# Sistema de Soporte a la Toma de Decisiones para el Análisis Multifractal del Genoma Humano

## Decision Support System for Multifractal Analysis of the Human Genome

**Martha Eliana Mendoza Becerra**
Ph.D. (c) en Ingeniería de Sistemas y Computación,
Universidad Nacional de Colombia
M.Sc. en Informática, Universidad Industrial de Santander
Docente Titular Tiempo Completo, Investigador Grupo de I+D en Tecnologías de la Información (GTI),
Universidad del Cauca, Popayán, Colombia
mmendoza@unicauca.edu.co

**Alba Viviana Camayo Otero**
Ingeniera de Sistemas, Universidad del Cauca
Auxiliar de Investigación en el Grupo de I+D en Tecnologías de la Información (GTI), Universidad del Cauca
Popayán, Colombia
acamayo@unicauca.edu.co

**Adrián Fernando Martínez Molina**
Ingeniero de Sistemas, Universidad del Cauca
Auxiliar de Investigación en el Grupo de I+D en Tecnologías de la Información (GTI), Universidad del Cauca
Popayán, Colombia
afmartinez@unicauca.edu.co

**Émber Ubéimar Martínez Flor**
M.Sc. (c) en Ingeniería de Sistemas, Universidad del Valle
Docente Asociado Tiempo Completo, Investigador Grupo BIMAC,
Universidad del Cauca, Popayán, Colombia
eumartinez@unicauca.edu.co

**Carlos Alberto Cobos Lozada**
Ph.D. (c) en Ingeniería de Sistemas y Computación,
Universidad Nacional de Colombia
M.Sc. en Informática, Universidad Industrial de Santander
Docente Titular Tiempo Completo, Coordinador del Grupo de I+D en Tecnologías de la Información (GTI),
Universidad del Cauca, Popayán, Colombia
mmendoza@unicauca.edu.co

**Resumen— En este artículo se presenta el modelo de datos multidimensional de dos data marts que forman parte de un Sistema de Soporte a la Toma de Decisiones en el área de la Genómica, el cual está basado en tecnologías de Bodegas de datos y OLAP. El primer data mart está relacionado con el "Análisis de unidades de información", que permite almacenar y consultar información sobre las unidades de información (Exón o Intrón) en la estructura de un gen, el orden y la posición inicial y final de las unidades de información. El segundo data mart llamado "Análisis fractal" permite almacenar y consultar información sobre los genes, por ejemplo, el número de unidades de información y longitud del gen, y medidas adicionales obtenidas del análisis fractal realizadas por una investigación previa. Finalmente, se presentan los problemas durante el proceso de cargue de datos y el modelado de los datos, junto con las soluciones planteadas a los mismos, y algunas interfaces de la herramienta desarrollada.**

**Palabras clave— Bodegas de Datos, OLAP, Bioinformática. Proceso ETL.**

**Abstract— This paper presents the data model of two multidimensional data marts that are part of a Decision Support System in the genomics' area, which is based on data warehousing and OLAP technologies. The first data mart is related to the "Units information analysis", which can store and retrieve information about units information (Exon or Intron) in a gene structure, the order, and the start and end position of information units. The second data mart called "fractal analysis" allows you to store and retrieve information about genes, for example, the number of information units and length of the gene, and measurements obtained by previews fractal analysis research. Finally, this paper presents the problems during the extraction, transformation and loading data process and data modeling, together with the proposed solutions to them, and some interfaces of the developed tool.**

**Keywords— Date warehouse, OLAP, Bioinformatics, ETL process.**

## I. INTRODUCCIÓN

En el año de 1990 se dio inicio al proyecto del genoma humano, un esfuerzo coordinado por el departamento de energía de los Estados Unidos y los institutos nacionales de salud de Japón, Francia y Alemania, entre otros [1]. Para suplir la necesidad de almacenar las secuencias y anotaciones generadas por este proyecto y realizar análisis

posteriores sobre estos datos, se usaron bases de datos como el *GenBank* (Base de datos pública de secuencias de nucleótidos y proteínas) [2], la Base de Datos de Secuencias de Nucleótidos del Laboratorio Europeo de Biología Molecular [3] y el Banco de Datos de DNA de Japón [4], entre otras.

Estas bases de datos, denominadas Bases de Datos Biológicas, comúnmente almacenan los datos en un archivo de texto plano, donde se guardan además de la secuencias de nucleótidos o proteínas, anotaciones, descripciones textuales, y atributos, entre otros, referentes a un organismo. Cada uno de los valores almacenados son organizados de forma tabular y contiene etiquetas o delimitadores que indican el inicio o fin de un conjunto de valores relacionados. Sin embargo, desde el punto de vista de las ciencias de la computación o la informática, esta forma de almacenar los datos corresponde con una base de datos no estructurada lo que dificulta el procesamiento y la búsqueda de información específica.

Las desventajas que presenta un sistema de gestión de datos basado en archivos planos son evidentes y se encuentran ampliamente discutidos por la comunidad científica, sin embargo, muchas de las bases de datos biológicas mantienen este enfoque. Esto obliga a los investigadores a generar aplicaciones específicas para la recuperación, limpieza, procesamiento y análisis de la información, tareas que generalmente demandan alta capacidad de cómputo.

Las investigaciones que se realizan en el área de la genómica, por lo general, empiezan con la descarga de las secuencias de nucleótidos y proteínas en archivos de texto plano. Posteriormente, se realizan los análisis requeridos por la investigación especifica y se genera un conjunto de datos que se almacenan por lo regular en archivos de texto plano y en algunas ocasiones en bases de datos relacionales; estos datos generados son luego analizados -generalmente en hojas de cálculo y por medio de un proceso manual, repetitivo y que requiere de mucho tiempo- para comprobar las hipótesis definida en cada investigación.

Al siguir este proceso general de gestión de datos, se pueden destacar tres problemas principales: (i) No se cuenta con un almacenamiento estructurado que permita centralizar la gran cantidad de archivos planos que se tienen y los datos que estos contienen, (ii) en un entorno de archivos planos no unificados, es difícil manejar los posibles errores de consistencia en los campos de los archivos provenientes del *GenBank*, y (iii) basados en esos archivos planos, se dificulta la tarea de encontrar relaciones o asociaciones entre las variables involucradas y contar con una herramienta que represente gráficamente los datos.

En un esfuerzo interinstitucional por entender ciertas particularidades del genoma humano, en el año 2004 se formuló el proyecto "Análisis Multifractal del Genoma Humano para la Búsqueda de Regularidades con Significado Biológico y una Contribución a la Generación de Biotecnología de la Información" (AMGH), que fue ejecutado por el Grupo de Biología Molecular, Ambiental y del Cáncer (BIMAC) y el Grupo de I+D en Tecnologías de la Información (GTI) de la Universidad del Cauca [5], en asocio con la Universidad del Valle, Universidad de Cantabria (Santander, España) y *Triesta Science*s (India) Pvt. Ltd. (Bangalore, India/Menlo Park, California).

En este proyecto (AMGH) se aplicaron diferentes técnicas y herramientas de Bioinformática y de Geometría Fractal para la construcción de un modelo teórico, no lineal e "integrativo" que explica cómo se estructura y funciona el genoma humano (ver principales resultados del proyecto en [6] y [7]). Este estudio se adelantó a partir de la información del genoma humano publicada en el *GenBank* [2, 5] y la gestión de datos también se basó en archivos planos, heredando de esta forma los problemas mencionados previamente.

En este artículo se presenta una solución computacional original para el almacenamiento y análisis de los datos biológicos generados en el proyecto de investigación AMGH. Para resolver los tres problemas previamente mencionados en el marco del proyecto AMGH, se desarrolló un sistema de soporte a la toma de decisiones (DSS, por sus siglas en inglés, *Decision Support Systems*) que permite a los investigadores del proyecto analizar los datos desde una perspectiva multidimensional. Para el desarrollo del DSS se modeló una bodega de datos que centraliza, homogeniza y unifica las diversas fuentes de datos y se adaptó una herramienta de procesamiento analítico en línea (OLAP, por sus siglas en inglés*, On Line Analytical Processing*) para obtener una mejor y más versátil forma de visualizar y explorar los datos generados en el proyecto AMGH.

Para el desarrollo de la bodega de datos se realizó inicialmente el proceso de extracción, transformación y carga (ETL, por sus siglas en inglés, *Extraction, Transform and Load*). La extracción se realizó tomando los archivos planos generados en el proyecto AMGH, luego estos archivos fueron depurados, integrados y cargados en un mismo repositorio (la bodega de datos). Esta bodega fue procesada y convertida en un cubo multidimensional; el cual es accedido, visualizado y analizado por parte de los investigadores a través de una herramienta OLAP adaptada. Los investigadores usan esta herramienta para encontrar relaciones y estadísticas importantes en los datos de una forma mucho más sencilla, clara y flexible; evitando realizar los procesos manuales y repetitivos que se debían realizar sin la solución planteada en este artículo.

A continuación, en la sección 2 se presentan algunos trabajos relacionados con el uso de bodegas de datos en bioinformática. Luego, en la sección 3 se presenta los aspectos más relevantes sobre el modelado y la presentación de los datos. Posteriormente en la sección 4, se comentan los problemas que se presentaron en el desarrollo de la investigación con sus respectivas soluciones. En la sección 5 se muestran algunas características de la herramienta OLAP y por último se presentan las conclusiones y algunas recomendaciones que pueden ser útiles para proyectos futuros de DSS que se realicen en el área de la bioinformática.

## II. TRABAJOS RELACIONADOS

En el contexto bioinformático cuando se habla de bodegas de datos, se hace referencia a bases de datos unificadas (*Unified databases*) o integradas, pero no de una bodega de datos "real", Por lo anterior, es difícil encontrar en bioinformática el uso de una metodológica reconocida de bodegas de datos o verdaderos modelos multidimensionales, estos últimos conformados por tablas de hechos, medidas y dimensiones, características comunes de las bodegas de datos.

Es importante tener en cuenta que una bodega de datos permite almacenar e integrar datos desde múltiples fuentes y permite a usuarios de alto nivel ejecutar consultas y análisis de dichos datos desde diferentes perspectivas (múltiples dimensiones o multidimensionalidad) [8] [9], sin tener que acceder a las fuentes originales. Para lograr lo anterior, en el desarrollo de una bodega de datos se hace especial énfasis en la homogenización de los datos [10], lo que implica que los datos cargados desde las diversas fuentes son transformados mediante un mapeo a un formato único y estándar antes de ser físicamente almacenados. Adicionalmente, durante este proceso se pueden llevar a cabo el filtrado, la validación y, de ser necesario, la modificación de datos obtenidos de las fuentes originales [11] [12]. También es importante tener claro que el análisis multidimensional de los datos en la bodega de datos es una característica clave para diferenciarla del concepto de base de datos unificada, ya que estas últimas no lo permiten y utilizan sentencias SQL estándar (*structured query language*) para mostrar resultados a los usuarios.

A continuación se presentan algunos de los proyectos de genómica más representativos donde se han usado bases de datos unificadas y bodegas de datos.

Un sistema de bodegas de datos para analizar familias de proteínas (DWARF, por sus siglas en inglés, *A data warehouse system for analyzing protein families*) [13], propone una base de datos unificada e integrada desde la perspectiva relacional, en lugar de una bodega de datos multidimensional. Este sistema integra datos acerca de las secuencias, estructuras y anotaciones funcionales de las familias de proteínas. Los modelos relacionales asociados a este sistema se componen de tres grandes secciones, representadas en entidades asociadas a la proteína, la estructura de la proteína y la secuencia de la proteína, es decir, esta base de datos unificada relaciona familias de proteínas y permite evaluar sus relaciones entre secuencia, estructura y función.

Una bodega de datos bioinformática integrada (ATLAS, *A data warehouse for integrative bioinformatics*) [14], es un sistema que presenta una base de datos unificada de datos biológicos. ATLAS almacena e integra secuencias biológicas, interacciones moleculares, información de homología (estudio comparativo de los seres vivos), anotaciones funcionales de genes y ontologías biológicas, El sistema busca proporcionar datos integrados desde diferentes fuentes heterogéneas, así como una infraestructura de software para la investigación y el desarrollo de la bioinformática.

Bodega de datos genéticos (GEWARE, por sus siglas en inglés, *Genetic Data Warehouse*) [15], es una bodega de datos que soporta grandes volúmenes de datos provenientes de experimentos y estudios que buscan el análisis de expresiones de genes. GEWARE almacena centralizadamente datos de expresiones junto con una variedad de anotaciones hechas previamente por investigadores con el objeto de soportar diferentes formas de análisis. GEWARE permite gran flexibilidad en el análisis de los datos basado en un modelo de datos multidimensional donde los datos son almacenados en varias tablas de hechos las cuales están asociadas con múltiples dimensiones jerárquicas que describen las anotaciones en los genes, muestras, experimentos y métodos de procesamiento. Las anotaciones son integradas y almacenadas de una forma genérica basada en una plantilla y vocabulario predefinido. El sistema cuenta con varios métodos de análisis, entre ellos el análisis de grupos de genes, de grupos de experimentos y de matrices de expresiones. Este sistema es totalmente operacional y soporta el desarrollo de varios proyectos de investigación.

En la Tabla I se realiza una comparación de estos proyectos, teniendo en cuenta el manejo de los principales conceptos del modelado multidimensional en bodegas de datos.

TABLA I.

CRITERIOS COMPARACIÓN TRABAJOS RELACIONADOS

| | ATLAS | DWARF | GEWARE |
|---|---|---|---|
| Modelo de datos | Base de datos relacional | Base de datos relacional | Esquema estrella no estándar |
| Dimensiones | No se mencionan | No se mencionan | Si existen |
| Tabla de Hechos | No se mencionan | No se mencionan | Si existen |
| Medidas | No existen | No existen | Existen, pero no se pueden identificar |
| Data Mart | No existen | No existen | Existen, pero no se pueden identificar |
| Información | Interacciones moleculares | Proteínas | Pacientes clínicos y Experimentos médicos |

Fuente: Autores del proyecto

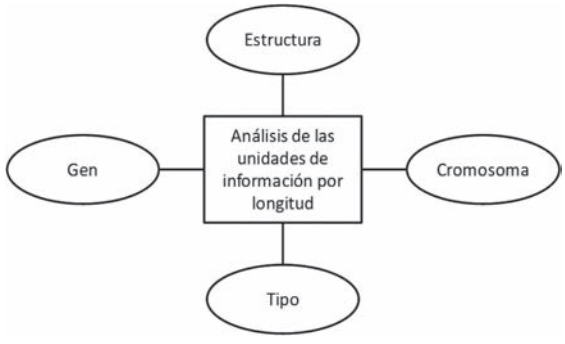
Como se puede apreciar, en ATLAS y DWARF se usa el término bodega de datos para describir bases de datos unificadas (o centralizadas), ya que los modelos de datos que presentan son modelos relacionales, y no utilizan los conceptos de "tablas de hechos" o "dimensiones", los cuales son característicos del modelo multidimensional de las bodegas de datos.

En GEWARE se observa el uso de conceptos relacionados con el modelo dimensional de bodega de datos, pero los modelos generados almacenan información clínica de los pacientes y no de secuencias de ADN, como es el caso de la investigación en AMGH, además estos modelos no se presentan en un modelo dimensional estándar, lo que los hace difíciles de entender.

## III. MODELADO DE DATOS

El desarrollo del DSS en el marco del proyecto AMGH, se centró en dos objetivos principales, el primero, permitir al investigador seleccionar el número de unidades de información que desea analizar, por ***estructura***, visualizar a qué ***cromosomas, genes*** ó familias (***tipos***) de genes pertenecen esas estructuras y, finalmente, poder organizar en un orden ascendente o descendente las longitudes de las unidades de información (Exón o Intrón). Para lograr lo anterior se modeló un Data Mart denominado Análisis de Unidades de Información (ver Fig. 1).

**Fig. 1.** MODELO DIMENSIONAL ANÁLISIS UNIDADES DE INFORMACIÓN

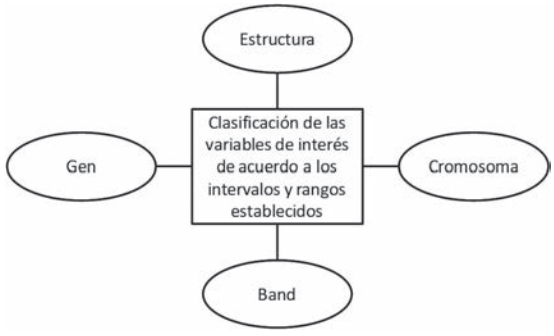



Fuente: Autores del proyecto

El segundo objetivo, permitir clasificar en rangos (***band***) una medida resultante de la investigación, la cual es obtenida por cada ***estructura*** y consultar dentro de cada uno de esos rangos las longitudes, número de unidades de información, sus porcentajes por ***gen*** y sus ***estructuras***, además calcular el número D (valor que corresponde al análisis fractal realizado al gen por los investigadores del proyecto AMGH [6]) por familia de Genes. Este segundo objetivo se modeló en el Data Mart denominado Análisis Fractal (ver Fig. 2).

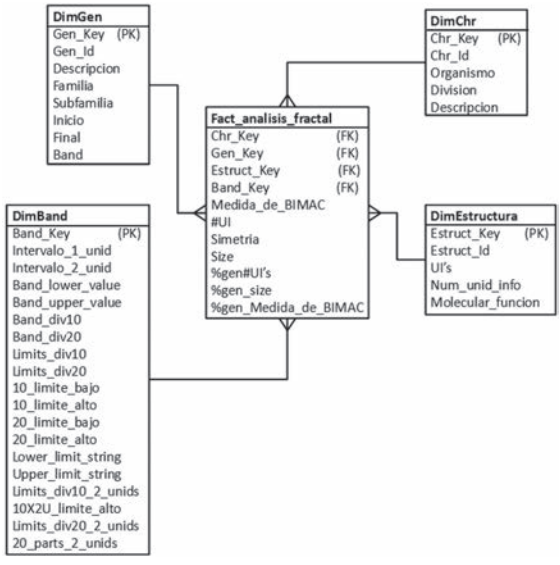
Este modelo se explica con mayor detalle (Ver Fig. 3, debido a que presenta un caso especial de modelado multidimensional. En el modelo se aprecia que todas las dimensiones presentan una clave sustituta. La dimensión gen contiene un campo de descripción, unos campos que determinan el inicio y el fin del gen en la secuencia, la hebra donde se encuentra el gen y la familia y subfamilia a la que pertenece. En la dimensión cromosoma se encuentra el organismo analizado y la división a la que pertenece y una descripción. La dimensión estructura contiene información acerca de la función molecular de dicha estructura, el número de unidades de información y un arreglo en el que se visualiza cada una de las unidades de información que conforman la estructura. La dimensión Band, contiene en términos generales, los intervalos y rangos de análisis de interés para los investigadores

**Fig. 2.** MODELO DIMENSIONAL ANÁLISIS FRACTAL




Fuente: Autores del proyecto

**Fig. 3.** *DATA MART* ANÁLISIS FRACTAL




Fuente: Autores del proyecto

La Tabla de hechos Análisis fractal tiene una granularidad definida por estructura, y presenta como medidas: el número de unidades de información (#UI), simetría y longitud de la estructura (size). Las medidas porcentaje de número de unidades de información (%gen#UI's) y porcentaje de longitud del gen (%gen_size), son relevantes al clasificar una de las medidas calculadas por los investigadores de AMGH, en intervalos y rangos. Además, la Tabla de hechos almacena otras medidas calculadas en el proyecto AMGH que no se muestran dado a una política de reserva sobre algunos resultados de la investigación [6].

## IV. PROBLEMAS PRESENTADOS EN EL DESARROLLO DE LA INVESTIGACIÓN

Los DSS están diseñados para tener en cuenta los sistemas operacionales (en este caso los archivos generados por las aplicaciones de cálculo fractal de AMGH), por lo tanto, para satisfacer las necesidades del proyecto, se presentaron algunos retos importantes, relacionados con: el proceso ETL, el diseño y construcción de la bodega de datos y la presentación de los datos.

### A. Problemas en el proceso ETL

El proceso ETL para esta investigación presentó algunas particularidades no soportadas por las herramientas de ETL tradicionales (transformaciones por medio de operadores básicos). Para poblar la bodega de datos no se utilizó como fuente de datos directa el GENBANK [2], debido a que los investigadores de AMGH obtenían la información de este repositorio público y realizaban diferentes procesos sobre estos datos, entre los que se encuentra la aplicación de técnicas de minería de datos. Por lo tanto, los datos fueron extraídos y cargados a la bodega de datos directamente desde los archivos generados por los investigadores de AMGH.

Debido a que el grupo de investigadores de AMGH realizó una limpieza exhaustiva de los datos del GENBANK (el porcentaje de los datos usados que presentaban error o ruido era insignificante), se presentaron pocos errores en el momento del cargue de los archivos proporcionados por los investigadores. Para resolver estos errores fue necesario realizar un análisis de los archivos de datos de los repositorios públicos basado en conceptos presentados en el taller de bases de datos: "*Issues in Biological Databases (DBiBD*) en el 2005 titula-

do “Una clasificación de los artefactos biológicos” [16]. En este estudio se presenta la naturaleza de los errores que se pueden presentar en estos archivos públicos y la razón por la cual no pueden ser cargados directamente en una base de datos unificada o en una bodega de datos. De acuerdo a la clasificación proporcionada en este estudio se encontraron dos tipos de errores que se presentaron durante el cargue de datos estos son:

- Se detectaron veintidós (22) registros erróneos de 29.200, con respecto a las descripciones de los genes en la dimensión gen. Dos de estos errores se relacionaron con abreviaturas y los 20 restantes con homónimos. Un homónimo se presenta cuando una o más secuencias diferentes reciben el mismo nombre, uno de los factores por lo que esto puede ocurrir es la existencia de secuencias que son comunes en diferentes organismos.

El problema con las abreviaturas se presenta cuando la abreviatura de una secuencia puede ser igual a muchas otras, por ejemplo, una secuencia denominada como GK, puede referirse a GLYCEROL KINASES, GLUTAMITE KINASES o GUANYLATE KINASES, entre otras. La presencia de homónimos y abreviaturas causa problemas en la identificación de secuencias y la búsqueda de palabras clave. Este tipo de error, hace referencia a un error de integridad referencial, que se presentó porque el archivo de origen sólo tenía las columnas de nombre del gen y descripción. Como un gen puede tener el mismo nombre, en el primer campo se podía encontrar varias veces el mismo nombre, esto hacía que al realizar la comparación no se lograra obtener un identificador único al cual asignarle la descripción. Para solucionar este problema, se reconstruyó el archivo que contenía las descripciones adicionándole el identificador del gen, que es único, de esta forma la comparación se podía hacer por ese identificador único.

- Otro error encontrado en los datos fue la violación de la estructura de la secuencia, encontrando entidades que no correspondían a la estructura lógica de una secuencia, por ejemplo, la estructura de la secuencia debe tener un número impar de unidades de información, al iniciar y finalizar el exón; si se encuentra un registro que no cumpla con estas características, se tiene un registro erróneo, por lo cual los registros que presentaron este error fueron eliminados.

## B. Problemas en el modelado de la solución

***Dimensión Rango*** (Tabla de Hechos Análisis Fractal): Para los investigadores de AMGH, es relevante generar intervalos y rangos de análisis dentro de los intervalos, de acuerdo a una variable establecida. Estos intervalos y rangos de análisis son seleccionados bajo criterios propios de los investigadores y a partir de estos se clasifica la variable de interés.

Al realizar el modelado de este requerimiento se presentaron algunas limitantes. A nivel teórico se encuentra parte de la solución a este problema de modelado, pero en la práctica las herramientas no soportan este caso de modelado.

Ralph Kimball [17], plantea para este caso de modelado una Tabla que permite la generación de reportes, de acuerdo a unos rangos de valores definidos por el usuario final, mostrando como ejemplo los saldos de las cuentas en una entidad bancaria (Ver Tabla II). En la Fig. 4, se presenta el modelo dimensional que permite que este tipo de consultas se pueda realizar [18]:

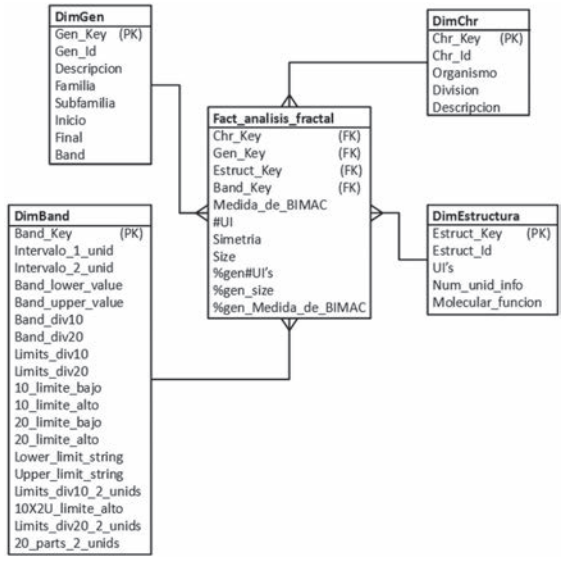
TABLA II.

EJEMPLO DE UN REPORTE POR INTERVALO

| Rango de Saldo | Número de Cuentas | Total de Saldos |
|---|---|---|
| 0 - 1.000 | 45.678 | $10.222.543 |
| 1.001 - 2.000 | 36.788 | $45.777.216 |
| 2.001 – 5.000 | 11.775 | $31.553.884 |
| 5.001 – 10.000 | 2.566 | $22.438.287 |
| 10.001 y más | 477 | $8.336.728 |

Fuente: Adaptado de [18]

**Fig. 4.** DIMENSIÓN BAND




Fuente: Tomado de [18]

En la solución propuesta por Kimball (ver Fig. 4), puede verse una particularidad: la Tabla que permite definir los rangos, no está asociada a la Tabla de hechos, lo cual se evidencia por la ausencia de una llave foránea en la Tabla de hechos, lo que involucra realizar un cruce entre la Tabla *BAND DEFINITION* y el hecho *PRIMARY MONTH ENDING BALANCE*, mediante el uso de un par de cruces de menor que y mayor que [18]. Esta solución no puede ser utilizada en herramientas de modelado dimensional, ya que en estas herramientas no se permite realizar una consulta que involucre una Tabla de hechos con una dimensión que no esté asociada a dicha Tabla.

En la Tabla III se muestra una alternativa de solución que toma como ejemplo la dimensión fecha. En ella se muestra la forma como se definen las diferentes agrupaciones que pueden existir para realizar los análisis, en este caso se puede agrupar la información por meses, por trimestres o por años, por medio de unos campos identificadores, que permiten agrupar la información según sea requerido, gracias a que asigna a cada fecha un conjunto de identificadores que permiten saber a qué mes, qué trimestre y qué año pertenece una fecha. De esta forma si se desea consultar los meses, lo que debe hacerse es agrupar por el identificador de mes, teniendo en cuenta que todos los registros que contengan el mismo identificador pertenecen al mismo mes [19].

TABLA III.
EJEMPLO DATOS DIMENSIÓN FECHA

| Fecha | Mes | Trimestre | Año | Nombre Mes | Nombre Trimestre |
|---|---|---|---|---|---|
| 02/01/2005 | 1 | 1 | 2005 | Ene | Trim_1 |
| 12/01/2005 | 1 | 1 | 2005 | Ene | Trim_1 |
| 10/02/2005 | 2 | 1 | 2005 | Feb | Trim_1 |
| 22/02/2005 | 2 | 1 | 2005 | Feb | Trim_1 |
| 22/04/2005 | 4 | 2 | 2005 | Abr | Trim_2 |
| 18/05/2005 | 5 | 2 | 2005 | May | Trim_2 |
| 20/06/2005 | 6 | 2 | 2005 | Jun | Trim_2 |
| 01/07/2005 | 7 | 3 | 2005 | Jul | Trim_3 |
| 04/11/2005 | 11 | 4 | 2005 | Nov | Trim_4 |

Fuente: Autor del proyecto

La forma como en la dimensión de fecha se manejan este tipo de agrupaciones, se convirtió en una mejor forma de manejar los intervalos que se usaron para el proyecto AMGH. Pero para aplicar este concepto se tuvo que tener en cuenta que los intervalos de análisis deben ser variables y no fijos como en el caso de la fecha (un año siempre tiene 365 días o 366 días en caso de los bisiestos).

En la Fig. 5 y en la Fig. 6, se puede observar que el intervalo que va de cero a uno puede dividirse de dos formas diferentes y que el intervalo con diez particiones contiene al que tiene veinte, igual que un trimestre contiene meses, por lo tanto, los intervalos fueron definidos de forma similar a como se hace en la dimensión fecha.

Para modelar los intervalos en la bodega de datos se usó la dimensión *Band* (ver Fig. 3), que contiene el valor del límite inferior y superior para la división más pequeña en cada intervalo, y el concepto que se usa para agrupar (similar a los meses, semestres y trimestre en la dimensión fecha). Además, un atributo que permite ordenar estas divisiones y un atributo para nombrar cada división de forma comprensible para el usuario final.

**Fig. 5.** INTERVALO DE 0 A 1 DIVIDIDO EN 10


Fuente: Autor del proyecto

**Fig. 6.** INTERVALO DE 0 A 1 DIVIDIDO EN 20


Fuente: Autor del proyecto

Para crear la Tabla de la dimensión *Band*, se recomienda seguir los siguientes pasos:

- Revisar cuáles son las agrupaciones y divisiones que se consideran relevantes para la investigación y listarlas, por intervalos de análisis y divisiones requeridas para cada uno de los intervalos.
- Determinar cuál es la división más pequeña que se pueda tener, o lo que es igual el máximo número de partes en que será dividido el intervalo más pequeño que se pueda tener en los intervalos de análisis. Esto determina el incremento más pequeño que se puede tener.
- Generar un conjunto de valores para la dimensión de la siguiente manera: se inicia desde el

menor valor posible (en este caso es 0), incrementar de acuerdo al valor resultante de la división que se realizó con anterioridad, para obtener así el siguiente valor hasta llegar al máximo valor requerido (en este caso 50). En resumen, al generar este conjunto de datos lo que se obtiene es el intervalo de mayor granularidad (más detalle). Estos valores son los que agrupados de diferentes formas permiten definir los intervalos y rangos de análisis.

- Con esta información se construyen las columnas de la dimensión, de acuerdo a los valores superior e inferior de cada división y los identificadores que permitan realizar las diferentes agrupaciones.

### C. Problemas en la presentación de datos

Se presentaron algunos inconvenientes en el momento de la presentación de los datos al usuario en la herramienta OLAP tradicional, ya que las consultas que solicitaban el grupo de investigadores, eran consultas especializadas, atípicas a las consultas realizadas en una bodega de datos tradicional. Los problemas presentados en la adaptación de la herramienta OLAP y las soluciones implementadas fueron las siguientes:

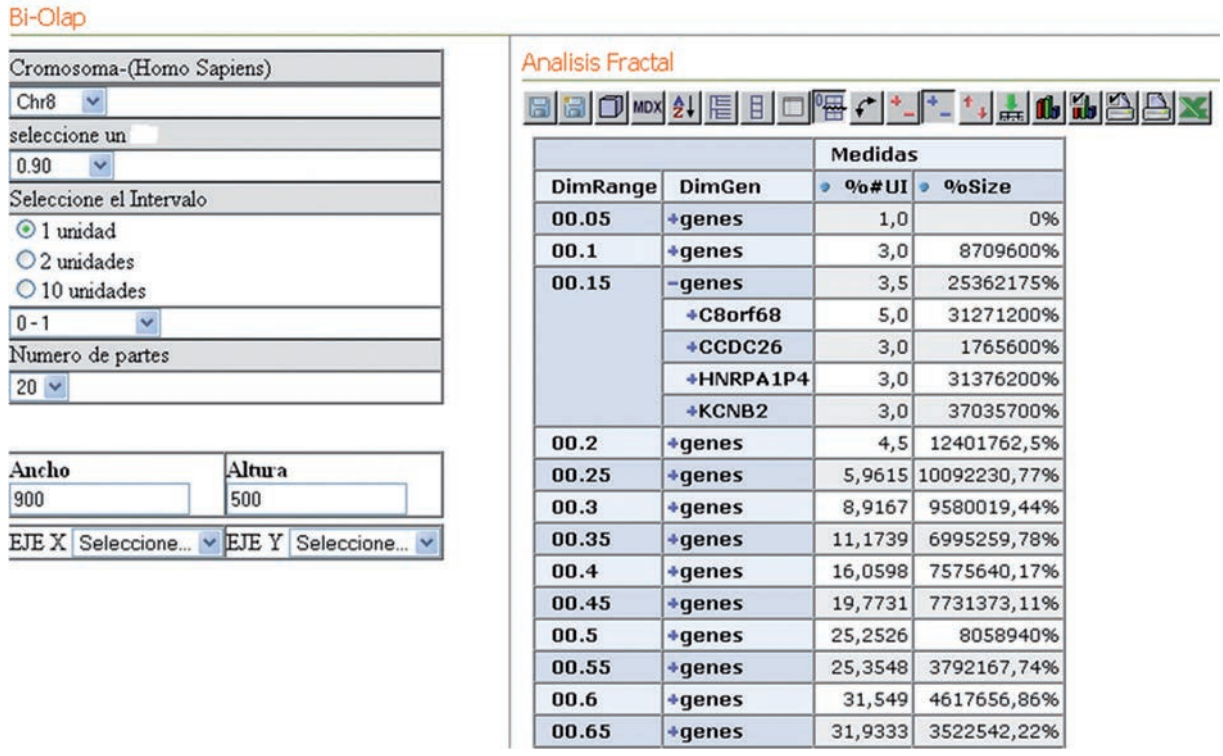
- Problemas Relacionados con los Filtros. Los filtros requeridos por los investigadores implicaban la capacidad de mostrar automáticamente valores que estén por encima o por debajo de un valor definido por el usuario, o la clasificación en rangos de las medidas, seleccionado un intervalo y el número de partes en las que se quiere dividir dicho intervalo. Para dar solución a este problema se adicionó funcionalidad a la aplicación creando los filtros por medio de una página que contenía las consultas MDX (MultiDimensional eXpressions) para realizar los filtros y que permitía que estos pudieran ser parametrizados.
- Problemas Relacionados con el Graficador. Las gráficas solicitadas requerían seleccionar las columnas que se deseaban graficar y el eje para cada una de las variables, sin embargo, los graficadores OLAP tradicionales no presentan estas opciones, las opciones gráficas se limitan a generar histogramas, tortas, líneas de relación, entre otros gráficos. Todas estas gráficas muestran únicamente la relación de dimensiones contra medidas, pero algunas de las consultas requeridas por los investigadores requerían gráficas de relaciones de medidas contra medidas. La solución a este problema consistió en: 1) adicionar un graficador externo (*Open Flash Chart*) a la herramienta OLAP tradicional usada (*jpivot* de Pentaho [20]), 2)dar la libertad de seleccionar columnas (atributos) que se deseaban graficar y los ejes de estas variables, y 3) usar consultas MDX parametrizadas en los filtros.
- Otro problema se presentó debido a que las dimensiones de los datos en una bodega de datos tradicional son típicamente cadenas de caracteres (por ejemplo: descripciones y nombres) mientras que en AMGH se tenían algunos valores numéricos y las herramientas OLAP los convertían automáticamente en cadenas de caracteres. Esto hacía que el visor les asignara un orden alfabético, cambiando el orden de presentación de los datos en los reportes. Para solucionar esto se adicionó una cantidad apropiada de ceros a la izquierda a estos valores, y de esta forma se logró conservar el orden numérico.

## V. HERRAMIENTA OLAP

El DSS propuesto en este artículo, se implementó en una suite Open Source, para el desarrollo de proyectos de inteligencia de negocios llamada *Pentaho* [20], debido a requerimientos de los investigadores del proyecto AMGH y a que en este tipo de comunidades académicas se utilizan herramientas Open Source para facilitar la publicación y distribución del conocimiento generado. A continuación se muestran algunas interfaces (screenshots) de una consulta sobre el data mart de Análisis Fractal y algunas gráficas producto de las consultas.

En la Fig. 8 se muestra una consulta que permite visualizar los genes clasificados en rangos y según sus unidades de información y sus longitudes. Al lado izquierdo de esta figura se puede observar el uso de filtros estándar (Cromosoma y Número de partes) y no estándar, para el manejo de los intervalos (dimensión Band). Al lado derecho superior se encuentra la barra de herramientas típica de un visor OLAP y en la parte central los datos del reporte solicitado al *data mart*.

Fig. 7. CONSULTA ANÁLISIS FRACTAL




Fuente: Autor del proyecto

En la Fig. 9 se muestra una gráfica con las unidades de información de todo el genoma humano expresado en el conteo de la medida fractal D (establecida por los investigadores AMGH) contra los genes clasificados en rangos.

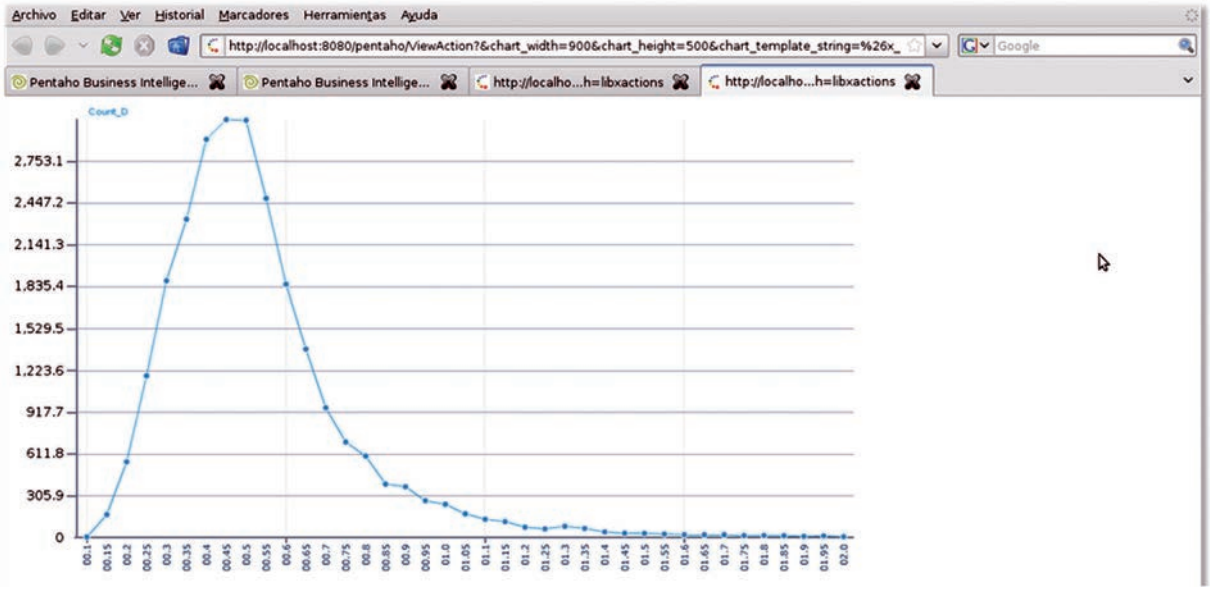
Finalmente, en la Fig. 10 se muestra una alternativa de visualización para el *data mart* Análisis fractal, donde los investigadores encuentran un conjunto de filtros que les permiten seleccionar un intervalo, dividirlo en un número de partes de acuerdo a su criterio, dicha selección es graficada contra las diferentes medidas del *data mart* para todos de los cromosomas simultáneamente, teniendo también la posibilidad de agregar o quitar cromosomas a la grafica.

## VI. CONCLUSIONES

En la revisión de los trabajos relacionados con los objetivos del presente trabajo se encontraron limitaciones a nivel de modelado y en las herramientas OLAP existentes para tratar datos bioinformáticos. El desarrollo del DSS presentado en este artículo muestra una alternativa viable para centralizar, almacenar y visualizar los datos provenientes del proyecto AMGH y se convierte en un referente para el uso de estas tecnologías en el área. Además, evidencia la necesidad de realizar más investigaciones que permitan acceder a soluciones de bodega de datos y OLAP en estos nuevos dominios de aplicación.

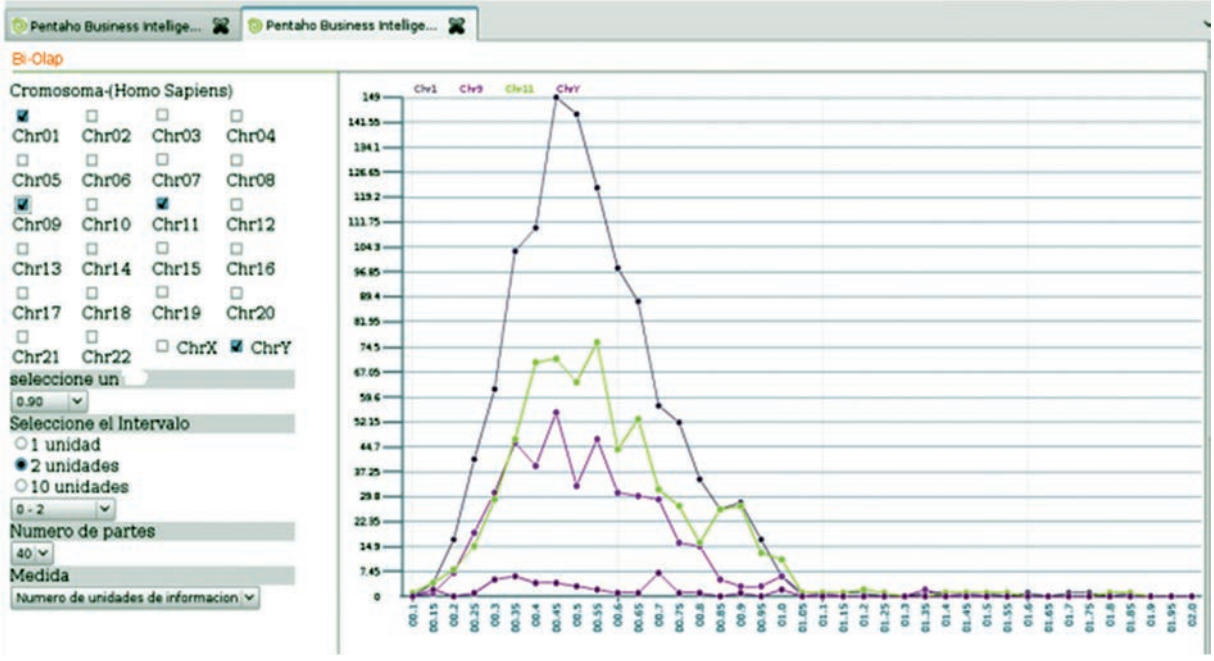
Las herramientas OLAP, deberían considerar ampliar el rango de posibilidades en cuanto a los análisis gráficos, permitiendo funcionalidades tales como: graficar medidas contra medidas y dar flexibilidad y facilidad a los usuarios de poder seleccionar las columnas que desean graficar. Facilidades que se incluyeron en el proyecto y mostraron su potencialidad. Además, ofrecer una mayor cobertura de funciones estadísticas para aplicar sobre los datos de la bodega de datos en el momento de la consulta. Esto permitiría entregar una mejor solución a las necesidades gráficas de este tipo de proyectos.

Fig. 8. RANGO VS. NÚMERO DE UNIDADES DE INFORMACIÓN TOTALES




Fuente: Autor del proyecto

Fig. 9. CROMOSOMAS VS 4 MEDIDAS




Fuente: Autor del proyecto

La solución planteada en esta investigación para modelar intervalos y rangos de estos intervalos, es una aproximación importante en el área de modelamiento multidimensional de datos, ya que permite obtener los resultados esperados por los investigadores aplicando conceptos de modelado ya existentes. Este trabajo puede ser una alternativa de solución para todos aquellos campos de aplicación (no solo en bioinformática) en los que sea una variable importante para los análisis de los usuarios, el manejo de intervalos y rangos de los mismos.

Los datos que se encuentran en el contexto biológico poseen particularidades que acarrean problemas y errores propios del área y no se encuentran reportados actualmente en la literatura tradicional de bodega de datos. Por lo anterior, es preciso tener en cuenta que para llevar a cabo un proceso exitoso de ETL en esta área, se deben entender los posibles errores que pueden llegar a presentar los datos de la fuente de datos -estos generalmente creados y almacenados en los repositorios públicos-, para mitigar el impacto que estos errores pueden llegar a tener en las consultas que arroje posteriormente la bodega de datos.

## VII. TRABAJO FUTURO

Ampliar el modelo actual al añadir nuevos atributos y medidas en la dimensión cromosoma que permitirá ubicar los genes con sus estructuras en un lugar determinado de los cromosomas y poder observar si existe alguna regularidad entre la posición de los genes, su función y su valor D. También añadir a la dimensión gen un atributo sobre oncogenes que permita observar con ayuda de la clasificación del valor D si existe alguna regularidad, información de interés para los investigadores del área.

La adición de nuevas dimensiones y tablas de hechos que completen el modelo biológico, como proteína, estructura de la proteína, *Alus, repeats,* la inclusión de la secuencia del gen, genoma y todas las posibles tablas de hechos que se puedan llegar a derivar, para de esta forma establecer que otras limitaciones a nivel de diseño e implementación se pueden presentar teniendo en cuenta las particularidades de análisis de los datos del área de la Bioinformática.

Extender el modelado de intervalos y rangos de estos intervalos, para hacerlo más dinámico, es decir, que permita que los intervalos y sus rangos puedan cambiar constantemente de acuerdo al criterio del investigador y la correspondiente implementación de una herramienta OLAP que permita visualizar adecuadamente estos intervalos.

## AGRADECIMIENTOS

El trabajo presentado en este artículo conto con la asesoría y apoyo del grupo de investigación BIMAC y GTI de la Universidad del Cauca y el grupo de investigación en Bioinformática de la Universidad del Valle. Se utilizaron recursos suministrados y financiados por COLCIENCIAS bajo el proyecto código 1103-12-16765.

## REFERENCIAS

[1] BERIS. (2012, July 2012). *Human Genome Project Information*. Available: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

[2] GenBank. (2012, Septiembre de 2012). *GenBank Overview*. Available: http://www.ncbi.nlm.nih.gov/genbank/

[3] EMBL. (2012, September 2012). *European Molecular Biology Laboratory*. Available: http://www.embl.org/

[4] DDBJ. (2012, October 2012). *DNA Data Bank of Japan*. Available: http://www.ddbj.nig.ac.jp/

[5] P. Vélez, "Propuesta para la participación en la convocatoria nacional para el concurso de proyectos de investigación programa nacional de biotecnología "Análisis Multifractal del Genoma Humano para la Búsqueda de Regularidades con Significado Biológico y una Contribución a la Generación de Biotecnología de la Información"," Universidad del Cauca 2004.

[6] P. Moreno*, et al.*, "The human genome: a multifractal analysis," *BMC Genomics,* Vol. 12, p. 506, 2011.

[7] P. E. Vélez*, et al.*, "The Caenorhabditis elegans genome: a multifractal analysis," *Genetics and Molecular Research,* Vol. 9, pp. 949-965, 2010.

[8] D. Florescu*, et al.*, "Database techniques for the World-Wide Web: a survey," *SIGMOD Rec.,* Vol. 27, pp. 59-74, 1998.

[9] S. B. Davidson*, et al.*, "Challenges in integrating biological data sources," *J Comput Biol,* Vol. 2, pp. 557-72, 1995.

[10] W. Sujansky, "Heterogeneous database integration in biomedicine," *Comput. Biomed. Res.,* Vol. 34, pp. 285-298, 2001.

[11] S. B. Davidson*, et al.*, "K2/Kleisli and GUS: experiments in integrated access to genomic data sources," *IBM Syst. J.,* vol. 40, pp. 512-531, 2001.

[12] J. Hammer and M. Schneider, "Genomics Algebra: A New, Integrating Data Model, Language, and Tool for Processing and Querying Genomic Information," presented at the 1st Biennial Conf. on Innovative Data Systems Research (CIDR), 2002.

[13] M. Fischer*, et al.*, "DWARF--a data warehouse system for analyzing protein families," *BMC Bioinformatics,* vol. 7, p. 495, 2006.

[14] S. P. Shah, *et al.*, "Atlas - a data warehouse for integrative bioinformatics," *BMC Bioinformatics,* vol. 6, p. 34, 2005.

[15] T. Kirsten, *et al.* (2004, Septiembre de 2012). *A Data Warehouse for Multidimensional Gene Expression Analysis*. Available: http://www.izbi.uni-leipzig.de/izbi/Working%20Paper/2004/01_geware.pdf

[16] J. L. Y. Koh, *et al.*, "A Classification of Biological Data Artifacts," presented at the ICDT Workshop on Database Issues in Biological Databases (DBiBD), Edinburgh, Scotland, UK, 2005.

[17] R. Kimball, *et al.*, *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing and Deploying Data Warehouses with CD Rom*: John Wiley & Sons, Inc., 1998.

[18] R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*: John Wiley & Sons, Inc., 2002.

[19] C. Imhoff, *et al.*, *Mastering Data Warehouse Design: Relational and Dimensional Techniques*: Wiley, 2003.

[20] Pentaho. (2012, July 2012). *Pentaho - Powerful Analytics Made Easy* Available: http://www.pentaho.com/