" indica si ha habido neoadyuvancia antes de la cirugía, lo que debe almacenarse en "post-neoadyuvancia" de "Evolución tumor", y además incluye una "c" o "p" indicando que los datos son "clínicos" (tomados antes de la cirugía) o "patológicos" (posteriores a la cirugía), lo que debe almacenarse en "tipo_estadio" de "Muestra_tumor". El tercer valor, que empieza por una "T", corresponde	Los valores de "clasif_tamano" deben ser: • Tis • T1 • T1a • T1b • T1c • T2 • T3 • T4

		a la "clasif_tamaño" de "muestra_tumor". El cuarto valor, que aparece entre paréntesis, corresponde al tamaño del tumor y debe almacenarse en "tamano" en "muestra_tumor". El quinto valor se desprecia, ya que es el mismo que el segundo valor. El sexto valor, precedido por una "N" indica el número de ganglios afectos (previo a la /) del total de ganglios extraídos (posterior a la /) lo que se almacena en la tabla "muestra_ganglios" en los atributos "positivos" y "extirpados" respectivamente. El séptimo valor, precedido por una "M", contiene las metástasis lo que generaría una instancia de la clase "metastasis" y se incluiría la localización en el caso en el que quede reflejada en la descripción posterior. El octavo y último valor se escapa de cualquier patrón, ya que es texto libre, y contiene información sobre el tratamiento llevado a	

		cabo. Puede contener información de las tablas "tratamiento_quimioterapia" y "tratamiento_quimio_neoadyuvante".	
Respuesta completa patológica	tratamiento_quimioterapia	patologica	• COMPLETA • PARCIAL
Radioterapia	tratamiento_radioterapia	Creamos una instancia de la clase Tratamiento_radioterapia	
Dosis radioterapia	tratamiento_radioterapia	dosis	
Nacionalidad	paciente	Nacionalidad	
Tabaco	habito_toxico	Tipo=Tabaco	
Alcohol	habito_toxico	Tipo=Alcohol	
Dieta	paciente	Dieta	
Peso	paciente	peso	
Talla(m)	paciente	talla	
IMC	paciente	IMC	
HTA	antecemedico_paciente	Tipo_anteced_medico=HTA	
DM	antecemedico_paciente	Tipo_anteced_medico=DM	
Menarquia	vida_reproductiva	Menarquia	
Menopausia	vida_reproductiva	FUR	
Nhijos	vida_reproductiva	GxAyPz Gestaciones=x Abortos=y Partos=z	

		Si pone IVE es Interrupción Voluntaria del Embarazo, lo que se traduce a 1 gestación y 1 aborto.	
EPE	vida _reproductiva	EPE	
Lactancia	vida _reproductiva	Lactancia	
ACOs	anticoncep-tivo	Nombre (si existe, si sólo aparece Si/No, ponerlo en Nombre) Tipo (si existe) Duración (de la tabla anticonceptivo _vidareproductiva)	
THS	vidareproductiva	THS	
BRCA	prueba_biologicas_ genetica	Gen = BRCA1 o BRCA2 o BRCA (este último no debería aparecer, pero si aparece lo ponemos) Resultado = mutado (si es el caso) Si aparece "BRCA no informativo" es "no mutado".	
Antec. familiares	anteceoncolo-gico_familiar	Parentesco (intentar identificar el parentesco)	

		Características _enfermedad (copiar el campo completo)	
Fecha ultimo seguimiento	episodio _seguimiento	Fecha (tabla Episodio) Descripción (si el campo es distinto a una fecha) Si aparece "dd/mm/aaaa (exitus)" incluir además en la tabla "Estado_paciente" la "fecha" y el "tipo" = "Exitus"	
Recaida 1 (año/lugar)	episodio _recaida	Fecha_episodio = fecha (si la hay) Lugar = localización del nuevo tumor (si lo indica) Descripción = copiar todo el texto del campo PE significa "Progresión de Enfermedad" ILE significa "Intervalo Libre Enfermedad"	
Nº Biopsia	muestra (asociada al	id_muestra	

	episodio de recaída)		
RE	prueba _biologica _inmunohis-toquimia (asociada al episodio de recaída)	RE (si el valor es 0 poner NEGATIVO, sino POSITIVO)	• POSITIVO • NEGATIVO
RP	prueba _biologica _inmunohis-toquimia (asociada al episodio de recaída)	RP (si el valor es 0 poner NEGATIVO, sino POSITIVO)	• POSITIVO • NEGATIVO
HER2	prueba _biologica _inmunohis-toquimia (asociada al episodio de recaída)	HER2 Si pone "AMPLIFICADO" o "FISH AMPLIFICADO" significa POSITIVO Si pone "C-erb.2+" y no hay más datos es NEGATIVO Si pone "C-erb2, Fish no amplificado" es NEGATIVO Si pone "HERB-2 POSITIVA" es POSITIVO	• POSITIVO • NEGATIVO
KI67	prueba _biologica	KI67	El valor es un %

	_inmunohis-toquimia (asociada al episodio de recaída)		
Tipo biopsia	muestra (asociada al episodio de recaída)	Técnica (Los datos que contiene son de localización, por tanto, no almacenar el valor)	
Órgano de biopsia	muestra (asociada al episodio de recaída)	Localización	
Tratamiento	tratamiento _quimioterapia (asociada al episodio de recaída)	Esquema. Tipo=Adyuvante	
Tipo de tratamiento	tratamiento (asociada al episodio de recaída)	Instancia del tipo que sea.	
Tipo de tratamiento biológico	tratamiento _biológico (asociada al episodio de recaída)	Tipo.	
Nº de ciclos	tratamiento _quimioterapia (asociada al episodio de recaída)	Num_ciclos.	
Esquema qt	tratamiento _quimioterapia (asociada	Esquema.	

	al episodio de recaída)		
Tipo hormonotera-pia	Tratamiento _hormonote-rapia (asociada al episodio de recaída)	Tipo.	
2ª Recaída	episodio _recaida	Descripción. Si hay fecha incluirla en "fecha" de la tabla "Episodio" correspondiente.	
Otro cáncer	antecemedico _paciente	Tipo_anteced_medi-co = Otros.	
Otras comorbilida-des	antecemedico _paciente	Tipo_anteced_medi-co = Otros	
Éxitus	estado _paciente	Fecha. Tipo = "Éxitus"	
Perdida	estado _paciente	Fecha. Tipo = "Pérdida de seguimiento"	
Gestante al dco	episodio _diagnostico	descripcion_dco. Gestante_al_dco =SI	
Lactancia al dco	episodio _diagnostico	descripcion_dco. Lactancia_al_dco =SI	

Tabla 5. Tabla de correspondencias de la base de datos con el Excel de datos de mujeres jóvenes.

FICHERO DE DATOS CLÍNICOS DE MUJERES MAYORES			
COLUMNA	TABLA	ATRIBUTO	DATOS PREDEFINIDOS
ID	muestra	id_muestra	
Edad al Dx	paciente	edad_al_dco	Edad en años
DCO	muestra_tumor	tipo_hist	• CDI • Carcinoma tubular • Medular atípico • Lobulillar • Cribiforme • CI tubular y cribiforme • CI Micropapilar • C Lobulillar y Ductal • C Mucinoso
Tipo	muestra_tumor	perfil_tumoral	• Luminal A • Luminal B • Her2 • Her2 Luminal • Triple negativo • Luminal / Triple negativo • Her2/Luminal B
Grado	muestra_tumor	grado_histologico	• I • II • III
Tamaño (cm)	muestra_tumor	tamano	Valor en centímetros
pT	muestra_tumor	El primer valor incluye una "c" o "p" indicando que los datos son "clínicos" (tomados antes de la cirugía)	Los valores de "clasif_tamano" deben ser: • Tis • T1 • T1a

		o "patológicos" (posteriores a la cirugía), lo que debe almacenarse en "tipo_estadio" de "Muestra_tumor". El segundo valor, que empieza por una "T", corresponde a la "clasif_tamaño" de "muestra_tumor".	• T1b • T1c • T2 • T3 • T4
N	muestra _ganglios	Siempre que en la columna siguiente no tengamos los valores exactos de ganglios positivos y extirpados: Si el valor es POS: Positivos = 1 Extirpados = 1 Si el valor es NEG: Positivos = 0 Extirpados = 1	
pN	muestra _ganglios	Positivos = Valor antes de la / Extirpados = Valor después de la /	
pM	metastasis	Si aparece "Si" o un valor mayor que 0 detrás de la "M"	

		creamos una instancia de la clase metástasis.	
RE	prueba _biologica _inmunohisto-quimia	RE = POSITIVO (Si el valor es "POS") RE = NEGATIVO (Si el valor es "NEG")	• POSITIVO • NEGATIVO
%RE	prueba _biologica _inmunohisto-quimia	No almacenamos el valor	
RE Allred	prueba _biologica _inmunohisto-quimia	No almacenamos el valor	
RP	prueba _biologica _inmunohisto-quimia	RP = POSITIVO (Si el valor es "POS") RP = NEGATIVO (Si el valor es "NEG")	• POSITIVO • NEGATIVO
%RP	prueba _biologica _inmunohisto-quimia	No almacenamos el valor	
RP Allred	prueba _biologica _inmunohisto-quimia	No almacenamos el valor	
HER2	prueba _biologica _inmunohisto-quimia	HER2 = POSITIVO (Si el valor es "POS")	• POSITIVO • NEGATIVO

		HER2 = NEGATIVO (Si el valor es "NEG")	
HER2 IHQ	prueba _biologica _inmunohisto-quimia	No almacenamos el valor	
HER2 FISH	prueba _biologica _inmunohisto-quimia	No almacenamos el valor	
%KI67	prueba _biologica _inmunohisto-quimia	KI67. Si hay dos valores poner el más alto.	El valor es un %
Radioterapia	tratamiento _radioterapia	Si el valor es "SI" creamos una instancia de la clase Tratamiento _radioterapia	
Dosis radioterapia	tratamiento _radioterapia	dosis	
Nacionalidad	Paciente	nacionalidad	
Tabaco	Habito_toxico	tipo=Tabaco	
Alcohol	Habito_toxico	tipo=Alcohol	
Dieta	Paciente	dieta	
IMC	Paciente	IMC	
HTA	antecemedico _paciente Farmaco _habitual Farmaco _habitual _paciente	tipo_anteced _medico = HTA Si hay valor entre paréntesis: Nombre (fármaco _habitual) = valor o valores entre	

		paréntesis si no existen Completar tabla fármaco_habitual _paciente	
DM	antecemedico _paciente	tipo_anteced _medico=DM Si hay valor entre paréntesis: Nombre (farmaco _habitual) = Valor o valores entre paréntesis si no existen Completar tabla farmaco_habitual_ paciente	
Menarquia	vida _reproductiva	menarquia (eliminar la palabra "años", si existe)	
Menopausia	vida _reproductiva antecemedico _paciente	FUR (eliminar la palabra "años", si existe) tipo_anteced_ medico = Otros valor = contenido del paréntesis (si existe)	
Nhijos	vida _reproductiva	partos (eliminar la palabra hijos, si existe)	

		abortos = valor antes de la palabra "abortos" (si existe) gestaciones = partos + abortos	
EPE	vida _reproductiva	EPE (eliminar la palabra "años", si existe)	
Lactancia	vida _reproductiva	lactancia. Valor entre paréntesis. Unificar a meses. Si únicamente aparece el valor "SI" unificar a "12". Si aparece el valor "NO" unificar a "0".	
ACOs	anticonceptivo	nombre (si existe) tipo ("ACO" si pone "SI" u otro tipo si lo especifica) duración (de la tabla anticonceptivo _vidareproductiva) el valor entre paréntesis	
THS	vida _reproductiva	THS. Valor entre paréntesis	
BRCA	prueba _biologicas _genetica	gen = BRCA1 o BRCA2 o BRCA (este último no debería aparecer,	

		pero si aparece lo ponemos) resultado = mutado (si es POSITIVO), no mutado (si es NEGATIVO)	
Antec. familiares	Anteceoncologi-co_familiar	parentesco = identificar parentesco o parentescos características_enfermedad = resto del contenido distinto a parentesco	
Recaída (año/lugar)	Episodio_recaida	fecha_episodio = fecha (si la hay, adaptarla a formato fecha) descripción = copiar todo el texto del campo	
Otras comorbilida-des	Antecemedico_paciente	tipo_anteced_medico = Otros	
Éxitus	Estado_paciente	Fecha. Tipo = "Éxitus"	

Tabla 6. Tabla de correspondencias de la base de datos con el Excel de datos de mujeres de edad avanzada

Una vez definidas las correspondencias entre la base de datos procederemos a la implementación y ejecución del script de carga de datos clínicos. Como se puede observar, la estructura de los ficheros de datos de mujeres jóvenes y mujeres mayores no es igual y como consecuencia, se han desarrollado dos scripts de carga con estructura diferente, adaptados a los datos de cada uno de los ficheros, pero con el mismo procedimiento. La razón de que los ficheros de datos no sean iguales se debe a que se han recogido los datos en momentos distintos, de forma retrospectiva, a partir de los datos extraídos de los informes clínicos de las pacientes, insertando columnas a medida que eran necesarias. El script recorre el fichero fila por fila, insertando un paciente en cada una de ellas.

En primer lugar, recoge e inserta los datos propios del paciente, como los datos personales, antecedentes médicos, estados del paciente, vida reproductiva... Posteriormente, crea un episodio de diagnóstico donde incluirá las pruebas, muestras y tratamientos que pertenezcan a la etapa de diagnóstico de la paciente. El siguiente paso será crear uno o varios episodios de recaída, siempre y cuando la paciente haya sufrido alguna recaída de la enfermedad. Para finalizar, se insertarán uno o varios episodios de seguimiento, según los datos que se hayan recogido de la paciente después de haberle dado el alta.

Una vez finalizado todo el proceso y el script haya recorrido todas las filas de los pacientes se procederá a ejecutar el mismo proceso con el otro script cargando el otro fichero Excel de pacientes. De esta manera, tendremos la base de datos preparada para que los sujetos puedan ejecutar el experimento con datos de pacientes reales.

4.3. Experimento

4.3.1. Preguntas de investigación y formulación de la hipótesis

Las preguntas de investigación que se plantean al inicio de este experimento y a las cuales se quiere dar respuesta con los resultados del mismo son las siguientes:

- **RQ1:** ¿Interfiere la utilización de Esquemas Conceptuales en el entorno clínico en la precisión de los datos de los pacientes y de los resultados de los análisis? Nos referimos a precisión cómo la cercanía del resultado de una medida con respecto al valor real [127]. La hipótesis nula de la que partimos para resolver esta pregunta de investigación es: *H_{01}: La precisión de los datos y análisis utilizando Esquemas Conceptuales es similar a la que se tiene utilizando el método tradicional.*
- **RQ2:** ¿La utilización de Esquemas Conceptuales en el entorno clínico mejora la eficiencia de los procedimientos de almacenamiento, gestión y análisis de datos de los pacientes? Nos referimos a eficiencia cómo el grado mediante el cual un sistema o componente lleva a cabo la función para la que fue diseñado con el mínimo consumo de recursos [127]. La hipótesis nula de la que partimos para resolver esta pregunta de investigación es: *H_{02}: La eficiencia de los procedimientos de almacenamiento, gestión y análisis de los datos de los pacientes utilizando software diseñado utilizando Esquemas Conceptuales es similar a la que tiene utilizando el procedimiento tradicional.*
- **RQ3:** ¿Mejora la satisfacción de los clínicos al utilizar el método diseñado en base a los Esquemas Conceptuales planteados en esta tesis respecto a sus procedimientos anteriores? La satisfacción se define como el cúmulo de sensaciones positivas que se sienten al utilizar el producto [127]. La hipótesis nula de la que partimos para

resolver esta pregunta de investigación es: *H_{03}: La satisfacción del usuario al almacenar, gestionar y analizar datos con el software diseñado utilizando Esquemas Conceptuales es similar a la obtenida al realizar el mismo proceso por el método manual.*

4.3.2. Factores y tratamientos

En este apartado se definen los factores y sus niveles para operacionalizar la causa del experimento. Los factores son variables cuyo efecto se quiere conocer a través de las variables respuesta [128]. El experimento estudia un factor: el método utilizado para introducir datos de un paciente y realizar un análisis de datos.

El mismo ejercicio se repetirá en los tres sujetos solicitándoles que realicen las mismas dos tareas utilizando uno u otro método. Las tareas a realizar son:

- Dada una historia clínica, localizar los datos relevantes y almacenarlos.
- Realizar un análisis de todos los datos disponibles para responder a tres cuestiones planteadas.

Para llevar a cabo el ejercicio, se establecen dos niveles o tratamientos: el método tradicional y el método utilizando la herramienta diseñada bajo las técnicas de Modelado Conceptual. El método tradicional se encuentra descrito de forma detallada en el capítulo 2.1.6 de esta tesis. Aplicando dicho método, se llevará a cabo el siguiente procedimiento:

- Tomaremos una Historia Clínica de una paciente en papel totalmente anonimizada y procederemos a introducir los datos que encontremos de forma manual en el Excel que manejan los clínicos con los datos de los pacientes.
- Categorizaremos esos datos para poder procesarlos de forma unificada asignándoles valores numéricos a los datos contenidos en las columnas que nos interesa medir, que según las cuestiones

planteadas son edad, perfil histológico tumoral, recaída, menarquia y edad del primer embarazo.
- Realizamos los cálculos estadísticos apropiados utilizando la tecnología utilizada actualmente para poder obtener el resultado solicitado al inicio de la prueba.

Por otra parte, los sujetos utilizarán la herramienta detallada en el capítulo 3.4 para conseguir el objetivo propuesto en el ejercicio. El procedimiento que se llevará a cabo utilizando la herramienta es el siguiente:

- Tomaremos una Historia Clínica de una paciente en papel para introducir los datos totalmente anonimizados en la herramienta a través de los formularios diseñados para tal fin.
- Utilizaremos la herramienta de análisis de los datos clínicos para establecer los parámetros de búsqueda y obtener el valor solicitado al inicio de la prueba.

4.3.3. Variables respuesta y métrica

Las variables respuesta son los efectos estudiados en el experimento causados por la manipulación de los factores [128]. Para medir la calidad de la solución planteada necesitaremos una variable. Siguiendo la norma ISO 9126-1 [129], la calidad se compone de varias variables: funcionalidad, fiabilidad, usabilidad, eficiencia, mantenibilidad y portabilidad. Para resolver la pregunta de investigación **RQ1** se ha escogido la funcionalidad, ya que se centra en resolver las expectativas del usuario final. Más concretamente, se estudiará su subvariable **precisión**, que se define en la ISO 9126-1 como "la capacidad de un producto software para proporcionar el resultado o efecto correcto o esperado". Mediremos la precisión como el porcentaje de acierto del ejercicio que se llevará a cabo.

En este caso, se ha definido un ejercicio que se ha dividido en dos tareas que se llevarán a cabo de forma correlativa. Las tareas definidas son las siguientes:

T1. Extraer los datos del paciente de una historia clínica en papel y almacenarlos en el sistema.

T2. Analizar los datos almacenados para resolver las tres preguntas planteadas en la hoja de ejercicios.

A partir de estas dos tareas, mediremos la precisión con la que se han llevado a cabo utilizando ambos procedimientos. En primer lugar, veremos si los datos se han introducido de forma correcta en el sistema, utilizando cada uno de los procedimientos. Además, evaluaremos la similitud entre el valor numérico obtenido de forma manual y utilizando el procedimiento propuesto, lo que nos dará finalmente la precisión de buscamos.

Además de la precisión, utilizaremos otra variable para medir la calidad de la herramienta que nos ayudará a resolver la **RQ2**. En este caso, esa variable es la **eficiencia**. Teniendo en cuenta que definimos eficiencia como la consecución de los resultados con el menor gasto de recursos [127], para poder medir esta variable llevaremos a cabo una medición de los tiempos empleados por los sujetos para llevar a cabo cada una de las tareas propuestas realizadas utilizando ambos métodos.

Por otro lado, la pregunta de investigación **RQ3** necesita una variable dependiente que mira la **satisfacción** del usuario. Definiendo la satisfacción como el cúmulo de sensaciones positivas que se sienten al utilizar el producto [127], mediremos esta variable utilizando un cuestionario Likert con una escala de 5 puntos. El instrumento utilizado para medir esta variable es un cuestionario de satisfacción construido utilizando el framework desarrollado por Moody [130]. Moody define una herramienta para evaluar la calidad del modelo en términos de utilidad percibida (UP), facilidad de uso percibida (FUP) e intención de uso (IU). Esta herramienta ha sido validada anteriormente y se usa en múltiples ámbitos y situaciones. Siguiendo el trabajo definido por Moody, hemos diseñado un cuestionario para cada uno de los dos métodos a utilizar (el tradicional y utilizando la herramienta). Aunque el significado de la pregunta sea el mismo en ambos

métodos, hemos adaptado las preguntas a cada uno de los métodos utilizando términos específicos.

En la siguiente Tabla 7 encontramos un resumen de las preguntas de investigación que vamos a resolver, las hipótesis y las variables respuesta utilizadas para comprobar dichas hipótesis:

RQs	Hipótesis	Variables respuesta	Métrica
RQ1	$H_0 1$	Precisión	% acierto
RQ2	$H_0 2$	Eficiencia	Tiempo
RQ3	$H_0 3$	Satisfacción	Cuestionarios

Tabla 7. Tabla resumen de las preguntas de investigación, hipótesis, variables respuesta y métrica utilizadas en la validación.

4.3.4. DISEÑO DEL EXPERIMENTO

Para ejecutar este experimento hemos escogido un **diseño de medidas repetidas** [128]. Para cada sujeto, el experimento se lleva a cabo en dos sesiones. En la primera sesión los sujetos resolverán el problema con el método tradicional. En la segunda, lo harán utilizando la herramienta planteada en esta tesis. Las ventajas de este diseño son que las condiciones que se van a comparar son las mismas. En este caso, tenemos cuatro sujetos que van a ejecutar el experimento. Cada sujeto ejecutará cada una de las dos tareas planteadas de forma consecutiva para cada experimento, ya que para completar la T2 necesitamos tener los resultados de la T1, y trabajarán con el mismo problema de partida en ambos métodos (introducirán en ambos sistemas la misma historia clínica de un paciente).

4.3.5. PROCEDIMIENTO DEL EXPERIMENTO

Para llevar a cabo la ejecución del experimento, nos reuniremos con cada uno de los sujetos individualmente para llevar a cabo la ejecución de las tareas.

Hay un total de 2 tareas. Se cronometra de forma oculta para el usuario el tiempo que éste tarda en completar cada una de las tareas, divididas en subtareas tal como muestra la siguiente Tabla 8:

<table>
<tr><th></th><th></th><th colspan="2">Tradicional</th><th>Herramienta</th></tr>
<tr><td>Tarea 1</td><td>Inserción Paciente</td><td colspan="2"></td><td></td></tr>
<tr><td rowspan="4">Tarea 2</td><td rowspan="2">Categorización</td><td>Jóvenes</td><td></td><td rowspan="2">(no procede)</td></tr>
<tr><td>Mayores</td><td></td></tr>
<tr><td>Cuestión 1</td><td colspan="2"></td><td></td></tr>
<tr><td>Cuestión 2</td><td colspan="2"></td><td></td></tr>
<tr><td>Cuestión 3</td><td colspan="2"></td><td></td></tr>
</table>

Tabla 8. Resumen de las tareas y subtareas a llevar a cabo por los sujetos del experimento

Se medirá el tiempo que ha tardado cada usuario para llevar a cabo cada subtarea. Podemos ver en la tabla que la categorización únicamente es necesaria en el método tradicional, ya que la herramienta no necesita realizar una categorización para poder analizar los datos. Además, esta categorización la dividiremos en "jóvenes" y "mayores" igual que los ficheros Excel que disponemos, ya que tienen que categorizar los dos ficheros Excel por separado para poder analizar los datos.

Por otra parte, para obtener el tiempo dedicado a resolver cada una de las cuestiones planteadas en la tarea 2 utilizando el método tradicional, es necesario sumarle al tiempo de resolución de cada una de las cuestiones el tiempo necesario para categorizar los datos. Ya que las cuestiones 1 y 3 corresponden a mujeres jóvenes y la 2 a mujeres mayores, a los tiempos de las cuestiones 1 y 3 le sumaremos la mitad del tiempo empleado para categorizar los datos de mujeres jóvenes. Por otro lado, ya que únicamente tenemos una cuestión en la que se solicitan los datos de mujeres mayores, al tiempo empleado para resolver la cuestión 2 le sumaremos el tiempo íntegro dedicado a categorizar el fichero Excel de mujeres mayores.

Al finalizar las dos tareas realizadas utilizando cada uno de los tratamientos, el usuario debe rellenar un cuestionario de escala Likert de 5 puntos donde se recoge la opinión de usabilidad de cada tratamiento.

4.3.6. Amenazas a la validez

En este punto se valoran los posibles factores que pueden hacer que el experimento pierda su validez, en el caso en el que ocurriesen, y como vamos a manejarlos para evitar que esto suceda.

- Validez de conclusión:
  - Sujetos de heterogeneidad aleatoria: Esta amenaza aparece si dentro del grupo de usuarios estudiados hay unos que tienen más experiencia en el uso de datos clínicos que otros. En la evaluación realizada, todos los sujetos están acostumbrados a tratar con datos clínicos y biológicos en su trabajo diario.

- Validez interna:
  - Maduración: Esta amenaza contempla la posibilidad de que los usuarios reaccionen de manera distinta conforme transcurre el tiempo del experimento, bien por cansancio, prisa o por aburrimiento. Esta amenaza se ha intentado solucionar haciendo una evaluación corta. La historia clínica seleccionada es de un proceso básico de diagnóstico y tratamiento de cáncer de mama. De la misma manera, las cuestiones planteadas en la tarea 2 son sencillas, sin necesidad de realizar cálculos estadísticos complejos. El tiempo máximo estimado de toda la prueba era de 90 minutos.
  - Instrumentación: A pesar de que tanto tareas como cuestionarios son los mismos para todos los usuarios, éstos pueden sufrir distintas interpretaciones por parte del usuario que los lea. Para evitar esta amenaza, se hizo una explicación detallada de las tareas y los cuestionarios previa a la realización del experimento.

- Validez de construcción:
  - Suposición de la hipótesis: Esta amenaza contempla la posibilidad de que el usuario suponga el propósito y el resultado esperado del experimento y base su comportamiento en esas suposiciones. Esta amenaza se ha minimizado realizando el experimento sin que los sujetos conozcan el resultado correcto de las cuestiones de la tarea 2 y utilizando primero el método tradicional (donde es más fácil equivocarse al realizar el análisis de los datos).

- Validez externa:
  - Interacción de la selección y los tratamientos: Este es el efecto de tener una población que no sea representativa de la población a la que queremos generalizar. Esta amenaza se minimiza tomando como sujetos personas de distinto perfil profesional (oncólogos, biólogos e ingenieros biomédicos, en este caso), pero que coinciden en la manipulación de datos clínicos y biológicos en sus tareas diarias. Nuestra población queda representada como aquellas personas que manipulan este tipo de datos, independientemente del perfil profesional.

4.3.7. Instrumentos utilizados en el experimento

Tal y como se ha comentado en la sección del procedimiento del experimento (sección 4.3.5) se han utilizado dos instrumentos: las tareas (integradas en dos ejercicios) y los cuestionarios para capturar la usabilidad percibida por el usuario.

Ejercicios

Hay un total de dos ejercicios compuestos por dos tareas en cada uno con las que se garantiza la interacción del usuario con cada uno de los

tratamientos. En el Anexo II se puede ver en detalle la información recogida en ambos ejercicios.

El primer ejercicio a realizar por los usuarios es el de llevar a cabo las dos tareas utilizando el **método tradicional**:

- Material proporcionado:
  - Historia Clínica del paciente
  - Archivo Excel con datos de mujeres jóvenes (utilizado por los oncólogos del INCLIVA para gestionar los datos de sus pacientes)
  - Archivo Excel con datos de mujeres mayores (utilizado por los oncólogos del INCLIVA para gestionar los datos de sus pacientes)

- **Tarea 1**: Introducir todos los datos contenidos en la Historia Clínica del paciente en el Excel de pacientes correspondiente. El sujeto deberá valorar si la Historia Clínica pertenece a una paciente joven o a una paciente mayor y completar el fichero Excel correspondiente.
- **Tarea 2**: Analizar los pacientes contenidos en los ficheros Excel utilizando las herramientas habituales para resolver las siguientes cuestiones:
  - Número de pacientes **menores de 35 años** con perfil tumoral **HER2** que hayan sufrido al menos una **recaída.**
  - Número de pacientes mayores de 60 años con perfil tumoral Luminal A con una menarquia menor o igual a 12 años.
  - Número de pacientes menores de 35 años que hayan tenido su primer embarazo entre los 20 y los 30 años (incluidos) con perfil tumoral Triple Negativo.

Para ello, se proporciona la siguiente tabla de categorización:

VARIABLE CLÍNICA	CATEGORIZACIÓN
Edad al diagnóstico	**1**: <35 años **2**: >60 años
Edad de la menarquia	**1**: ≤ 12 años **2**: 13-14 años **3**: ≥ 15 años **4**: Desconocido
Recaída de la enfermedad	**1**: Sí **2**: No **3**: Desconocido
Perfil tumoral inmunohistoquímico	**1**: Luminal A **2**: Luminal B **3**: Luminal (Ki 67 no especificado) **4**: HER 2 **5**: Luminal/HER2 **6**: Triple negativo **7**: Desconocido
Edad del primer embarazo	**1**: < 20 años **2**: 20-30 años **3**: > 30 años **4**: Desconocido

Tabla 9. Tabla de categorización de las variables clínicas

El siguiente ejercicio a llevar a cabo, es la realización de las mismas dos tareas utilizando la **herramienta** propuesta en esta tesis:

- Material proporcionado:
  - Historia Clínica del paciente
  - Herramienta de gestión de datos diseñada en esta tesis.

- **Tarea 1**: Introducir todos los datos contenidos en la Historia Clínica del paciente en la base de datos utilizando el formulario de "Crear nuevo paciente".
- **Tarea 2:** Analizar los pacientes disponibles en la base de datos utilizando el formulario "Búsquedas en pacientes" para resolver las

siguientes cuestiones:

- Número de pacientes **menores de 35 años** con perfil tumoral **HER2** que hayan sufrido al menos una **recaída.**
- Número de pacientes mayores de 60 años con perfil tumoral Luminal A con una menarquia menor o igual a 12 años.
- Número de pacientes menores de 35 años que hayan tenido su primer embarazo entre los 20 y los 30 años (incluidos) con perfil tumoral Triple Negativo.

CUESTIONARIOS

Tras ejecutar las tareas de cada tratamiento, el usuario debe rellenar las preguntas del cuestionario referentes a la usabilidad del mismo. Es importante remarcar que el cuestionario consiste en una escala Likert de 5 puntos. Cada pregunta del cuestionario se redacta como una sentencia en afirmativo. El usuario debe marcar en la escala la casilla que está más cerca de su opinión. En el Anexo I encontramos los cuestionarios proporcionados a los sujetos. Las posibilidades de respuesta del usuario son las siguientes:

- Si el usuario marca la primera casilla, es que está totalmente en desacuerdo con la afirmación.
- Si el usuario marca la segunda casilla, es que está bastante en desacuerdo con la afirmación.
- Si el usuario marca la tercera casilla es que no lo tiene claro o está indeciso, término medio.
- Si el usuario marca la cuarta casilla es que está bastante de acuerdo con la afirmación.
- Si el usuario marca la quinta casilla es que está totalmente de acuerdo con la afirmación.

4.4. Resultados

4.4.1. Análisis de datos

Esta sección analiza los datos obtenidos en el experimento. El primer paso en el análisis de datos es el de eliminar de los datos capturados en el experimento aquellos que son anómalos. En este caso, nos encontramos con 3 datos anómalos. Dos de ellos son datos de categorización del fichero de Excel de mujeres jóvenes y el tercero corresponde al tiempo empleado en resolver la cuestión 1 de la tarea dos del método tradicional. Estos tiempos son más altos que la media, debido a que los usuarios intentaron realizar el análisis de los datos utilizando la herramienta de análisis estadísticos SPSS con un conocimiento limitado de la misma, pasando finalmente a la utilización de la herramienta Excel para la realización de la categorización de las variables y de los cálculos, al igual que hicieron desde un principio los otros dos sujetos.

En base a estos resultados, se estudia si las hipótesis nulas definidas en el diseño experimental son o no ciertas. El total de hipótesis incluidas en el diseño experimental son tres:

- H_{01}: La precisión de los datos y análisis utilizando Esquemas Conceptuales es similar a la que se tiene utilizando el método tradicional.
- H_{02}: La eficiencia de los procedimientos de almacenamiento, gestión y análisis de los datos de los pacientes utilizando software diseñado utilizando Esquemas Conceptuales es similar a la que tiene utilizando el procedimiento tradicional.
- H_{03}: La satisfacción del usuario al almacenar, gestionar y analizar datos con el software diseñado utilizando Esquemas Conceptuales es similar a la obtenida al realizar el mismo proceso por el método manual.

A continuación, se estudia cada una de estas hipótesis en una subsección distinta.

H01: La precisión de los datos y resultados de los análisis

Como hemos comentado en la sección 4.3.3, para poder conocer la precisión de los datos y resultados del análisis tendremos que estudiar las métricas recogidas en el experimento.

En primer lugar, comprobaremos si los datos se han introducido de forma correcta en el sistema, utilizando cada uno de los procedimientos. Para ello, hemos contado los datos que se han introducido desde la misma Historia Clínica en el sistema utilizando ambos tratamientos. Estos datos quedan recogidos en la fila "Datos insertados" de la tabla disponible en el Anexo II o en formato resumido en la Tabla 10.

Analizando los datos, podemos observar que, a partir de una misma Historia Clínica, se han almacenado, de media, 26 datos utilizando el método tradicional y 75 utilizando la herramienta propuesta (lo que supone unos 49 datos de los 79 que hay en la historia clínica que no se han almacenado utilizando la herramienta tradicional). Esto es debido a varios factores.

El primero de ellos es que utilizando el método tradicional hay algunos datos que se almacenan como uno solo en la misma celda. Por ejemplo, el caso de 3 gestaciones, 2 abortos y 1 parto, se almacena en el método tradicional codificado en una celda como "G3A2P1", o bien indicándolo como "1 hijo" (perdiendo en este caso la información sobre gestaciones y abortos). También nos encontramos casos como los antecedentes oncológicos familiares, donde en el método tradicional quedarían todos los familiares almacenados como texto libre de forma consecutiva en una misma celda. Teniendo en cuenta este factor a la hora de almacenar los datos, podríamos decir que el total de datos a almacenar siguiendo el método tradicional es de 66 datos en lugar de los 79 de los que hablábamos en el párrafo anterior (por lo cual, volviendo a hacer el cálculo, se siguen quedando sin almacenar 40 datos de la historia clínica si utilizamos el método tradicional). De esta manera, se pierde precisión en los datos y se

complica su posterior análisis debido a la heterogeneidad de datos que nos encontramos en el método tradicional.

RESULTADOS CUANTITATIVOS VALIDACIÓN	MEDIAS y PORCENTAJES	
	Trad.	Herram.
Tpo. Inserción Paciente	10'43" (24"/dato)	26'22" (21"/dato)
Datos insertados (79 datos en HC Herr. 66 datos en HC Trad.)	26 (40%)	75 (94'9%)
Tpo. Categorización	Jov. 13'31" May. 4'15"	----------
Tpo. Ejercicio 2.1	7'55"	25"
Tpo. Ejercicio 2.2	5'32"	35"
Tpo. Ejercicio 2.3	8'35"	25"
Resultado Ejercicio 2.1	75%	100%
Resultado Ejercicio 2.2	75%	100%
Resultado Ejercicio 2.3	75%	100%

Tabla 10. Tabla resumen con las medias de los tiempos dedicados y porcentajes de acierto en la realización de los ejercicios de la validación.

El segundo factor que afecta en la baja cantidad de datos almacenados utilizando el método tradicional es la falta de columnas para albergar esos datos en el fichero Excel. La mayoría de los datos que aparecen en la historia clínica no tienen cabida en el fichero Excel. Hemos detectado con el experimento que muchos datos relevantes como fechas, pruebas diagnósticas realizadas (como mamografías o BAGs), o pruebas pertenecientes a estudios de extensión (como TAC, mamografías o analíticas), entre otros no tienen cabida en el fichero Excel. Esto provoca una acusada falta de precisión de los datos, ya que hay mucha información que se queda sin representación en el Excel y con la que no podremos

contar a la hora de realizar los análisis. Obviamente, este factor afecta considerablemente a la precisión del método, haciendo que únicamente podamos trabajar con los datos disponibles en las columnas y complicando la vida del investigador, que tendrá que volver a acceder a las historias clínicas si necesita alguno de los datos que no tuvo en cuenta en su día para poder completar alguna investigación.

En términos generales, detectamos que utilizando la metodología tradicional se han introducido una media de 26 datos de los 66 disponibles en la historia clínica proporcionada, lo que corresponde a un 40% de datos insertados. En cambio, introduciendo los datos utilizando la herramienta, se han introducido una media de 75 datos de 79 disponibles en la historia clínica, lo que quiere decir que se han introducido un 94,9% de los datos disponibles. Comparando estos porcentajes, podemos decir que la precisión de la herramienta propuesta en esta tesis es un 54,9% mayor que el método tradicional.

Por otra parte, para evaluar la precisión también tenemos que tener en cuenta la similitud entre los valores numéricos obtenidos de forma manual y utilizando la herramienta para resolver las cuestiones planteadas en la tarea 2. En este caso, si observamos los resultados obtenidos por el método tradicional observamos que uno de los sujetos tiene los resultados totalmente diferentes al resto. Esto fue provocado por un error en la manipulación manual del fichero Excel que contiene los datos de los pacientes. Sorprendentemente, al sujeto se le alteró el orden de algunas de las filas y columnas de los pacientes, provocando que los datos no coincidiesen con el fichero original proporcionado y, por tanto, los resultados del análisis fueron incorrectos. Esto nos hace ver que el método tradicional no es tan preciso como pensábamos de antemano y los errores humanos en la manipulación de datos es algo bastante común. Hablando de porcentajes, podríamos decir que, en este caso, el método tradicional tiene un 75% de aciertos (dato que podría variar si se amplía la población de estudio), frente a un 100% de aciertos que hemos registrado utilizando la herramienta.

Evaluando la precisión de la herramienta en términos generales, podemos determinar que, para insertar pacientes en el sistema, la precisión de la herramienta propuesta en esta tesis es un 54,9% mayor que el método tradicional y para llevar a cabo análisis de datos de pacientes la ventaja en la precisión de la herramienta propuesta frente al método tradicional es de un 25%. De esta manera, podemos determinar que **la hipótesis de partida H_{01} no es correcta, siendo bastante más precisa la tecnología diseñada utilizando Modelos Conceptuales.**

H_{02}: La eficiencia de los procedimientos de almacenamiento, gestión y análisis de los datos de los pacientes.

De la misma manera que hemos hecho con la hipótesis H_{01}, vamos a analizar los valores recogidos durante el experimento para valorar la eficiencia de los procedimientos de almacenamiento, gestión y análisis de los datos de los pacientes utilizando cada uno de los tratamientos planteados.

Para evaluar la eficiencia, nos basaremos en los tiempos recogidos durante la ejecución del ejercicio de validación por cada uno de los sujetos.

Empezando por el tiempo dedicado a insertar los datos de la historia clínica proporcionada en cada uno de los sistemas (fila *Tpo. Inserción Paciente* de la Tabla 10), podemos detectar fácilmente que se tarda de media algo más del doble de tiempo en insertar un paciente utilizando la herramienta, que utilizando el método tradicional. Evaluando estos valores aisladamente podríamos determinar que la herramienta tradicional es más eficiente para insertar a un paciente, pero tendríamos una medida sesgada. Para poder definir la eficiencia es importante tener en cuenta el número de datos que se almacenan en cada caso. Si dividimos el tiempo medio dedicado a insertar un paciente con el método tradicional por la media de la cantidad de datos que se insertan utilizando este método, podemos determinar que se tardan unos 24 segundos en insertar un dato en el sistema. Si realizamos

el mismo cálculo para la herramienta propuesta, obtenemos que tardamos 21 segundos por dato insertado (un 12'5% menos de tiempo). De esta manera, sí que podemos determinar que la herramienta es más eficiente en la inserción de datos de pacientes que el método tradicional.

Los siguientes datos a analizar son los tiempos dedicados a la resolución de los ejercicios (podemos ver estos datos en las *filas Tpo. Ejercicio 2.1, 2.2 y 2.3* de la Tabla 10). En este caso, como hemos comentado al inicio de esta sección, tenemos varios datos que vamos a descartar del análisis (marcados en la tabla con un *) porque se han visto alterados por intentar utilizar la herramienta SPSS para resolver las cuestiones con conocimientos limitados sobre la utilización de la misma (se pueden ver en la tabla completa que se encuentra en el Anexo II). Sobre los datos restantes hemos calculado las medias de los tiempos obtenidos y podemos ver una notable reducción del tiempo empleado para resolver las tres cuestiones planteadas utilizando la herramienta propuesta, con una mejora en los tiempos de 19, 10 y 20 veces menor en la resolución de las cuestiones 1, 2 y 3 respectivamente.

Con estos resultados tan llamativos, podemos afirmar que **la hipótesis de partida H_{02} no es correcta, siendo más eficiente (en un orden del 93% de reducción de tiempo) el procedimiento de almacenamiento, gestión y análisis de los datos de los pacientes utilizando el software diseñado mediante Modelos Conceptuales.**

H_{03}: La satisfacción del usuario al almacenar, gestionar y analizar datos

La última de las variables respuesta que vamos a analizar es la satisfacción del usuario al almacenar, gestionar y analizar datos tanto utilizando la metodología tradicional, como la herramienta propuesta.

Para poder medir esta variable, hemos utilizado los cuestionarios que hemos mencionado en apartados anteriores y que se encuentran en el Anexo I en los que los sujetos indican su percepción respecto a la facilidad de uso (FUP), la utilidad (UP) y la intención de uso (IU) de cada uno de

los tratamientos planteados. La valoración de las cuestiones, como se ha comentado en secciones anteriores, oscila entre 1 y 5, siendo 1 "Totalmente en desacuerdo" y 5 "Totalmente de acuerdo".

En primer lugar, analizaremos la FUP de ambos tratamientos observando los resultados obtenidos en las preguntas 1, 4, 5 y 13 de los cuestionarios como se puede ver en la siguiente Tabla 11:

FACILIDAD DE USO PERCIBIDA

	TRADICIONAL	HERRAMIENTA
1	1, 2, 3, 4	4, 5, 5, 5
4	2, 2, 3, 4	4, 5, 5, 5
5	2, 3, 4, 5	5, 5, 5, 5
13	1, 2, 2, 2	3, 5, 5, 5

Tabla 11. Tabla con la facilidad de uso percibida por los usuarios en la validación

Hay varios datos llamativos en los resultados. En una visión general de los datos podemos observar que las puntuaciones obtenidas para el método tradicional raramente sobrepasan la puntuación 3, a diferencia de la valoración de la herramienta donde en su mayoría todas la superan, mayoritariamente con puntuación de 5.

Además, respecto al método tradicional, destacan las dos puntuaciones de 1 en las preguntas 1 y 13 indicando la disconformidad de los sujetos respecto a la facilidad de uso del método. Resaltan, sobre todo, las puntuaciones obtenidas en la pregunta 13, ya que ninguna de ellas llega al 3, lo que indica la falta de confianza en que el método tradicional sea el más rápido y sencillo con respecto a otros procedimientos. Por otro lado, cabe destacar el 5 otorgado en la pregunta 5 a la herramienta tradicional por parte de la oncóloga del grupo. Esto se debe a que ella es la que principalmente maneja los datos de los pacientes utilizando a día de hoy el método tradicional, por lo que lo considera sencillo de explicar a otra persona que no lo conociese de antemano.

Respecto a las puntuaciones recibidas por la herramienta, destaca la mayoría de puntuaciones por encima de 3, principalmente 5. Además, cabe

resaltar los resultados de la pregunta 5, donde refleja clarísimamente lo fácil que consideran los sujetos que les sería explicar el funcionamiento de la herramienta a una persona que no la conociese.

Estos resultados denotan **una alta percepción de facilidad de uso de la herramienta** por parte de los usuarios frente al método tradicional.

En segundo lugar, estudiaremos la UP de ambos tratamientos por parte de los sujetos, observando los resultados que aparecen en la siguiente Tabla 12:

UTILIDAD PERCIBIDA

	TRADICIONAL	HERRAMIENTA
2	1, 2, 2, 3	4, 5, 5, 5
3	1, 1, 3, 3	4, 5, 5, 5
6	1, 1, 3, 4	4, 5, 5, 5
7	1, 2, 2, 3	3, 4, 5, 5
8	1, 2, 2, 3	5, 5, 5, 5
10	1, 1, 2, 3	5, 5, 5, 5
11	1, 1, 2, 4	4, 5, 5, 5
12	1, 2, 2, 3	4, 5, 5, 5
14	1, 2, 2, 3	4, 5, 5, 5

Tabla 12. Tabla con utilidad percibida por los usuarios en la validación

En vistas generales, podemos observar la alta presencia de puntuaciones bajas en las cuestiones del método tradicional y de puntuaciones altas en las de la herramienta propuesta.

Centrándonos en los resultados del método tradicional, podemos observar que en todas las preguntas hay al menos una puntuación de 1, lo que indica la poca percepción de utilidad del método tradicional para los sujetos de nuestro experimento. En las nueve preguntas planteadas, únicamente han valorado con un 4 dos de ellas, la 6 y la 11, siendo el resto de puntuaciones negativas mayoritariamente.

Sin embargo, los resultados obtenidos por la herramienta son mucho más positivos. Todas las preguntas planteadas para este tratamiento han sido

valoradas con puntuación de 5 por dos o más de los sujetos, lo que indica la alta satisfacción del usuario respecto a utilidad. Además, cabe resaltar los resultados de las preguntas 8 y 10, donde todas las valoraciones han sido de 5, indicando que la satisfacción de los sujetos respecto a la precisión de los datos percibida utilizando la herramienta es máxima.

Con estos datos podemos resumir que la percepción de utilidad y precisión de los datos y resultados de los análisis utilizando la herramienta es mucho más alta que la percibida utilizando el método tradicional.

Finalmente, analizaremos las valoraciones de los sujetos sobre la IU de ambos tratamientos para gestionar y analizar los datos de sus pacientes. Los resultados obtenidos de las valoraciones se resumen en la siguiente Tabla 13:

	INTENCIÓN DE USO	
	TRADICIONAL	HERRAMIENTA
9	1, 1, 3, 3	5, 5, 5, 5
15	1, 1, 2, 3	4, 4, 5, 5
16	1, 1, 2, 3	4, 5, 5, 5

Tabla 13. Tabla con la intención de uso de ambos métodos planteados en la validación

Observando la tabla a simple vista se puede detectar que la intención de uso del método tradicional ha sido valorada de forma bastante negativa, predominantemente por 1, mientras que la intención de uso de la herramienta ha sido bastante positiva, con valoraciones de 5 principalmente.

Si nos paramos a observar los resultados en detalle, veremos que para el método tradicional ninguna de las valoraciones supera el 3, y en todas las cuestiones hay al menos dos valoraciones con un 1, indicando poco interés en la utilización de este tratamiento en un futuro.

En cambio, si visualizamos las valoraciones obtenidas en las cuestiones relacionadas con la intención de uso de la herramienta, detectamos una alta satisfacción por parte del usuario. En la pregunta 9 se han marcado con 5

las cuatro valoraciones, lo que indica que los sujetos definitivamente usarían la herramienta para almacenar y gestionar los datos de sus pacientes.

Estos resultados nos permiten afirmar que la intención de uso de la herramienta para almacenar y gestionar datos de los pacientes es mucho más alta que la del método tradicional.

Finalmente, después de haber analizado en detalle los datos obtenidos en los cuestionarios, podemos afirmar que **la hipótesis de partida H_{03} no es correcta, siendo mucho más satisfactorio el almacenamiento, gestión y análisis de los datos de los pacientes utilizando el software diseñado utilizando Esquemas Conceptuales.**

4.4.2. SESIÓN DE FOCUS-GROUP

Para completar esta validación hemos llevado a cabo una sesión de *focus-group* que nos permita información cualitativa adicional sobre las percepciones de los sujetos en relación a ambos tratamientos. Además, el *focus-group* nos permite incluir en la validación la parte de "Análisis de microARNs" la cual no era viable incluir en el experimento cuantitativo por la gran dedicación de tiempo que supone la realización de dicho análisis de forma manual. De esta manera, podemos obtener la valoración de ambos métodos (el tradicional y la herramienta) por parte de los sujetos sin necesidad de llevar a cabo un ejercicio de análisis de microARNs completo.

Los participantes en el *focus-group* fueron los mismos sujetos del experimento cuantitativo exceptuando a la ingeniera biomédica, que no pudo acudir por causas ajenas al experimento. El *focus-group* tuvo una duración de 2 horas, que fueron grabadas en audio, y transcurrió siguiendo el siguiente diseño:

- **Tarea 1**: Evaluar la parte de datos clínicos: Valorar los ejercicios realizados en las sesiones anteriores.
    - Plantear los resultados obtenidos en las encuestas y en los

ejercicios realizados en una presentación para empezar con el intercambio de impresiones.
    - Pros y contras de la metodología antigua y de la nueva para almacenar la información de un paciente y analizar los datos. Para ello utilizaremos unos post-its en una pizarra. Dejaremos unos minutos para que escriban algunos y los pongan en la pizarra.
    - Seguidamente, procederemos a comentar cada uno de ellos para dejar claras sus impresiones.

- **Tarea 2**: Evaluar la parte de los datos de microARNs: demostración
    - Hacer la demostración de un análisis de microARNs utilizando la herramienta propuesta para que los sujetos puedan valorarlo. Como la metodología tradicional la conocen de sobra, no es necesaria una demostración de la misma para poder valorarla.
    - Pros, contras, mejoras y posible potencial de la herramienta de análisis de microARNs y del método tradicional. Utilizaremos la técnica de los post-its de nuevo y el intercambio de impresiones posterior.

- **Tarea 3**: Plantear posibles mejoras y potencial de explotación del Esquema Conceptual
    - Mostrar el Esquema Conceptual del Cáncer de Mama
    - Utilizar la técnica de los post-it para definir posible potencial de explotación y mejoras. Plantear posibles extensiones o interés en investigar nuevas tecnologías.

El esquema diseñado en la pizarra para la colocación de los post-its fue el que aparece en la Figura 91 que aparece a continuación:

Almacenamiento y análisis de datos de pacientes					Análisis microARNs				
Tradicional		Herramienta			Tradicional		Herramienta		
Pros	Cons	Pros	Cons	Mejoras	Pros	Cons	Pros	Cons	Mejoras

Esquema Conceptual
Mejoras y potencial

Figura 91. Esquema dibujado en la pizarra durante la sesión de *focus-group* para pegar los post-its con los pros, contras y mejoras.

En la Figura 92 y la Figura 93 vemos el resultado de la pizarra con todos los post-its colocados sobre ella. La Figura 92 contiene las valoraciones de los dos métodos para la gestión y análisis de datos de pacientes y la Figura 93 las valoraciones de los métodos para el análisis de microARNs.

Empezamos el análisis de los resultados de este *focus-group* detallando los pros y contras del almacenamiento y análisis de datos de pacientes utilizando el método tradicional:

- PROS:
  - **Es más asequible para los investigadores**. Las aplicaciones informáticas utilizadas en el método tradicional son, principalmente, Word y Access (Microsoft) y SPSS para la realización de estudios estadísticos más complejos. Estas herramientas normalmente las tienen previamente instaladas en el ordenador, y de no ser así, las licencias de compra no son demasiado elevadas, lo que las hace accesibles a investigadores con baja financiación.

Figura 92. Pizarra con post-its de la valoración de la gestión de datos y análisis de pacientes

Figura 93. Pizarra con post-its de la valoración del análisis de microARNs

- **La interfaz de introducción de datos es sencilla.** Se trata de una sencilla hoja de cálculo utilizada de forma común por médicos y biólogos en multitud de situaciones.
- **Personalización y versatilidad.** Para los sujetos es muy sencillo poder introducir cualquier dato en una celda, aunque no sea su cometido, o poder insertar columnas si lo consideran necesario. Pueden manipular el fichero Excel a su antojo.

- CONTRAS:
  - **La posibilidad de cometer errores humanos es muy alta.** Los datos pueden modificarse o eliminarse de forma muy simple por lo que su manipulación de forma manual puede llevar a errores fácilmente.
  - **Datos menos robustos.** En cada una de las celdas del fichero Excel es posible incluir cualquier dato, lo que provoca una gran heterogeneidad de datos en las celdas de una misma columna.
  - **Categorización engorrosa (si el fichero está mal estructurado).**

El mismo problema comentado en el punto anterior lleva a una categorización bastante engorrosa, teniendo en cuenta que tienes que revisar los datos uno a uno para asegurarte de que la categorización se hace correctamente.

- **Columnas poco informativas y/o complejas.** Si el fichero no se ha diseñado previamente de forma correcta puede haber columnas redundantes, otras que tengan poca información o incluso columnas que contengan muchos datos que deberían estar en columnas separadas.
- **Es imprescindible definir un buen diseño de la tabla Excel**. Para poder almacenar los datos es esencial diseñar previamente la tabla correctamente estructurada con los datos más relevantes, tarea que no suele ser sencilla para los usuarios, provocando celdas sobre explotadas, columnas vacías y falta de información.

Continuaremos con los resultados obtenidos sobre el almacenamiento y análisis de datos de paciente utilizando la herramienta propuesta:

- PROS:
  - **Solidez de los datos introducidos.** Los datos que se introducen en el sistema son muy estables. La posibilidad de que ocurran errores humanos a la hora de introducir los datos es muy baja.
  - **Datos muy informativos y correctamente clasificados.** Buscar información sobre una determinada prueba o tratamiento se convierte en algo sencillo y hace que toda la información de la base de datos sea útil y accesible.
  - **Entorno cómodo, sencillo e intuitivo para el usuario.** La interfaz de usuario es intuitiva y sencilla de utilizar.
  - **Base de datos con mucho potencial.** Se puede utilizar la herramienta para añadir nuevos pacientes, pero también para analizar los pacientes de forma retrospectiva, pudiendo inferir nueva información a largo plazo.
  - **Análisis robusto de los datos.** Se pueden incluir en el análisis

muchas variables diferentes y hacer muchas combinaciones.
    - **Análisis rápido, seguro y fiable.** Las consultas a la base de datos y los análisis se realizan de forma ágil y segura.

- CONTRAS:
    - **Mayor inversión de tiempo inicial.** Como se ha demostrado en las pruebas cuantitativas, introducir un nuevo paciente supone una dedicación de tiempo mucho mayor, debido a que se introducen muchos más datos que con el método tradicional.
    - **Se debe invertir tiempo en el aprendizaje.** Explorar las opciones disponibles o localizar el lugar correcto para insertar los datos requiere de un tiempo inicial de aprendizaje.

- MEJORAS Y/O POTENCIAL FUTURO
    - **Incluir más parámetros en el análisis.** Los parámetros establecidos en el formulario de análisis para la implementación del prototipo de esta herramienta quedaron definidos por las cuestiones que se iban a resolver en los ejercicios de validación. Hay que extender estos parámetros para poder incluir cualquier dato de la base de datos en un futuro.
    - **Exportar a Excel ciertos campos seleccionados.** Poder realizar una selección de los datos que interesa exportar en formato Excel para su manipulación externa.
    - **Más formatos para exportar el informe de los pacientes.** Poder exportar en otros formatos de archivo, como Word, por ejemplo.

A continuación, detallamos los resultados obtenidos sobre el análisis de microARNs utilizando el método tradicional:

- PROS:
    - **Análisis muy versátil.** La utilización de herramientas como Excel y R abren una gran ventana de posibilidades ante en

análisis de los datos de microARNs.

- **Ayuda de la comunidad muy dinámica.** Al ser un método ampliamente utilizando en la comunidad científica, existen muchos foros y páginas de ayuda donde pueden ayudarte a resolver problemas que puedan ir surgiendo.
- **Salida final más gráfica.** Los paquetes estadísticos incluyen funcionalidades que permiten a los usuarios generar gráficas con sus datos.

- CONTRAS:
  - **Necesidad de conocer múltiples páginas web de análisis, herramientas y bases de datos.** La utilización de múltiples recursos web hace complicado el aprendizaje y la comprensión de todos y cada uno de ellos.
  - **Tarda mucho más tiempo.** El proceso de análisis de los datos es muy largo y tedioso debido, entre otras cosas, a la utilización de múltiples herramientas y recursos, pudiéndose dedicar a ello una mañana completa (entre 4 o 5 horas).
  - **Necesidad de conocimientos bioinformáticos.** La necesidad de usar paquetes estadísticos implementados en lenguajes utilizados por los profesionales de la bioinformática (como R) hace imprescindible el aprendizaje de estas tecnologías para poder llevar a cabo el análisis.

En los siguientes puntos se detallan los resultados obtenidos sobre el análisis de datos de microARNs utilizando la herramienta propuesta:

- PROS:
  - **Posibilidad de personalizar las búsquedas.** Se pueden incluir distintos parámetros en el análisis y realizar varios análisis cambiando los parámetros de forma sencilla, permitiendo un análisis más completo y versatil.
  - **Exportar datos fácilmente.** Con un simple botón se pueden exportar los datos del análisis realizado a un fichero Excel.

- **Intuitivo, fácil y rápido de utilizar.** Los análisis se realizan de forma rápida, sencilla y cómoda para el investigador.
- **No se necesitan conocimientos previos especializados.** No es necesario tener conocimientos bioinformáticos o estadísticos para llevar a cabo el análisis utilizando la herramienta.
- **Integración en una única plataforma de la información necesaria para la interpretación de los resultados.** Disponer de toda la información integrada y relacionada con los datos del análisis facilita notablemente el trabajo de los investigadores.

- CONTRAS:
  - **Reducida versatilidad de los análisis estadísticos.** Los estudios estadísticos que se llevan a cabo están preestablecidos por código. Es posible alterar los parámetros del análisis, pero no el tipo de análisis a realizar.
  - **Necesita conversión de datos crudos desde la plataforma genómica.** Cada plataforma genómica tiene un formato distinto en su fichero de salida de datos que es necesario preprocesar antes de introducir en la herramienta.

- MEJORAS Y/O POTENCIAL FUTURO
  - **Ampliar los parámetros de análisis.** Incluir más parámetros configurables en el formulario de realización del análisis, entre ellos, datos clínicos que permitan filtrar las muestras. El prototipo actual únicamente filtra por umbral y por rango de edad.
  - **Actualizar las bases de datos automáticamente.** Disponer de un sistema de carga automático de datos de forma periódica desde las fuentes para tener los datos constantemente actualizados.
  - **Ampliar los formatos de entrada de datos.** Permitir la inserción de los datos directamente desde los ficheros de salida de las plataformas genómicas sin preprocesado previo.

Para finalizar, los últimos resultados obtenidos de este *focus-group* han sido las mejoras y/o potencial de explotación del Esquema Conceptual diseñado en esta tesis.

En primer lugar, como ya se ha comentado en el apartado de mejoras de la herramienta, ampliar los parámetros tanto de análisis de datos clínicos como de análisis de microARNs supone una mayor explotación del potencial del Esquema Conceptual diseñado en esta tesis, permitiendo hacer consultas complejas sobre todos los datos contenidos en la base de datos e incluir cualquier dato clínico para seleccionar los grupos de pacientes a comparar en el análisis de microARNs.

Además, las asistentes han comentado posibles futuras extensiones del modelo conceptual, empezando por incluir la extensión de la vista se Secuenciación Masiva que se ha diseñado en esta tesis para incluir estudios de genomas completos secuenciados utilizando técnicas de secuenciación masiva. Comentan los sujetos que esta es la ampliación del esquema conceptual que más se ajusta a la realidad actual ya que, aunque todavía se desconoce el significado de la mayoría de variaciones del genoma humano, la tendencia a secuenciar genomas o exomas completos va en aumento.

Por otro lado, también comentan que sería útil incluir entre las técnicas utilizadas para medir la expresión de los microARNs la técnica de RNASeq [131]. Esta novedosa técnica, que está empezándose a utilizar en muchos laboratorios, funciona secuenciando el ARN de la muestra del paciente y contando las copias de cada ARN.

Finalmente, se proponen otras vistas interesantes a tener en cuenta en futuras ampliaciones del modelo, entre ellas una vista de análisis de proteínas, de estudios de metilación o la generalización de las pruebas de estudios de anticuerpos por inmunohistoquímica, permitiendo incluir cualquier anticuerpo y el valor obtenido. Las posibilidades de ampliación del modelo son muy grandes, y con ellas su potencial de explotación para gestionar datos sobre el cáncer de mama.

5. Conclusiones

Esta tesis doctoral se centra en el planteamiento innovador de la creación de Sistemas de Información basados en Modelos Conceptuales en un entorno clínico y bioinformático donde las soluciones existentes carecen de estas bases cuya efectividad ha sido demostrada en múltiples dominios. En nuestro caso en particular, centramos el trabajo en el dominio del Cáncer de Mama. La contribución general de este trabajo es el planteamiento de una base metodológica para la construcción de sistemas de información en el ámbito sanitario y de investigación médica. Como se ha demostrado en la validación, el sistema de información propuesto mejora considerablemente la gestión de datos clínicos y biológicos de los pacientes, la realización de análisis y estudios con los mismos y la obtención de resultados.

Este trabajo se inició haciendo un **profundo estudio del dominio**, donde se analizó, primeramente, el procedimiento de diagnóstico y tratamiento que se lleva a cabo en los hospitales para poder identificar la información que maneja, cómo clasifican sus datos, las fuentes de datos clínicas o biológicas donde consultan información o las herramientas que utilizan para almacenar y gestionar la información que consideran relevante. Tras este primer estudio salieron a la luz los procedimientos de manejo y análisis de datos que se llevan a cabo en los hospitales, y se pudieron determinar las carencias a las que se enfrentan los profesionales médicos y biólogos

respecto a la gestión de sus datos, lo que nos sirvió como captura de requisitos para la realización de este sistema de información.

Además, se estudiaron las principales fuentes de información donde se consultan los datos para la realización de diagnósticos genómicos y estudios de expresión génica, resaltando la dificultad de manejo de estas fuentes, principalmente provocada por la falta de una estandarización de la información que facilite las interrelaciones de los datos. De este estudio se detectó la necesidad de un sistema integrador de la información procedente de las distintas fuentes de datos que permita gestionar la información de forma conjunta y estructurada.

Como último apartado del estado del arte, se estudiaron distintas tecnologías existentes para la gestión de datos en entornos clínicos y genómicos. Se detectó la falta de sistemas de información estructurados utilizando modelos conceptuales, así como la carencia de sistemas integradores de información clínica y biológica que faciliten el estudio de enfermedades complejas como el Cáncer de Mama.

Tras la detección de las carencias en gestión de datos existentes en el dominio de trabajo, se inició el diseño de la solución que cumpla con los requisitos definidos. Para ello, y como núcleo central y principal contribución de este trabajo de tesis doctoral, se ha **diseñado un Esquema Conceptual del Cáncer de Mama** que representa los datos clínicos sobre diagnóstico y tratamiento de los pacientes con cáncer de mama y los datos genómicos relacionados procedentes de los estudios llevados a cabo en los laboratorios de biología molecular. Incluso relaciona toda esta información con la información procedente de bases de datos públicas, permitiendo la integración de la información pública y privada en un único sistema de información.

Además, el esquema conceptual diseñado se ha utilizado de base para el **diseño de arquetipos de historia clínica electrónica siguiendo el estándar ISO 13606**. Estos arquetipos diseñados a partir del esquema conceptual contienen la información relativa a los distintos episodios clínicos de un

paciente en diagnóstico o tratamiento del cáncer de mama, incluyendo estudios genómicos o de expresión génica. La utilización de modelos conceptuales como base para el diseño de arquetipos de historia clínica electrónica es otra de las contribuciones de este trabajo de tesis. Este planteamiento supone una simbiosis en la que se permite una óptima gestión de los datos incluidos en los informes clínicos gracias al sistema de información correctamente estructurado al que hacen referencia y una interoperabilidad semántica entre sistemas de información clínica que utilicen el mismo estándar de información.

Por otro lado, y también tomando como base el esquema conceptual diseñado, **se ha desarrollado una base de datos que integre la información clínica, biológica y genómica** representada en el esquema. En ella, se ha integrado información sobre el cáncer de mama a partir de diversas fuentes y orígenes que se encuentra totalmente relacionada entre sí. El planteamiento de esta solución de almacenamiento de datos clínicos, genómicos y biológicos integrada y correctamente estructurada es otra de las contribuciones de esta tesis.

Para poder realizar una prueba de concepto de la solución planteada en esta tesis, se ha **implementado un prototipo de herramienta de gestión de datos sobre el cáncer de mama**, basándonos en el Esquema Conceptual diseñado y que permita almacenar y explotar la información incluida en la base de datos desarrollada. Gracias a este prototipo, incluido como una de las contribuciones de este trabajo, los datos pueden ser gestionados y analizados de manera sencilla, rápida, segura y robusta por el personal clínico e investigador. La utilización de esquemas conceptuales hace que el sistema de información sea flexible y actualizable a demanda de los investigadores. Actualmente, el prototipo está pensado para el análisis de datos clínicos y de expresión génica de microARNs, está actualmente preparado para incorporar expresión génica global y sería extensible de forma sencilla y rápida a nuevos datos genómicos procedentes de estudios de secuenciación masiva y metilación.

Finalmente, se ha llevado a cabo la **validación del prototipo** que integra el trabajo realizado en esta tesis y nos permite evaluar si la solución planteada en este trabajo de tesis ofrece algún beneficio a la comunidad médica y científica. En este caso, la herramienta ha generado impresiones positivas en los investigadores que van a manejar este tipo de datos, tanto médicos como biólogos moleculares. La buena acogida de la herramienta ha permitido que actualmente se esté utilizando en el laboratorio de biología molecular del INCLIVA, donde se realizó la validación de la misma, como soporte a la realización de sus estudios de expresión génica de microARNs.

5.1. Trabajo futuro

El trabajo llevado a cabo en esta tesis es un primer paso que demuestra la eficiencia de los sistemas de información basados en modelos conceptuales en la práctica clínica, en concreto en el estudio, diagnóstico y tratamiento de una enfermedad como es el cáncer de mama. Sin embargo, queda mucho trabajo por delante para conseguir que los profesionales médicos y biólogos trabajen con este sistema de información desde las consultas médicas hasta los laboratorios.

En el apartado 4.4.2 de esta tesis, ya se han comentado algunas mejoras de la herramienta propuestas por los sujetos de la validación. Está claro que la herramienta es un prototipo que permite demostrar la validez de estos sistemas en el dominio estudiado, pero es necesario crear una herramienta que permita realizar análisis más precisos, ampliando los parámetros de búsqueda y análisis para conseguir una mayor explotación del potencial del Esquema Conceptual diseñado en esta tesis. Además, se ha detectado la necesidad de ampliar los formatos de exportación de la información, así como los de entrada de los datos, y un sistema de actualización de la base de datos a medida que salgan nuevas versiones en las fuentes de datos origen. Estas ampliaciones ofrecerían a los usuarios mayor versatilidad y utilidad, lo que convertiría la herramienta en un sistema más maduro, dejando de ser un prototipo de prueba de concepto.

Además, los sujetos han propuesto posibles futuras extensiones del modelo conceptual, empezando por incluir en la herramienta la gestión de datos de secuenciación masiva, que ya se ha tenido en cuenta en el esquema conceptual. Los expertos también han sugerido la ampliación del esquema conceptual con una vista de estudios de metilación, trabajos que piensan abordar a corto plazo y de los que esperan extraer grandes cantidades de datos para analizar.

Por otro lado, consideramos que un sistema de información maduro para su aplicación en la práctica clínica debería disponer de un apartado de gestión de usuarios que permita controlar los accesos a información tan

delicada como la clínica y la genómica de los pacientes. Además, habría que incluir también sistemas de encriptación de datos que mejorasen la seguridad de los mismos y que nos permitiesen pasar de manejar datos anónimos a manejar datos identificativos de los pacientes con total seguridad.

La migración de la herramienta a una tecnología de programación más avanzada, también sería un paso a tener en cuenta. Access es una tecnología muy sencilla y cómoda para la generación de formularios a modo de prototipo, pero para una herramienta final supone muchas limitaciones. Una opción bastante actual sería la utilización de una arquitectura web basada en la creación de APIs (*Application Programming Interface*) en el servidor para facilitar la interoperabilidad y el uso de REST para facilitar el acceso de las aplicaciones cliente. En la implementación de los servicios del lado servidor podría utilizarse Java o node.js y en el lado cliente tecnologías web (HTML5, CSS3 y frameworks o librerías Javascript como Angular, React o Ionic) que posibilitan el uso de aplicaciones multiplataforma para que sea accesible desde Windows, IOS y Android.

La inclusión de información de nuevas fuentes a nuestra base de datos sería una de las ampliaciones a tratar. Actualmente la base de datos se ha cargado con información de algunas de las bases de datos más relevantes, con información tanto general, como específica de la enfermedad o de expresión génica, como muestra de la posibilidad de integrar información diversa en un único repositorio y poder determinar las ventajas que esto supone ante la dispersión y heterogeneidad de datos ante la que nos encontramos actualmente. La ampliación de esta base de datos con información de más fuentes de datos públicas mejoraría la calidad de la información contenida, ampliando la cantidad de datos interrelacionados y ofreciendo la posibilidad a los profesionales que trabajan con esta información de tener toda la información que necesitan relacionada entre sí en una misma fuente de datos, facilitándoles la tarea de consulta de información sobre sus estudios.

Finalmente, es interesante resaltar que se están aplicando los resultados de esta tesis doctoral en dos proyectos financiados por el Gobierno de Paraguay en los que participa el Centro de Investigación PROS, titulados "Proyecto multicéntrico de formación multidisciplinar en cáncer y aplicación de la Historia Clínica Electrónica (HCE) con el fin de integrar los datos clínico-moleculares y orientar la estrategia terapéutica" (PINV15-149) y "Proyecto multicéntrico de determinación del perfil mutacional de pacientes con cáncer de tumores sólidos para guiar la estrategia terapéutica hacia una medicina personalizada" (PINV15-156). Estos proyectos tienen como objetivo implantar la Historia Clínica Electrónica en las unidades de Oncología de dos hospitales de Asunción utilizando el estándar ISO13606 y esquemas conceptuales para recopilar y gestionar la información relacionada con el diagnóstico y tratamiento de los tumores sólidos y analizarla estadísticamente para intentar obtener conclusiones relevantes que mejoren el tratamiento de la enfermedad. Teniendo en cuenta que el cáncer de mama es un tipo de tumor sólido y que en esta tesis se demuestra la mejora de la calidad y la eficiencia en la gestión y análisis de datos clínicos y genómicos utilizando modelos conceptuales, se va a aplicar la metodología propuesta en esta tesis en estos nuevos proyectos que han empezado en abril de este año 2017.

5.2. Aportación a la comunidad científica

5.2.1. Publicaciones

Internacionales

Pastor, M., Burriel, V. and Pastor, O. "Conceptual Modeling of Human Genome Mutations: A Dichotomy Between What we Have and What we Should Have." BIOSTEC Bioinformatics (BIOINFORMATICS), Valencia, Spain, (2010).

Burriel Coll, V. "Design and development of a genomic information system to manage breast cancer data." Sixth International Conference on Research Challenges in Information Science (RCIS), IEEE. Valencia, Spain (2012).

Bosca, D., Marco, L., Burriel, V., Jaijo, T., Millán, J.M., Levin, A.M., Pastor, O., Robles, M. and Maldonado, J.A. "Genetic testing information standardization in HL7 CDA and ISO13606." 14th World Congress on Medical and Health Informatics (MedInfo) Copenhagen, Denmark (2013).

Burriel, V., Pastor, M.Á., Celma, M., Casamayor, J.C. and Mota, L. "Design and Development of an Information System to Manage Clinical Data about Usher Syndrome Based on Conceptual Modeling." The Fifth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO) Lisbon, Portugal (2013). *Best Paper Award.*

Burriel, V. and Pastor, O. "Conceptual Schema of Breast Cancer: The background to design an efficient information system to manage data from diagnosis and treatment of breast cancer patients." IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), IEEE. Valencia, Spain (2014).

Lichtnow, D., Alves, R., Pastor, O., Burriel, V. and Moreira de Oliveira, J.P. "BION2SEL: An Ontology-Based Approach for the Selection of Molecular Biology Databases." Brazilian Symposium on Bioinformatics (BSB), Advances in Bioinformatics and Computational Biology, Springer: 83-90. Belo Horizonte, Brazil (2014).

Burriel, V., Pastor, O., Peña-Chilet, M., Martínez, M.T. and Ribas, G. "Conceptual Schema of miRNA's Expression." IEEE Tenth International Conference on Research Challenges in Information Science (RCIS). Grenoble, France (2016).

León, A., Reyes, J., Burriel, V. and Valverde, F. "Data Quality Problems When Integrating Genomic Information." International Conference on Conceptual Modeling (ER), Springer International Publishing. Gifu, Japan (2016).

Nacionales

Marco Ruiz, L., Boscá Tomás, D., Burriel Coll, V., Jaijo, T., Millán, J.M., Levin, A., Pastor López, O., Robles Viejo, M., Maldonado Segura, J.A. "Estandarización de información genética en la HCE según CDA. Síndrome de Usher" XXX Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB) San Sebastián (2012).

Burriel, V. "Diseño del Esquema Conceptual del Cáncer de Mama que sirva de base para un Sistema de Información eficiente que gestione datos clínicos y biológicos sobre el Cáncer de Mama." I Congreso Biomedicina Predocs Valencia. Valencia (2014).

Burriel, V. "Diseño y desarrollo de un Sistema de Información para la gestión de catos sobre Cáncer de Mama" VII Reunión del Foro de Interoperabilidad en Salud. Salamanca (2017)

5.2.2. Participación en proyectos de investigación

"Integración de un motor de búsqueda tipo BLAST a una aplicación de análisis de variaciones de secuencias genómicas" Financiado por la Universidad Politécnica de Valencia. Fecha de inicio: 15/12/2009. Duración: 12 meses. Cuantía total: 12.000 €

"Incorporación de la información genética a la Historia clínica electrónica (HCE) mediante Modelos Conceptuales" Financiado por la Universidad Politécnica de Valencia. Fecha de inicio: 01/12/2011. Duración: 12 meses. Cuantía total: 6.000€

"FUTURE CLINIC – Preparación del escenario de la Medicina Genómica Personalizada" Financiado por Indra y GEM BioSoft. Fecha de inicio: 01/03/2011. Duración: 22 meses. Cuantía total: 21.000 €

"SINB: Diseño e implementación de un Sistema de Información para la gestión de datos clínicos, biológicos y evolutivos sobre el Neuroblastoma" Financiado por la Universidad Politécnica de Valencia y el IIS La Fe. Fecha de inicio: 19/12/2012. Duración: 6 meses. Cuantía total: 3.000€

"TROMBORISK- Modelo predictivo del riesgo de trombosis venosa combinando perfiles polimórficos y efectos funcionales en el sistema hemostático." Financiado por la Universidad Politécnica de Valencia y el IIS La Fe. Fecha de inicio: 19/12/2012. Duración: 6 meses. Cuantía total: 3.000€

"Diseño y Desarrollo de un Sistema de Información para la Gestión Eficiente de los Datos Referentes al Cáncer de Mama en Mujeres Jóvenes." (PI13/02247) Financiado por el Instituto de Salud Carlos III. Fecha de inicio: 01/01/2014. Duración: 3 años. Cuantía total: 15.125 €

"DATAME - Desarrollo de aplicaciones Big Data" Financiado por la Universidad Politécnica de Valencia. Fecha de inicio: 01/09/2016. Duración 1 año. Cuantía total: 18.800 €

"IDEO - Innovative services for Digital Enterprises with ORCA (Servicios Innovadores para Empresas Digitales con ORCA)" (PROMETEO/2014/039) Financiado por la Generalitat Valenciana. Fecha de inicio: 01/06/2014. Duración 3 años y 6 meses. Cuantía total: 143.180 €

"DATAME - Un Método de producción de software dirigido por modelos para el desarrollo de aplicaciones Big Data" (TIN2016-80811-P) Financiado por el Ministerio de Economía y Competitividad del gobierno de España. Fecha de inicio: 20/12/2016 Duración: 4 años. Cuantía total: 99.099 €

"Proyecto multicéntrico de formación multidisciplinar en cáncer y aplicación de la Historia Clínica Electrónica (HCE) con el fin de integrar los datos clínico-moleculares y orientar la estrategia terapéutica" Financiado por CONACYT (Paraguay) y el Gobierno Nacional de Paraguay. Fecha de inicio: 01/04/2017. Duración: 15 meses.

"Proyecto multicéntrico de determinación del perfil mutacional de pacientes con cáncer de tumores sólidos para guiar la estrategia terapéutica hacia una medicina personalizada" Financiado por CONACYT (Paraguay) y el Gobierno Nacional de Paraguay. Fecha de inicio: 01/04/2017. Duración: 15 meses.

5.2.3. Organización de congresos

"Jornadas de Tecnologías para la Salud-INDRA". Jornada de carácter anual con ponentes nacionales e internacionales. 4 Ediciones (del 2009 al 2013). Miembro del equipo organizador.

"25th International Conference on Advanced Information Systems Engineering (CAISE 2013)" del 17 al 21 de junio de 2013 en Valencia, España. Miembro del equipo organizador.

"IEEE International Conference on Biomedical and Health Informatics (BHI 2014)" del 1 al 4 de junio de 2014 en Valencia, España. *Session Chair*.

"8th IFIP WG 8.1 working conference on the Practice of Enterprise Modelling (POEM 2015)" del 10 al 12 de noviembre de 2015 en Valencia, España. Miembro del equipo organizador.

6. Referencias bibliográficas

[1] B. W. Stewart and C. P. Wild, "World Cancer Report 2014," International Agency for Research on Cancer. World Health Organization 978-92-832-0443-5, 2014.

[2] J. Ferlay, *et al.*, "GLOBOCAN 2012 v1.0," International Agency for Research on Cancer, Lyon, France 2013.

[3] "Las cifras del cáncer en España 2014," Sociedad Española de Oncología Médica (SEOM)2014.

[4] F. S. Collins and H. Varmus "A New Initiative on Precision Medicine," *New England Journal of Medicine,* vol. 372, pp. 793-795, 2015.

[5] P. Economopoulou, *et al.*, "Beyond BRCA: New hereditary breast cancer susceptibility genes," *Cancer Treatment Reviews,* vol. 41, pp. 1-8, 2015.

[6] P. D. Pharoah, *et al.*, "Polygenic susceptibility to breast cancer and implications for prevention," *Nat Genet,* vol. 31, pp. 33-6, May 2002.

[7] D. F. Easton, *et al.*, "Genome-wide association study identifies novel breast cancer susceptibility loci," *Nature,* vol. 447, pp. 1087-93, Jun 28 2007.

[8] *Genética del cáncer de mama de peor pronóstico*. Available: https://ayudacancer.wordpress.com/2007/04/24/genetica-del-cancer-de-mama-de-peor-pronostico/

[9] R. Wieringa, "Design science methodology: principles and practice," in *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 2*, 2010, pp. 493-494.

[10] R. H. Von Alan, *et al.*, "Design science in information systems research," *MIS quarterly,* vol. 28, pp. 75-105, 2004.

[11] R. Wieringa, "Empirical research methods for technology validation: Scaling up to practice," *Journal of systems and software,* vol. 95, pp. 19-31, 2014.

[12] R. Wieringa and A. Moralı, "Technical action research as a validation method in information systems design science," in *International Conference on Design Science Research in Information Systems*, 2012, pp. 220-238.

[13] S. R. Lakhani, *et al.*, *WHO Classification of Tumours of the Breast, Fourth Edition*, 4th ed. vol. 4: International Agency for Research on Cancer.

[14] J. Asad, *et al.*, "Does oncotype DX recurrence score affect the management of patients with early-stage breast cancer?," *Am J Surg,* vol. 196, pp. 527-9, Oct 2008.

[15] M. Buyse, *et al.*, "Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer," *J Natl Cancer Inst,* vol. 98, pp. 1183-92, Sep 6 2006.

[16] P. Villanova, *et al.*, "Orion Clinic. Un sistema de información orientado a transformar el uso de la información en la práctica clínica, administrativa y asistencial de los hospitales," *IS Informática y Salud,* vol. 85, pp. 18-26, 2011.

[17] *Proyecto Orion-Clinic de la Agencia Valenciana de la Salud.* Available: http://www.everis.com/spain/WCLibraryRepository/References/Proyecto%20Orion-Clinic%20de%20la%20Agencia%20Valenciana%20de%20la%20Salud.pdf

[18] I. N. M. Day, "dbSNP in the detail and copy number complexities," *Human Mutation,* vol. 31, pp. 2-4, 2010.

[19] S. Sherry, *et al.*, "dbSNP: the NCBI database of genetic variation," *Nucleic acids research,* vol. 29, p. 308, 2001.

[20] NCBI. *dbSNP Handbook*. Available: http://www.ncbi.nlm.nih.gov/books/NBK21088/

[21] D. A. Benson, *et al.*, "GenBank," *Nucleic Acids Res,* vol. 41, pp. D36-42, Jan 2013.

[22] P. Flicek, *et al.*, "Ensembl 2011," *Nucleic acids research,* vol. 39, p. D800, 2011.

[23] T. Hubbard, *et al.*, "The Ensembl genome database project," *Nucleic acids research,* vol. 30, p. 38, 2002.

[24] E. Birney, *et al.*, "An overview of Ensembl," *Genome research,* vol. 14, p. 925, 2004.

[25] D. Smedley, *et al.*, "BioMart–biological queries made easy," *BMC genomics,* vol. 10, p. 1, 2009.

[26] D. N. Cooper, *et al.*, "The human gene mutation database," *Nucleic acids research,* vol. 26, pp. 285-287, 1998.

[27] P. Stenson, *et al.*, "The human gene mutation database: 2008 update," *Genome medicine,* vol. 1, p. 13, 2009.

[28] M. Krawczak, *et al.*, "Human gene mutation database—a biomedical information and research resource," *Human Mutation,* vol. 15, pp. 45-51, 2000.

[29] M. Magrane and U. Consortium, "UniProt Knowledgebase: a hub of integrated protein data," *Database,* vol. 2011, 2011.

[30] B. E. Suzek, *et al.*, "UniRef: comprehensive and non-redundant UniProt reference clusters," *Bioinformatics,* vol. 23, pp. 1282-1288, 2007.

[31] C. Wu, *et al.*, "The Universal Protein Resource (UniProt): an expanding universe of protein information," *Nucleic acids research,* vol. 34, p. D187, 2006.

[32] C. E. Cook, *et al.*, "The european bioinformatics institute in 2016: Data growth and integration," *Nucleic acids research,* vol. 44, pp. D20-D26, 2016.

[33] P. Artimo, *et al.*, "ExPASy: SIB bioinformatics resource portal," *Nucleic acids research,* vol. 40, pp. W597-W603, 2012.

[34] W. C. Barker, *et al.*, "The protein information resource (PIR)," *Nucleic acids research,* vol. 28, pp. 41-44, 2000.

[35] A. D. Johnson and C. J. O'Donnell, "An open access database of genome-wide association results," *BMC medical genetics,* vol. 10, p. 1, 2009.

[36] S. A. Forbes, *et al.*, "COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer," *Nucleic acids research,* vol. 38, pp. D652-D657, 2010.

[37] S. A. Forbes, *et al.*, "COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer," *Nucleic acids research,* vol. 39, p. D945, 2011.

[38] R. Shepherd, *et al.*, "Data mining using the catalogue of somatic mutations in cancer BioMart," *Database,* vol. 2011, 2011.

[39] L. A. Hindorff, *et al.*, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proceedings of the National Academy of Sciences,* vol. 106, pp. 9362-9367, 2009.

[40] D. Welter, *et al.*, "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations," *Nucleic acids research,* vol. 42, pp. D1001-D1006, 2014.

[41] J. S. Amberger, *et al.*, "OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders," *Nucleic acids research,* vol. 43, pp. D789-D798, 2015.

[42] A. Hamosh, *et al.*, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic acids research,* vol. 33, p. D514, 2005.

[43] M. Kanehisa, *et al.*, "KEGG for representation and analysis of molecular networks involving diseases and drugs," *Nucleic Acids Res,* vol. 38, pp. D355-60, Jan 2010.

[44] A. Kozomara and S. Griffiths-Jones, "miRBase: integrating microRNA annotation and deep-sequencing data," *Nucleic acids research,* vol. 39, pp. D152-D157, 2011.

[45] A. Kozomara and S. Griffiths-Jones, "miRBase: annotating high confidence microRNAs using deep sequencing data," *Nucleic acids research,* vol. 42, pp. D68-D73, 2014.

[46] S. Griffiths-Jones, *et al.*, "miRBase: tools for microRNA genomics," *Nucleic acids research,* vol. 36, pp. D154-D158, 2008.

[47] M. D. Paraskevopoulou, *et al.*, "DIANA-microT web server v5. 0: service integration into miRNA functional analysis workflows," *Nucleic acids research,* vol. 41, pp. W169-W173, 2013.

[48] I. S. Vlachos, *et al.*, "DIANA-TarBase v7. 0: indexing more than half a million experimentally supported miRNA: mRNA interactions," *Nucleic acids research,* vol. 43, pp. D153-D159, 2015.

[49] M. D. Paraskevopoulou, *et al.*, "DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs," *Nucleic acids research,* vol. 41, pp. D239-D245, 2013.

[50] I. S. Vlachos, *et al.*, "DIANA-miRPath v3. 0: deciphering microRNA function with experimental support," *Nucleic acids research,* vol. 43, pp. W460-W466, 2015.

[51] V. Agarwal, *et al.*, "Predicting effective microRNA target sites in mammalian mRNAs," *Elife,* vol. 4, 2015.

[52] H. R. Chiang, *et al.*, "Mammalian microRNAs: experimental evaluation of novel and previously annotated genes," *Genes & development,* vol. 24, pp. 992-1009, 2010.

[53] E. Mosca, *et al.*, "A multilevel data integration resource for breast cancer study," *BMC Systems Biology,* vol. 4, p. 1, 2010.

[54] M. Ashburner, *et al.*, "Gene Ontology: tool for the unification of biology," *Nature Genetics,* vol. 25, pp. 25-29, 2000.

[55] B. Rhead, *et al.*, "The UCSC genome browser database: update 2010," *Nucleic acids research,* 2009.

[56] D. Shen and J. V. Vadgama, "BRCA1 and BRCA2 gene mutation analysis: visit to the Breast Cancer Information Core (BIC)," *Oncology research,* vol. 11, p. 63, 1999.

[57] C. Szabo, *et al.*, "The breast cancer information core: database design, structure, and scope," *Human Mutation,* vol. 16, pp. 123-131, 2000.

[58] R. A. Baasiri, *et al.*, "The breast cancer gene database: a collaborative information resource," *Oncogene,* vol. 18, 1999.

[59] D. L. Steffen, *et al.*, "Digital reviews in molecular biology: approaches to structured digital publication," *Bioinformatics,* vol. 16, pp. 639-649, 2000.

[60] ISO/CEN, "ISO 13606-1:2008 Health informatics -- Electronic health record communication -- Part 1: Reference model," ed, 2008, p. 83.

[61] ISO/CEN, "ISO 13606-2:2008 Health informatics -- Electronic health record communication -- Part 2: Archetype interchange specification," ed, 2008, p. 124.

[62] ISO/CEN, "ISO 13606-3:2009 Health informatics -- Electronic health record communication -- Part 3: Reference archetypes and term lists," ed, 2009, p. 46.

[63] ISO/CEN, "ISO/TS 13606-4:2009 Health informatics -- Electronic health record communication -- Part 4: Security," ed, 2009, p. 30.

[64] ISO/CEN, "ISO 13606-5:2010 Health informatics -- Electronic health record communication -- Part 5: Interface specification," ed, 2010, p. 15.

[65] *Clinical Information Model Manager (CIMM).* Available: http://www.en13606.org/component/content/article/1-latest-news/94-new-archetype-repository-announcement

[66] *LinkEHR.* Available: http://www.linkehr.com/

[67] M. D. Maldonado J.A., Boscá D., Fernández-Breis J.T., Angulo C., Robles M.,, "LinkEHR-Ed: A multi-reference model archetype editor based on formal semantics," *Int J Med Inform,* vol. 78, pp. 559 - 570, 2009.

[68] *Recursos de Modelado Clínico (arquetipos).* Available: http://www.msssi.gob.es/profesionales/hcdsns/areaRecursosSem/Rec_mod_clinico_arquetipos.htm

[69] A. Muñoz Carrero*, et al.*, "Manual práctico de interoperabilidad semántica para entornos sanitarios basada en arquetipos," *Unidad de investigación en Telemedicina y e-Salud. Instituto de Salud Carlos III-Ministerio de Economía y Competitividad,* 2013.

[70] L. Zheng*, et al.*, "A clinical omics database integrating epidemiology, clinical, and omics data for colorectal cancer translational research," in *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)*, 2011, pp. 2031-2035.

[71] O. Dancea*, et al.*, "Information System for Multimodal Parameter Analysis Applied in Early Detection of Prostate Cancer," in *2006 IEEE International Conference on Automation, Quality and Testing, Robotics*, 2006.

[72] M. L. Gjerstorff, "The Danish cancer registry," *Scandinavian journal of public health,* vol. 39, pp. 42-45, 2011.

[73] *Línea de Servicios Atención Especializada: Sistema de información para la gestión de la atención especializada.* Available: http://www.everis.com/spain/WCLibraryRepository/caso_exito_orion.pdf

[74] G. Gómez Soriano. *Soluciones de eSalud en la Comunidad Valenciana.* Available: http://www.socinfo.info/seminarios/sanidad4/valencia.pdf

[75] J. Blanquer-Gregori*, et al.*, "Informatización de la atención primaria. Experiencia de implantación en un área de salud mediante la metodología de gestión por procesos," *Atención Primaria,* vol. 37, pp. 360-361, 2006.

[76] R. Baeza-Yates, "Big Data or Right Data?," in *25th International Conference on Advanced Information Systems Engineering CAiSE 2013*, Valencia, 2013.

[77] D. Kalra, "Electronic health record standards," in *IMIA Yearbook of Medical Informatics*, I. M. I. A. a. Schattauer, Ed., ed Stuttgart, Germany, 2006, pp. 136-144.

[78] S. R. Muñoz A., Pascual M., Fragua J.A., González M.A., Monteagudo J.L., Salvador C.H.,, "Proof-of-concept Design and Development of an EN13606-based Electronic Health Care Record Service," *JAMIA,* vol. 14, pp. 118-129, 2007.

[79] R. Lenz*, et al.*, "Semantic integration in healthcare networks," *International journal of medical informatics,* vol. 76, pp. 201-207, 2007.

[80] V. Burriel*, et al.*, "Design and Development of an Information System to Manage Clinical Data about Usher Syndrome Based on Conceptual Modeling," in *BIOTECHNO 2013, The Fifth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies*, 2013, pp. 30-35.

[81] K. V. Voelkerding*, et al.*, "Next-generation sequencing: from basic research to diagnostics," *Clinical chemistry,* vol. 55, pp. 641-658, 2009.

[82] M. L. Metzker, "Sequencing technologies—the next generation," *Nature Reviews Genetics,* vol. 11, pp. 31-46, 2009.

[83] O. Pastor*, et al.*, "A Conceptual Modeling Approach to Improve Human Genome Understanding," in *Handbook of Conceptual Modeling*, ed: Springer, 2011, pp. 517-541.

[84] O. Pastor, "Conceptual Modeling Meets the Human Genome," *Conceptual Modeling-ER 2008,* pp. 1-11, 2008.

[85] O. Pastor*, et al.*, "Enforcing conceptual modeling to improve the understanding of human genome," in *Research Challenges in Information Science (RCIS), 2010 Fourth International Conference on*, 2010, pp. 85-92.

[86] O. Pastor*, et al.*, "Model driven-based engineering applied to the interpretation of the human genome," *The Evolution of Conceptual Modeling. Springer, Heidelberg,* 2010.

[87] N. Paton*, et al.*, "Conceptual modelling of genomic information," *Bioinformatics,* vol. 16, p. 548, 2000.

[88] M. Pastor, *et al.*, "Conceptual Modeling of Human Genome Mutations: A Dichotomy Between What we Have and What we Should Have," in *BIOSTEC Bioinformatics*, Valencia, 2010, pp. 160-166.

[89] A. Martin and M. Celma, "Integrating Human Genome Variation Data: An Information System Approach," in *Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on*, 2011, pp. 65-69.

[90] M. J. Villanueva, *et al.*, "Applying Conceptual Modeling to Alignment Tools One Step towards the Automation of DNA Sequence Analysis," in *Bioinformatics*, 2011, pp. 137-142.

[91] M. J. Villanueva, *et al.*, "Diagen: A Model-Driven Framework for Integrating Bioinformatic Tools," in *IS Olympics: Information Systems in a Diverse World*, ed: Springer, 2012, pp. 49-63.

[92] A. M. Martínez, *et al.*, "Facing the Challenges of Genome Information Systems: A Variation Analysis Prototype," in *Information Systems Evolution*, ed: Springer, 2011, pp. 222-237.

[93] A. Olivé, *Conceptual modeling of information systems*: Springer Science & Business Media, 2007.

[94] D. W. Embley and B. Thalheim, *Handbook of conceptual modeling: theory, practice, and research challenges*: Springer, 2011.

[95] O. Pastor and J. C. Molina, *Model-driven architecture in practice: a software production environment based on conceptual modeling*: Springer, 2007.

[96] O. Pastor, *et al.*, "The OO-Method approach for information systems modeling: from object-oriented conceptual modeling to automated programming," *Information Systems,* vol. 26, pp. 507-534, 2001.

[97] J. Rumbaugh, *et al.*, *Unified modeling language reference manual, the*: Pearson Higher Education, 2004.

[98] C. C. Compton, *et al.*, *AJCC Cancer Staging Atlas* vol. 2: Springer-Verlag New York, 2012.

[99] D. Maglott, *et al.*, "Entrez Gene: gene-centered information at NCBI," *Nucleic acids research,* vol. 39, p. D52, 2011.

[100] K. D. Pruitt, *et al.*, "NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy," *Nucleic acids research,* vol. 40, pp. D130-D135, 2012.

[101] *SQLite*. Available: www.sqlite.org

[102] S. Haldar, "SQLite Database System: Design and Implementation," *Motorola Mobility, Inc,* 2011.

[103] *DB-Engines: Knowledge Base of Relational and NoSQL Database Management Systems - DB-Engines Ranking*. Available: http://db-engines.com/en/ranking

[104] *Microsoft Office Access*. Available: https://products.office.com/es-es/access

[105] *Soporte de Microsoft Office: Conceptos básicos sobre bases de datos*. Available: https://support.office.com/es-es/article/Conceptos-b%C3%A1sicos-sobre-bases-de-datos-a849ac16-07c7-4a31-9948-3c8c94a7c204

[106] *Microsoft Store: Office*. Available: https://products.office.com/es-es/buy/office

[107] M. Kofler, *The definitive guide to MySQL 5*: Apress, 2006.

[108] *MySQL: The world's most popular open source database*. Available: https://www.mysql.com/

[109] P. D. Stenson, *et al.*, "Human gene mutation database (HGMD®): 2003 update," *Human Mutation,* vol. 21, pp. 577-581, 2003.

[110] *GRAMPS: Relational Database Comparison*. Available: https://gramps-project.org/wiki/index.php?title=Relational_database_comparison

[111] *Redis*. Available: https://redis.io/

[112] *Apache Cassandra*. Available: http://cassandra.apache.org/

[113] A. Lakshman and P. Malik, "Cassandra: a decentralized structured storage system," *ACM SIGOPS Operating Systems Review,* vol. 44, pp. 35-40, 2010.

[114] *MongoDB Spain - MongoDB:Characteristics and future*. Available: http://www.mongodbspain.com/en/2014/08/17/mongodb-characteristics-future/

[115] *MongoDB: Reinventando la gestión de datos*. Available: https://www.mongodb.com/es

[116] K. Chodorow, *MongoDB: the definitive guide*: " O'Reilly Media, Inc.", 2013.

[117] P. Danecek, *et al.*, "The variant call format and VCFtools," *Bioinformatics,* vol. 27, pp. 2156-2158, 2011.

[118] *VCF Variant Call Format*. Available: http://www.internationalgenome.org/wiki/Analysis/vcf4.0/

[119] *NCBI - Variation Glossary - VCF INFO tag terms*. Available: https://www.ncbi.nlm.nih.gov/variation/docs/glossary/#ui-ncbiinpagenav-heading-2

[120] *HashMap (Java Platform SE 8)*. Available: https://docs.oracle.com/javase/8/docs/api/java/util/HashMap.html

[121] *Sequence Variant Nomenclature: HGVS-nomenclature*. Available: http://varnomen.hgvs.org/

[122] S. F. Altschul, *et al.*, "Basic local alignment search tool," *Journal of molecular biology,* vol. 215, pp. 403-410, 1990.

[123] S. McGinnis and T. L. Madden, "BLAST: at the core of a powerful and diverse set of sequence analysis tools," *Nucleic acids research,* vol. 32, pp. W20-W25, 2004.

[124] E. Y. Chen, *et al.*, "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool," *BMC bioinformatics,* vol. 14, p. 1, 2013.

[125] D. Croft, *et al.*, "The Reactome pathway knowledgebase," *Nucleic acids research,* vol. 42, pp. D472-D477, 2014.

[126] *Affymetrix - Microarray Solutions*. Available: http://www.affymetrix.com/estore/browse/level_one_category_template_one.jsp?category=35796

[127] IEEE, "Systems and software engineering—Vocabulary ISO/IEC/IEEE 24765: 2010," *ISO/IEC/IEEE,* vol. 24765, pp. 1-418, 2010.

[128] N. Juristo and A. M. Moreno, *Basics of software engineering experimentation*: Springer Science & Business Media, 2013.

[129] ISO, "IEC 9126-1: Software Engineering-Product Quality-Part 1: Quality Model," *Geneva, Switzerland: International Organization for Standardization,* p. 27, 2001.

[130] D. L. Moody, "The method evaluation model: a theoretical model for validating information systems design methods," *ECIS 2003 proceedings,* p. 79, 2003.

[131] Z. Wang, *et al*., "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics,* vol. 10, pp. 57-63, 2009.

Anexo I: Cuestionarios diseñados para la validación

Cuestionario del método tradicional

Este cuestionario le ofrece la oportunidad de expresar su opinión sobre el seguimiento del **método tradicional** para almacenar y analizar datos clínicos de los pacientes. Por favor, lea cada sentencia y puntúela en base a su opinión. Los posibles valores de la puntuación son:

- 1= Totalmente en desacuerdo
- 2= Bastante en desacuerdo
- 3= Neutral
- 4= Bastante de acuerdo
- 5= Totalmente de acuerdo

Sentencias	1	2	3	4	5
1. Encontré el método tradicional sencillo y fácil de seguir para alguien que es la primera vez que lo utiliza.	O	O	O	O	O
2. Creo que el método utilizado facilita la tarea de almacenar datos de los pacientes.	O	O	O	O	O
3. El método tradicional reduce el esfuerzo necesario para analizar datos de los pacientes en conjunto	O	O	O	O	O
4. En general, encuentro el método tradicional cómodo y fácil de usar	O	O	O	O	O
5. Podría explicar el funcionamiento del método tradicional fácilmente a otra persona que no lo conociera	O	O	O	O	O
6. En general, encontré el método tradicional útil	O	O	O	O	O

7. En general, el método es práctico para visualizar los datos de uno de los pacientes y encontrar la información que busco.	O	O	O	O	O
8. En mi opinión, esta método me parece fiable y preciso para realizar los análisis de datos de mis pacientes	O	O	O	O	O
9. Definitivamente, usaría el método tradicional para almacenar y gestionar los datos de mis pacientes.	O	O	O	O	O
10. Creo que este método mejora la precisión de los datos almacenados con respecto a otros métodos	O	O	O	O	O
11. En general, creo que el método proporciona una solución efectiva para el almacenamiento y gestión de los pacientes	O	O	O	O	O
12. Usando este método puedo analizar los datos de mis pacientes de forma eficiente	O	O	O	O	O
13. Estoy seguro de que con este método puedo realizar un análisis de datos de los pacientes de manera más rápida y sencilla que utilizando otros procedimientos	O	O	O	O	O
14. En general, creo que este método es una mejora con respecto a otros procedimientos de gestión de pacientes	O	O	O	O	O
15. Intentaré usar preferiblemente este método con respecto a otras técnicas de gestión de pacientes en un futuro	O	O	O	O	O
16. Prefiero seguir utilizando este método para realizar los análisis de datos de mis pacientes	O	O	O	O	O

Cuestionario de la herramienta propuesta

Este cuestionario le ofrece la oportunidad de expresar su opinión sobre el seguimiento de la **herramienta propuesta** para almacenar y analizar datos clínicos de los pacientes. Por favor, lea cada sentencia y puntúela en base a su opinión. Los posibles valores de la puntuación son:

- 1= Totalmente en desacuerdo
- 2= Bastante en desacuerdo
- 3= Neutral
- 4= Bastante de acuerdo
- 5= Totalmente de acuerdo

Sentencias	1	2	3	4	5
1. Encontré la herramienta sencilla y fácil de utilizar para alguien que es la primera vez que la utiliza.	O	O	O	O	O
2. Creo que la herramienta utilizada facilita la tarea de almacenar datos de los pacientes.	O	O	O	O	O
3. La herramienta reduce el esfuerzo necesario para analizar datos de los pacientes en conjunto	O	O	O	O	O
4. En general, encuentro la herramienta propuesta cómoda y fácil de usar	O	O	O	O	O
5. Podría explicar el funcionamiento de la herramienta fácilmente a otra persona que no la conociera	O	O	O	O	O
6. En general, encontré la herramienta propuesta útil	O	O	O	O	O
7. En general, la herramienta es práctica para visualizar los datos de uno de los pacientes y encontrar la información que busco.	O	O	O	O	O
8. En mi opinión, esta herramienta me parece fiable y precisa para realizar análisis de datos de pacientes	O	O	O	O	O

9. Definitivamente, usaría la herramienta propuesta para almacenar y gestionar los datos de mis pacientes.	O	O	O	O	O
10. Creo que la herramienta propuesta mejora la precisión de los datos almacenados con respecto a otros métodos	O	O	O	O	O
11. En general, creo que esta herramienta proporciona una solución efectiva para el almacenamiento y gestión de los pacientes	O	O	O	O	O
12. Usando la herramienta propuesta puedo analizar los datos de mis pacientes de forma eficiente	O	O	O	O	O
13. Estoy seguro de que con esta herramienta puedo realizar un análisis de datos de los pacientes de manera más rápida y sencilla que utilizando otros procedimientos	O	O	O	O	O
14. En general, creo que esta herramienta es una mejora con respecto a otros procedimientos de gestión de pacientes	O	O	O	O	O
15. Intentaré usar preferiblemente esta herramienta con respecto a otras técnicas de gestión de pacientes en un futuro	O	O	O	O	O
16. Me gustaría utilizar la herramienta para realizar los análisis de datos de mis pacientes en un futuro	O	O	O	O	O

Anexo II: Tabla de resultados de la validación

RESULTADOS CUANTITATIVOS VALIDACIÓN	GLORIA		MARÍA		MAITE		ANNA		MEDIAS y PORCENTAJES	
	Trad.	Herram.	Trad.	Herram.	Trad.	Herram.	Trad.	Herram.	Trad.	Herram.
Tpo. Inserción Paciente	13'49''	32'33''	10'38''	27'13''	8'07''	18'27''	10'16''	27'13''	10'43''	26'22''
Datos insertados (79 datos en HC Herr. 66 daros en HC Trad.)	25	68	27	78	30	75	24	79	26	75
Tpo. Categorización	Jov 30'30''* May 3'06''	----------	Jov 10'40'' May 4'30''	----------	Jov 28'56''* May 3'29''	----------	Jov 15'43'' May 5'53''	----------	Jov 13'31'' May 4'15''	----------
Tpo. Ejercicio 2.1	7'26''* (+6'45'' = **14'11''**)	26''	1'49'' (+5'20'' = **7'09''**)	33''	1'10'' (+6'45'' = **7'55''**)	20''	50'' (+7'51'' = **8'41''**)	22''	7'55''	25''
Tpo. Ejercicio 2.2	1'51'' (+3'06'' = **4'57''**)	34''	1'10'' (+4'30'' = **5'40''**)	53''	33'' (+3'29'' = **4'02''**)	15''	1'35'' (+5'53'' = **7'28''**)	38''	5'32''	35''
Tpo. Ejercicio 2.3	1'04'' (+6'45'' = **7'49''**)	35''	3'32'' (+5'20'' = **8'52''**)	16''	2'12'' (+6'45'' = **8'57''**)	16''	49'' (+7'51'' = **8'40''**)	35''	8'35''	25''
Resultado Ejercicio 2.1	3	5	5	5	5	5	5	5	75%	100%
Resultado Ejercicio 2.2	2	4	4	4	4	4	4	4	75%	100%
Resultado Ejercicio 2.3	4	2	2	2	2	2	2	2	75%	100%
* Algunos problemas con el SPSS provocaron un retraso en el tiempo de los ejercicios										