Modelo de una bodega de datos para el soporte a la investigación bioinformática

Data Warehouse Model for Bioinformatic Investigation

Ligia Stella Bustos Ríos[1], Ricardo Moreno Laverde[2], Néstor Darío Duque[3]
Universidad Tecnológica de Pereira, Pereira, Colombia
ligia.bustosr@gmail.com
rmoreno@utp.edu.co
ndduqueme@unal.edu.co

***Resumen*— La Bioinformática es el uso de herramientas computacionales que permiten analizar, depurar y agilizar el manejo de grandes cantidades de datos de la biología en términos fisicoquímicos y permitir comprender y organizar la información asociada. La bioinformática parte de datos encontrados experimentalmente, los cuales son almacenados y sobre estos se aplican técnicas de consulta, de análisis y de extracción de conocimiento.**

Uno de los asuntos centrales la actualidad en este campo es definir el esquema de almacenamiento y las herramientas de análisis de los grandes volúmenes de datos generados y disponibles.

Este artículo presenta una revisión de las necesidades en la investigación bioinformática y de los enfoques de almacenamiento y procesamiento de datos, donde se muestra que una preocupación de los investigadores está relacionada con la posibilidad de aprovechar los datos obtenidos y poder extraer información y conocimiento subyacente. El análisis realizado conduce a concluir que aprovechar las tecnologías asociadas a Data Warehouse se presenta como una prometedora solución, pero requiere tomar decisiones que definirán el resultado final..

***Palabras clave*—:** **Bioinformática, bodegas de datos, herramientas de análisis, minería de datos.**

***Abstract*— Bioinformatics is the use of computational tools for analyze, debug and streamline the handling of large amounts of data in biology in terms physicochemical and help to understand and organize the information associated with it. Bioinformatics starts from data found experimentally, which are stored and apply later techniques for consultation, analysis and knowledge extraction.**

One of the central issues in this field today is to define the storage schema and tools for analyzing large volumes of data generated and available.

This article presents a review of bioinformatics research needs and approaches to data storage and processing, which shows that a concern of researchers is related to the ability to leverage the data and to extract information and underlying knowledge. The analysis leads to the conclusion that take advantage of the technologies associated with data warehouse is presented as a promising solution, but requires decisions that will shape the final outcome.

***Key Word* — Bioinformatics, data warehouse, OLAP, dataminig.**

I. INTRODUCCIÓN

En 1866 Mendel descubre los genes. Posteriormente, en 1871 se descubren los ácidos nucleicos: la gran molécula de la vida. Los primeros pasos en la genética fueron lentos y hasta el siglo siguiente no se hicieron descubrimientos nuevos, como por ejemplo en 1953, año en que se descubrió la estructura del ADN. A partir de este descubrimiento se empezó a buscar los genes en dicha estructura y, por consiguiente, la existencia de un código genético.

Entre 1975 y 1979, se aísla el primer gen humano. A partir de este momento la genómica da un salto espectacular y se pasa de estudiar un sólo gen a tener descifrados códigos genéticos sencillos pertenecientes a bacterias para finalmente llegar a conseguir la secuenciación completa del genoma humano. Una gran cantidad de datos generados gracias a la tecnología y que necesita de ésta para poder ser manejada [1].

Los avances que se han producido en biología molecular en los últimos años han provocado tal explosión de la información, que la comunidad científica se ha visto obligada a informatizar todo su conocimiento y aquí es donde aparece la bioinformática.

La mayor dificultad probablemente se encuentra en la tarea de capturar y modelar los diversos objetos biológicos y su complejo de relaciones. El análisis realizado conduce a concluir que aprovechar las tecnologías asociadas a Data Warehouse para la

[1] Ingeniera Industrial, Ingeniera de Sistemas, MSc. En Administración. , Doctorando en Ingeniería Informática
[2] Ingeniera Eléctrico, MSc. en Administración. Doctorando en Ingeniería Informática
[3] Ph.D. en Ingeniería – Sistemas, Profesor Asociado, Universidad Nacional de Colombia. Sede Manizales.

Fecha de Recepción: 26 de Agosto de 2011
Fecha de Aceptación: 11 de Octubre de 2011

investigación bioinformática, se presenta como una prometedora solución, pero requiere tomar decisiones que definirán el resultado final.

Este artículo pretende dar una visión de la propuesta orientada al diseño e implementación de un Almacén de Datos (Data Warehouse) que sirva como apoyo a la investigación bioinformática, partiendo de una mirada al marco conceptual de Data Warehouse y del campo de aplicación específico en bioinformática. Luego se hace una revisión del estado del arte en la aplicación de bodegas de datos para el almacenamiento de datos biológicos y su uso en obtención de información y conocimiento. A continuación se determinan los datos producidos en el proceso investigativo y sus fuentes primarias, a la vez que, las necesidades básicas de información y las expectativas de conocimiento por parte de los investigadores, enmarcados en la perspectiva del Data Warehouse, lo que permite llegar al modelo propuesto. Finalmente se presentan las conclusiones y trabajo futuro.

II. MARCO CONCEPTUAL

Para Rodríguez y otros (2006) la Bioinformática representa un campo científico muy amplio que resumen a partir de tres perspectivas distintas. La primera de sus perspectivas es la célula [1].

A partir de la célula suben de nivel de abstracción hasta los organismos individuales, los cuales representan la segunda perspectiva de la Bioinformática. Los genes, lejos de ser entidades estáticas, son regulados dinámicamente en respuesta al paso del tiempo, la región y el estado fisiológico. Por último, desde el más alto nivel de abstracción posible, proponen la tercera perspectiva de la Bioinformática: el árbol de la vida.

Relacionada con la segunda de las perspectivas plantean el análisis de datos de expresión genética, es decir, el estudio de los ARN transcritos por un conjunto de genes en distintas condiciones experimentales.
Los datos se almacenan en bases de datos, dentro de ellas se puede referir tres grandes bases de datos públicas que almacenan grandes cantidades de secuencias de Nucleótidos y Proteínas: GenBank, en el Centro Nacional de Biotecnología de los Estados Unidos (NCBI, http://www.ncbi.nih.gov), la Base de Datos de ADN de Japón (DDBJ, http://www.ddbj.nig.ac.jp/) y el Instituto Europeo de Bioinformática (EBI, http://www.ebi.ac.uk/) en Inglaterra. Estas tres instituciones intercambian sus secuencias diariamente como parte de una colaboración internacional.

En la tabla 1, extraída de [2] y tomada de Front Line Stratgic management Consulting (FLSMC), se aprecia el crecimiento de esta área, con ratas del 35%:

2000	2001	2002	2003	2004	2005	2010	CGRM % (2000-2005)	CGRM % (2005-2010)
468	609	824	1120	1508	1987	5421	33.5	22

Tabla 1. Cifras del mercado Bioinformático mundial. Ingresos en Millones de pesos alrededor del mundo.

La Bioinformática es un área de investigación multidisciplinaria, la cual puede ser ampliamente definida como la interface entre dos ciencias: Biología y Computación y está impulsada por la incógnita del genoma humano [3].

La Bioinformática es un considerada como una disciplina científica emergente que utiliza tecnología de la información para organizar, analizar y distribuir información biológica con la finalidad de responder preguntas complejas en biología [4].

Según el *National Institute of Health* de Estados Unidos, la Bioinformática es la investigación, desarrollo o aplicación de herramientas computacionales y propuestas científicas para extender y facilitar el uso de datos biológicos, médicos o sanitarios, incluyendo la adquisición, almacenamiento, organización, análisis y visualización de los mismos. Existe una disciplina científica relacionada con la Bioinformática, la Biología Computacional, que el mismo instituto define así: el desarrollo y la aplicación de datos analíticos y métodos teóricos, modelado matemático y técnicas de simulación por computador para estudiar sistemas biológicos, de conducta y sociales. La Biología Computacional es un campo más restringido basado en desarrollos matemáticos y métodos computacionales concretos.

A continuación se relacionan varios de los trabajos más importantes y desde diversas visiones que se han desarrollado o aun están en desarrollo en este campo.

III. REVISION DEL ESTADO DEL ARTE

Entre de las aplicaciones que se desarrollan actualmente o se prevén a corto plazo están el descubrimiento de drogas, portales de información genética, medicina forense, mejoramiento agrícola y ecológico, entre otros. En este aparte se incluyen algunas propuestas importantes encontradas en la revisión bibliográfica, lo que permite dar un vistazo a los diferentes enfoques y alternativas válidas para el modelo propuesto.

- Desde el año 1980, las bases de datos del Laboratorio de Biología Molecular Europeo, EMBL (European Molecular Biology Laboratory), del NCBI (Estados Unidos) y del laboratorio japonés DDBJ (DNA Databank of Japan) han recopilado las secuencias nucleotídicas publicadas hasta hoy. Actualmente existe una colaboración entre todas ellas, de forma que cada nueva entrada es automáticamente intercambiada con las otras dos restantes.

  Las secuencias proteicas son almacenadas y distribuidas por las bases de datos SWISS-PROT. Es una base de datos no redundante y mantiene numerosas referencias cruzadas con 26

bases de datos diferentes (BIB-GEN. Bases de datos genéticas). Las secuencias nucleotídicas son incorporadas a las bases de datos a un ritmo de 210 millones de pares de bases de datos al año. Sus datos se encuentran divididos en entradas, cada una de las cuales tiene un número de acceso, un conjunto de anotaciones que incluyen la descripción de la secuencia, información taxonómica del organismo del que deriva, lista de nombres de autores, referencias bibliográficas, características generales así como regiones de interés biológico y finalmente, la secuencia en sí [5].

Pero este campo ha abierto muchos caminos y opciones a investigadores de diferentes latitudes y disciplinas, mostrándose como un espacio de gran dinamismo. Esto se aprecia en la siguiente reseña.

- En un llamativo artículo, [6] muestra las posibilidades y ventajas que brindan las nuevas herramientas de virtualización de objetos reales con fines educativos, para ser usados en la generación de animaciones tridimensionales virtuales que permitan transmitir de manera audiovisual la información anatómica, fisiológica y quirúrgica, con el fin de simplificar y complementar el proceso educativo tradicional de la medicina y ciencias de la salud. Resume trabajos previos sobre el Corazón Virtual Animado, la Técnica de Prostatectomía Laparoscópica Dedo Asistida y el Sistema de información a pacientes.

- En [7] se plantea una solución al problema de descentralización y diagnóstico de las diferentes divisiones hospitalarias, enfocada en cuatro aspectos fundamentales como son: procesamiento de imágenes para generar diagnóstico, soporte de interconectividad física para compartir la información, administración eficiente de información referente a los estudios y administración de la interconexión. Reconocen la dificultad e importancia de administrar los datos recolectados en los procedimientos médicos y la necesidad de su centralización y administración.

- Wang y otros en artículo presentado describen su propuesta de modelamiento multidimensional para datos biomédicos, basados en una bodega de datos. Desarrollan un nuevo modelo llamado esquema BioStar que puede capturar la rica semántica de datos biomédicos y proporcionar una mayor extensibilidad y flexibilidad para la rápida evolución de las metodologías de investigación biológica. Esto se garantiza con el almacenamiento de las diferentes medidas en m-tablas separadas, las cuales son usadas para manejar las relaciones de muchos-a-muchos entre la entidad central y las dimensiones y pueden estar diseñados para soportar características específicas de una medida (por ejemplo, soporte bi-temporal de algunos datos clínicos). Además, es más eficiente el proceso para actualizar una m-tabla para relaciones con incertidumbre o con datos imprecisos en las entradas que para una tabla central de hechos de un esquema tradicional en estrella [8].

- Darmont y Olivier (2006) proponen e implementan un Data Warehouse para personalización de procesos en medicina. Plantean que el creciente uso de las nuevas tecnologías genera cambios significativos en las ciencias de la salud, tales como los registros electrónicos, que permiten personalizar la asistencia en saludo de por vida y el tratamiento pre-sintomático aprovechando varios análisis sobre una población dada de pacientes. Su objetivo es hacer que las personas administren como su capital su propia salud, formulando recomendaciones en relación con, por ejemplo, estilo de vida, nutrición o actividad física. Para lograr este objetivo, el sistema de apoyo a las decisiones deberá permitir análisis transversal de una población determinada y el almacenamiento de datos médicos globales biométricos tales como, datos biológicos, cardiovasculares, clínicos y psicológicos. En este trabajo se presenta el diseño del data warehouse con datos complejos que afectan a los deportistas de alto nivel de rendimiento [9].

- Ligand Depot es una fuente de datos integrado para encontrar información acerca de las moléculas pequeñas a las proteínas y los ácidos nucleicos. La versión inicial (versión 1.0, noviembre, 2003) se centra en proporcionar información química y estructural para pequeñas moléculas encontradas como parte de las estructuras depositadas en el Banco de Datos de Proteína (PDB). Ligand Depot acepta consultas basadas en palabras clave y también proporciona una interfaz gráfica para la realización de búsquedas en subestructura química. Una amplia variedad de recursos Web que contienen información sobre las moléculas pequeñas pueden accederse a través de ligand Depot [10]. Ligand Depot posee una interfaz de usuario y ha sido implementado como una aplicación Web cliente/ servidor de tres capas. Cuenta con navegador web en el cliente, un servidor de base de datos MySQL como el back-end y un servidor Tomcat en la aplicación servidor como nivel medio. El back-end tiene un conjunto de normalizado de tablas que almacenan las direcciones URL y otras informaciones sobre los sitios web relacionados con pequeñas moléculas. La lógica de procesamiento que ocurre en el nivel medio y es manejado por el servidor de aplicaciones usando Java Servlets. Ofrece capacidades flexibles de consulta una herramienta de dibujo para la realización de búsquedas de subestructura y un medio para importar y exportar archivos gráficos de moléculas pequeñas. Como un Data Warehouse Ligand Depot optimiza la consulta y reporte ligando la información presente en el PDB. Plantean como trabajos futuros la implementación de capacidades mejoradas de búsqueda y la incorporación de una más sofisticada interfaz gráfica de usuario.

- En un llamativo y avanzado trabajo presentado en [11] se expone el software llamado EMAAS (Extensible MicroArray Analysis System) que es una rica aplicación multi-usuario en Internet con una facilidad simple y robusta para acceso a los recursos actualizados a un microarray de almacenamiento de datos y análisis, combinado con herramientas integradas para optimización en tiempo real el apoyo a los usuarios y la formación. El framework EMAAS permite a los usuarios importar datos de microarrays de diversas fuentes hacia una base de datos subyacente, pre-procesar, evaluar y analizar la calidad de los datos, realizar análisis funcionales. Un número de paquetes de análisis, incluidos R-Bioconductor y Affymetrix Power Tools se han integrado en el servidor y está disponible programación mediante librerías Postgres-PLR o en clustres Grid. Los recursos integrados distribuidos incluyen la herramienta de anotación funcional DAVID, GeneCards y los repositorios de datos de microarrays GEO, CELSIUS y Mimir.

IV. DATA WAREHOUSE PARA BIONFORMATICA. MODELO PROPUESTO

A. Ámbito de la propuesta.

Para llegar al modelo de parte de las características particular de área de estudio y de las alternativas ofrecidas tecnológicamente y así proponer un modelo promisorio que desde lo tecnológico soporte la investigación.

Las principales dificultades que se obtienen en el proceso de investigación a la hora de administrar los datos y la información, son:

- Interés en manejar grandes volúmenes de datos.
- Múltiples y variadas fuentes de información
- Información dispersa y no oportuna con una alta probabilidad de inconsistencias
- Altos volúmenes de información no estructurada que requieren análisis
- Dificultad en acceso a la información histórica
- Falta de flexibilidad en la manipulación de información

A partir de la revisión de los proyectos, se encuentra que los requerimientos de este campo exigen el almacenamiento de grandes volúmenes de datos, con múltiples dimensiones, de periodos de tiempo extensos y con formatos heterogéneos al igual que sus fuentes. Por todo lo anterior la solución propuesta se basa en las tecnologías de Bodegas de Datos (Data Warehouse).

Un data warehouse es un conjunto de datos integrados orientados a una materia, que varían con el tiempo y que no son transitorios, los cuales soportan el proceso de toma de decisiones de la administración. (W. H. Inmon, considerado como el padre del data warehouse) [12]. Data Warehouse es un concepto relativamente nuevo, orientado al manejo de grandes volúmenes de datos, provenientes de diversas fuentes, de muy diversos tipos. Estos datos cubren largos períodos de tiempo, lo que trae consigo que se tengan diferentes esquemas de las bases de datos fuentes. Su misión consiste en, a partir de estos datos y apoyado en herramientas sofisticadas de análisis, obtener información útil para el soporte a la toma de decisiones [13]. El data warehousing o almacenamiento de datos es el proceso de reunir información histórica de una organización en una(s) base(s) de datos central(es) [14].

Los procesos asociados a Data Warehouse [13], [15] como se muestra en la figura 1 son:

- Población (Cargue inicial, actualizaciones)
- Almacenamiento (Estrategias para lograr eficiencia y disponibilidad)
- Uso de herramientas para obtención de información y extracción del conocimiento.

Este último proceso reviste gran importancia para la propuesta, pero se fundamenta en el contenido de la bodega de datos, con posibilidades de obtener información a partir de simples consultas o aplicando herramientas OLAP (On-Line Analytical Processing), que permiten obtener información relacional y multidimensional y mejor aún apoyarse en técnicas de minería de datos para extraer conocimiento oculto y realizar tareas descriptivas e incluso predictivas

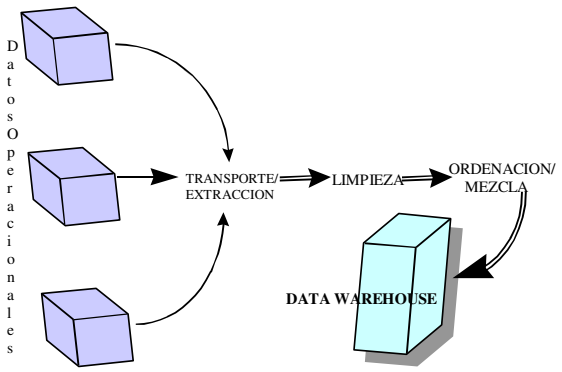

Figura 1. Proceso en bodegas de datos [13].

Estas son algunas de las tareas que deben ser sorteadas por el equipo encargado del diseño, implementación y montaje del sistema de bodegas de datos [16]: La integración de datos y metadatos de diferentes fuentes y épocas; limpieza, filtrado y refinación de los datos; en los sistemas de procesamiento en línea (OLTP) el detalle de las operaciones son muy importantes mientras que el Data Warehouse se busca almacenar datos en forma condensada y agrupada.

Otra decisión a tomar en la propuesta del modelo informático es la selección de la arquitectura. La arquitectura enfoca el proyecto como componentes (Fuente de datos, bodega de datos, datamart y el acceso y uso). La correcta definición de la misma es una condición para el éxito del proyecto [16].

Sobre este aspecto, los proveedores ofrecen diferentes modelos, de los cuales varios enfoques son elegibles:

- Consultas desde un esquema virtual hacia los datos operacionales. Normalmente una bodega de datos se asocia con un almacén donde se hacen copias de datos de aplicaciones en producción y de carácter histórico. En esta arquitectura se elimina la copia y actualización y se usan los datos de las bases de datos operacionales, a partir del metamodelo del Data Warehouse, los cuales se accesan al momento de la consulta.

- Almacenamiento propio a partir de varias fuentes. Bodega de datos empresarial, no necesariamente centralizada. Se apoya en la normal necesidad de preprocesar los datos desde las fuentes en operación y aboga por realizar esta tarea una vez y almacenarlos en bases propias, que serán actualizadas periódicamente. A partir de éstas se aplican las herramientas de análisis. Esta estrategia asegura la consistencia, pero es complejo de crear.

- Datamarts o mercado de datos únicamente. Plantea y reconoce las particularidades de cada área o departamento de una organización y la imposibilidad de ser satisfechos sus requerimientos por un solo Data Warehouse. El concepto de datamarts es una analogía a tiendas de vecindario que sirven a la población del sector, en lugar de un gran supermercado que abastece toda la ciudad. Los Datamarts son sub-bodegas, organizadas por temas a nivel de departamentos. Esta arquitectura solo usa datamart.

- Data Warehouse y mercado de datos. Es una combinación de las dos anteriores. El Data Warehouse corporativo es un recopilador y distribuidor de la información sin desconocer las particularidades específicas de cada área. Esta estrategia permite posibles inconsistencias en los datos.

- Cliente Servidor en dos capas. Solo existen servidores de datos y clientes que los usan. En el servidor (o servidores) residen las fuentes de datos, el Data Warehouse y los datamarts. En los clientes, se ejecutan las herramientas de acceso del usuario final; éstas son generalmente aplicaciones gráficas.

- Cliente Servidor en tres capas. Las tareas se dividen en tres niveles.
  - Un servidor de datos, que contiene las fuentes de los datos.
  - Un servidor de aplicaciones, que contienen los datos de la bodega de datos y manejan el software de Data Warehouse y datamarts.
  - La porción cliente, que manejan las aplicaciones de consulta y reporte.

Tomada la decisión de la arquitectura, y siendo la bodega de datos el resultado de la importación de datos de diferentes fuentes, las cuales son dinámicas, cambian con el tiempo, se requiere generar mecanismos que garanticen la sincronización y aseguren la actualización a partir de los cambios en las fuentes.

Para una correcta operación de la bodega de datos es necesario tener correcta información sobre los datos que se tienen almacenados y entonces la administración de metadatos toma importancia.

El diseño de las bodegas de datos incluye el modelamiento dimensional, el análisis de fuentes de datos, el diseño físico y el diseño de la arquitectura técnica. La definición de los requerimientos de análisis como un modelamiento dimensional, permite identificar las tablas de hechos y las dimensiones asociadas, incluyendo el detalle de atributos y jerarquías.

B. Modelo propuesto.

1. Datos en el objeto de estudio.

Para el objeto de la propuesta y por definición de los expertos en el campo de estudio, los datos de la bioinformática consisten en información biológica y medica de diversos tipos: Identificación del paciente, factores de riesgo: vivienda, el entorno, síntomas, enfermedades, características del paciente, examen físico: mediciones, exámenes paraclínicos, marcadores moleculares: secuencias, exámenes especializados: imagen, radiografías.

Estos datos incluyen secuencias biológicas (DNA, ARN, y proteínas), genes o expresión de proteína, características funcionales, interacciones moleculares, datos clínicos, descripciones de sistemas, y publicaciones relacionadas.

Los datos aparecen como secuencias, anotaciones de secuencias, modelos estructurales, mapas físicos, expedientes clínicos, caminos de interacción, genes y expresiones de la proteína, interacciones de la proteína-proteína, y otras fuentes tales como bases de datos, colecciones de los datos confidenciales, y publicaciones relacionadas.

2. Selección de Arquitectura del Data Warehouse.

Retomando conceptos de [12] y [16] es necesario reconocer que otro elemento que reviste importancia al momento de implementar una bodega de datos, es la selección de la arquitectura. La arquitectura enfoca el proyecto como componentes (Fuente de datos, bodega de datos, data mart y el acceso y uso).

A partir de la información recolectada, el estado del arte y las consultas de los expertos se propone el siguiente esquema básico para el Data Warehouse, siendo un sistema hibrido de Data warehouse y data marts en el marco de un sistema cliente/servidor. La figura 2 recoge la propuesta.

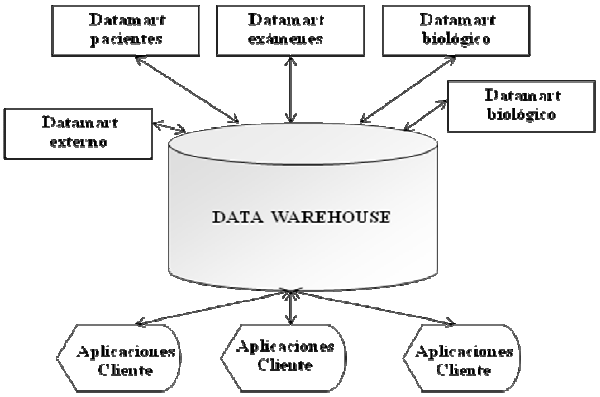

Figura 2. Modelo propuesto.

3. Aplicación de Descubrimiento de Conocimiento (KDD) en los datos almacenados.

Hernández y otros definen KDD como el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos. La minería de datos (DM) como parte del proceso, aunque en ocasiones se trata indistintamente con KDD, integra técnicas de análisis de datos y extracción de modelos. Se basa en varias disciplinas, algunas de ellas más tradicionales como la estadística y el aprendizaje auto-mático, se diferencia de ellas en la orientación más hacia el fin que hacia los medios [17]. La DM tiene como objetivo analizar los datos para extraer conocimiento, en forma de relaciones, patrones o reglas inferidos de los datos y (previamente) desconocidos, o bien en forma de una descripción más concisa (es decir, un resumen de los mismos). Estas relaciones o resúmenes constituyen el modelo de los datos analizados.

Dentro de DM se distinguen dos tipos de tareas: predictivas, donde se tratan problemas en los que hay que predecir uno o más valores para uno o más ejemplos; y descriptivas, que buscan describir y arrojar luces a la interpretación de los datos. Para la solución de dichas tareas, se emplean diferentes métodos y técnicas entre ellas: las técnicas algebraicas y estadísticas, técnicas bayesianas, técnicas basadas en árboles de decisión y sistemas de aprendizaje de reglas, técnicas basadas en redes neuronales artificiales y difusas. También se pueden combinar las técnicas ya mencionadas, para aprovechar las ventajas que pueda ofrecer cada una de ellas [18].

La figura 3 muestra las diferentes fases del proceso KDD tomado de [19]

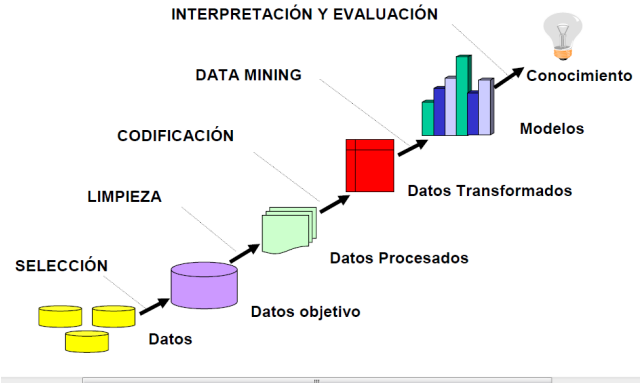

Figura 3. Etapas del KDD.

Dentro de estas fases, es posible distinguir el proceso denominado ETL (Extraer, Transformar y Cargar) que reviste gran importancia e implica transferir, formatear, limpiar y cargar datos desde los datos fuentes al esquema de la Bodega de datos. Las tareas son particulares para cada caso y ocupan entre el 40% al 60% del esfuerzo del proceso total.

Para esta fase se evaluaron varias herramientas y en el modelo se seleccionó, no de forma excluyente, Talend Open Studio, una potente y versátil solución open source para integración de datos a través de un ambiente de desarrollo grafico fácil de usar. Talend soporta todos los tipos de datos, su migración y las operaciones de sincronización. La figura 4 muestra su amigable interfaz.

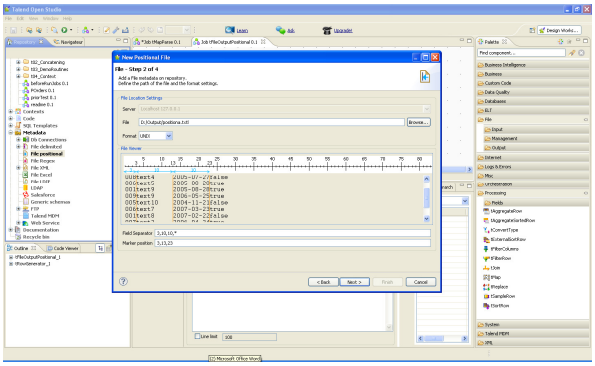

Figura 4. Interfaz de Talend.

Luego de tener los datos según las decisiones de diseño, se debe avanzar en aplicar técnicas de minería de datos, lo que puede hacerse en forma iterativa según los intereses del analista, con el fin de construir los modelos que permitan extraer conocimiento oculto y que sirva para avanzar en las investigaciones. La figura 5 agrupa algunos algoritmos explorados con los tipos de datos disponibles y que han demostrado que cumplen la función esperada. Se aclara que no es resultado definitivo ni una visión terminada, es una investigación en marcha.

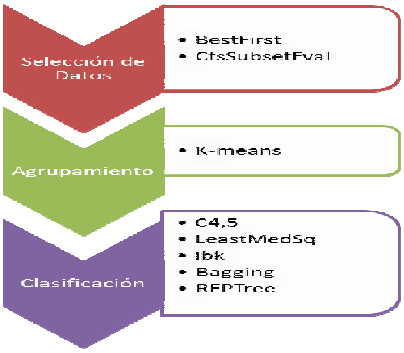

Figura 5. Fases propuestas en el modelo.

V. CONCLUSIONES Y TRABAJO FUTURO

Las bases de datos que existen en todo el mundo contienen datos de más de millón y medio de proteínas y con tendencia al crecimiento. La administración de estos recursos exige contar con grandes capacidades de procesamiento en hardware y software. Cada vez más, los estudios biológicos parten de la conexión de múltiples bases de datos y de complejos algoritmos de análisis de datos para formular hipótesis que versan sobre la organización de los genes, el análisis de su secuenciación y la predicción de su estructura y comportamiento.

La propuesta presentada es un acercamiento probado a la organización de dichos datos provenientes de diversas fuentes y su posterior procesamiento orientado a descubrir patrones que se traduzcan en conocimiento.
Como trabajo futuro se requiere evaluar las plataformas tecnológicas y la implementación de los procesos asociados a la solución propuesta, que permita cargar todos los datos disponibles en la investigación y obtener resultados que se puedan generalizar. No obstante en esta fase los investigadores ven representados sus intereses y requerimientos lo que es condición fundamental para el éxito del sistema planteado.

REFERENCIAS.

[1] Rodríguez Baena, D. ; Riquelme Santos, José C.; Aguilar Ruiz, Jesús S. . Análisis de datos de Expresión Genética mediante técnicas de Biclustering. Sevilla. 2006

[2] Barraza, F. Cluster de Bioinformática en el Valle del Cauca. Parquesoft. Cali. 2010.

[3] Hernández, Eugenio. Bioinformática: una nueva área de oportunidad. México, 2003, p. 1.

[4] MKM Editorial. Bases de Datos Bioinformáticas Disponible en línea: http://www.mkm-pi.com/mkmpi.php?article201. Consultado en septiembre de 2009..

[5] Bib-Gen. Biblioteca Virtual en salud. España. 2010. Disponible en http://bvs.isciii.es/bib-gen/Actividades/curso_virtual/ Ftes_informacion/fteinformacion4.htm. Consultado mayo de 2010.

[6] Escobar Roa, Juan Miguel. Aplicaciones virtuales en biomedicina. Revista Digital Vol 1 No. 1 Facultad de Ingeniería Electrónica. Universidad El Bosque. Bogotá. 2006.

[7] Prieto Reyes Sandy Johana, Salcedo López Dennys Marcela, Torres Romero Oscar Mauricio. Central de procesamiento de imágenes médicas para General Médica de Colombia S.A. Revista Digital Vol 1 No. 1 Facultad de Ingeniería Electrónica. Universidad El Bosque. Bogotá. 2006.

[8] Liangjiang Wang Murali Ramanathan y Aidong Zhang. BioStar models of clinical and genomic data for biomedical data warehouse design. Bioinformatics Research and Applications. 2005.

[9] Darmont, Jérôme; Olivier, Emerson . A Complex Data Warehouse For Personalized, Anticipative Medicine. University of Lyon. 2006.

[10] Zukang Feng, Li Chen, Himabindu Maddula, Ozgur Akcan, Rose Oughtred, Helen M. Berman and John Westbrook. Ligand Depot: a data warehouse for ligands bound to macromolecules. Bioinformatics Applications Note Vol. 20 no. 13. 2004.
Disponible en http://bioinformatics.oxford journals.org/

[11] Barton, J Abbott, N Chiba, DW Huang, Y Huang, M Krznaric, J Mack-Smith, A Saleem, BT Sherma, B Tiwari, C Tomlinson, T Aitman, J Darlington, L Game, MJE Sternberg and SA Butcher. EMAAS: An extensible grid-based Rich Internet Application for microarray data analysis and management. BMC Bioinformatics. 2008.
Disponible en: http://www.biomedcentral.com/1471-2105/9/493

[12] Harjinder, S. Gill. Prakash, C. Rao. Data Warehousing. La Integracion de Informacion para la Mejor Toma de Decisiones. Prentice Hall. Mexico, 1996.

[13] Duque, Néstor Darío. Tamayo, Alonso. Data Warehouse: Herramienta para la toma de decisiones. Parte II.. Revista NOOS. Número 13. Universidad Nacional de Colombia. Manizales, 2001.

[14] Orfali, Robert. Harkey, Dan. Edwards, Jeri. Cliente/Servidor. Guía de Supervivencia. Segunda edicion. McGraw-Hill. México, 1997.

[15] Escalante, Iván. Data Warehouse. Revista Soluciones Avanzadas No.34. Universidad de la Habana. Cuba. 1996.

[16] Duque, Néstor Darío. Tamayo, Alonso. Data Warehouse: Herramienta para la toma de decisiones. Parte I. Revista NOOS. Número 12. Universidad Nacional de Colombia. Manizales, 2001.

[17] Hernández Orallo, José, Ramírez Quintana, María José y Ramírez Ferri, César. 2004. Introducción a la Minería de Datos. Madrid : Pearson, Prentice Hall.

[18] Osorio Z. Germán Augusto; Sánchez G., Luis Gonzalo; Duque M. Néstor Darío. 2009. Técnicas de minería de datos aplicadas a la valoración de ambientes creativos. Revista Respuestas. Año 14. No. 1. Junio de 2009. Universidad Francisco de Paula Santander. Cúcuta.

[19] Mensalvas, E y Millan, S. El proceso de Data Mining. Universidda del Valle. 2005.