METODOLOGÍA DE EVALUACIÓN DE TÉCNICAS DE LA MINERÍA DE DATOS APLICADAS A DATOS BIOLÓGICOS

TESIS

Que para obtener el grado de

MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

presenta

Mara Alejandra León Pinto

Ivo Humberto Pineda Torres

Asesor de Tesis

Benemérita Universidad Autónoma de Puebla

Septiembre 2006

Contenido

Índice de tablas

Índice de figuras

Resumen

La Bioinformática es un nuevo campo en las ciencias de la computación que ha nacido de la necesidad de altos requerimientos de recursos computacionales, los cuales ayudan a organizar, analizar y almacenar información biológica. Se establece una metodología de evaluación de técnicas de minería de datos aplicadas a datos biológicos a través de herramientas de Bioinformática disponibles de manera pública.

Capítulo 1

Introducción

1.1. Introducción

Dada la complejidad de los sistemas biológicos, y la gran cantidad de investigación dedicada a las ciencias biológicas, el uso de las computadoras y los algoritmos asociados como ayuda en el manejo de la información es indispensable. Se requiere de análisis de datos para generar conocimientos generales e integrales.

El poder desarrollar algoritmos, bases de datos, interfaces de usuario, y herramientas estadísticas hace que los resultados sean potencialmente significativos. Estas nuevas herramientas dan la oportunidad de interpretar datos y asignar algún significado donde no existía.

Partiendo de una pregunta de interés biológico, la informática/computación da soporte a través de diferentes técnicas y herramientas tales como bases de datos, visualización, programas de búsqueda y comparación de patrones, servidores con la capacidad de procesamiento y almacenamiento, análisis exhaustivos, etc.

Este trabajo se encuentra estructurado de la siguiente manera:

Capítulo 1: Presenta una introducción conceptual del tema, antecedentes y situación actual.

Capítulo 2: Establece la vinculación entre la Minería de Datos y la Bioinformática, se definen las técnicas de minería de datos aplicadas a la Bioinformática.

Capítulo 3: Define las categorías de análisis para la evaluación de las técnicas de minería de datos en base a los requerimientos de las herramientas de Bioinformática.

Capítulo 4: Presenta la descripción de las herramientas de Bioinformática utilizadas para la evaluación de las técnicas de minería de datos y se definen los parámetros de evaluación de cada técnica.

Capítulo 5: Muestra los resultados obtenidos por cada herramienta después de la evaluación de las técnicas de minería de datos sobre las bases de datos biólogicas seleccionadas.

Capítulo 6: Presenta el análisis de los resultados y se establece la metodología a seguir para la evaluación de las técnicas de minería de datos sobre datos biológicos.

Capítulo 7: Presenta las conclusiones y propuestas para el trabajo futuro.

1.2. Descripción del Proyecto

1.2.1. Antecedentes del Proyecto

En la década de los años 50, con el descubrimiento de la estructura de doble hélice del ADN por el biofísico *Francis Crick* y el bioquímico *James Watson*, comenzó una nueva época en el desarrollo de la biología molecular. A mediados de los años 90, la prensa internacional difundió la noticia de la publicación del genoma humano, abriendo definitivamente el paso a la era genómica. La palabra *genómica* surgió producto de la fusión entre el prefijo *gen* y el sufijo *omica* que significa conjunto.

La disponibilidad de genomas completos, el volumen de información ubicado actualmente en las bases de datos públicas y los ambiciosos proyectos masivos de estudio sobre la interacción entre proteínas, ha generado un nuevo paradigma consistente en la aplicación de los métodos computacionales de análisis, procesamiento y almacenamiento de datos al área de la biología molecular. El enfoque clásico, que consistía en conocer una determinada función y buscar el gen responsable, se transformó y creó un nuevo escenario donde se dispone de un importante número de genes desconocidos a los que es necesario asignar una función. Esta nueva forma de análisis dio lugar al desarrollo de la *Bioinformática* [22]. La Bioinformática ocupa un papel central como el *elemento* que une a diversas áreas de la ciencia. Es por esto que cada vez más, grupos de investigación en México se unen a los grupos ya reconocidos internacionalmente para desarrollar nuevos proyectos en esta disciplina.

El doctor en ciencias biomédicas, Julio Collado Vides, al término de una estancia posdoctoral de tres años en Boston, regresó a México a instalar un laboratorio de biología computacional (Bioinformática), en el Centro de Investigación sobre Fijación de Nitrógeno (CIFI) que la UNAM tiene en Cuernavaca, Morelos. En ese laboratorio, y con la participación de otros colegas, el doctor Collado Vides se ha dedicado a desarrollar programas de cómputo derivados de su propia tesis doctoral, una teoría linguística de la regulación de la expresión genética modelo matemático-gramatical creado por el propio investigador.

Dicha metodología lo llevó a colaborar recientemente en el equipo internacional que descifró (mediante un reconocedor sintáctico) el genoma completo de *Escherichia Coli K-12*, bacteria que ocupa un lugar único en las ciencias biológicas como la célula autónoma sobre la que más conocimiento se dispone, y cuyo estudio incubó el nacimiento de la biología molecular. El 16 de febrero de 1997 se depositó en las bases de datos la secuencia completa del cromosoma, es decir, 4 millones 639 mil 221 nucleótidos de E. Coli. Los resultados fueron reportados en un artículo publicado por la revista Science [3] y fue la primera vez que un investigador y una institución mexicana participa en un artículo sobre un genoma completo.

Actualmente el segundo proyecto principal es el de coordinar la Bioinformática asociada al proyecto del genoma de *Rhizobium Etli*. Este trabajo se está realizando en colaboración con otros colegas del Centro de Investigación sobre Fijación de Nitrógeno (CIFI) y otras instituciones nacionales e internacionales. Este trabajo implica la anotación de las secuencias del genoma en número de marcos de lectura abiertos (ORFs) y las proteínas predichas; el análisis y la predicción de elementos reguladores; predicción de elementos repetidos, así como análisis evolutivo y comparativo con otras bacterias relacionadas.

En el estado de Puebla, el Centro de Investigación en Tecnologías de Información de la Universidad de las Américas (UDLA), en el grupo de Matemáticas Aplicadas y Computabilidad desarrollan el proyecto denominado *Bioinformática: análisis del DNA*.

1.2.2. Objetivos del Proyecto

El objetivo principal de este trabajo es hacer una evaluación teórico-práctica de técnicas de minería de datos mediante el uso de herramientas de Bioinformática disponibles de manera pública que permita establecer una metodología de evaluación de éstas técnicas aplicadas en datos de tipo biológico. De este objetivo se derivan los siguientes objetivos particulares:

- Realizar una investigación teórica de la Bioinformática y Minería de Datos, así como la vinculación que existe entre estas áreas.
- Realizar un estudio teórico de la aplicación de técnicas de Minería de Datos en el área de Bioinformática.
- Realizar un estudio teórico-práctico de las herramientas de Bioinformática disponibles de manera pública con el fin de utilizarlas para realizar pruebas en la aplicación de las técnicas de minería de datos sobre bases de datos biológicas.
- Elaborar una metodología que funcione como guía para los usuarios que deseen realizar una evaluación de técnicas de minería de datos aplicadas sobre datos biológicos.

Capítulo 2

Minería de Datos y Bioinformática

2.1. Historia de la Bioinformática

La historia de la bioinformática inicia a partir de la historia de la biología. En realidad son los biólogos y los bioquímicos quienes hacen su primer acercamiento a la tecnología computacional como elemento fundamental para su trabajo diario.

Desde los siglos XVIII y XIX, los biólogos se enfrentaron a problemas relacionados con el procesamiento masivo de la información. *Darwin*, por ejemplo en su viaje en el Beagle, recolectó y procesó manualmente multitud de datos sobre las especies. En aquellos tiempos, los taxonomistas catalogaron más de 50.000 plantas. El desarrollo de la genética con la formulación de las *Leyes de Mendel* hace más de 100 años y el descubrimiento de la estructura del ADN en 1953, abrieron las puertas de la investigación que desembocó en el proyecto *Genoma humano* en el año 1990. Desde los años 60, el crecimiento en el número de secuencias conocidas de aminoácidos de las proteínas impulsó la aplicación pionera de las computadoras en biología molecular.

El desarrollo de la genética como una disciplina científica, basada en claros principios como las *Leyes de Mendel* y el descubrimiento de la estructura del ADN condujo a nuevas investigaciones que crearon un volumen enorme de información que era necesario guardar y analizar. Así, al

principio de los años 60, el número creciente de secuencias de aminoácidos era uno de los factores principales que contribuyó al desarrollo de la biología computacional. En esta década también, aparecieron los primeros signos de una convergencia entre la biología, bioquímica, ingeniería e informática que conduciría después al nacimiento de la Bioinformática.

No obstante, el uso de las computadoras para la investigación biológica durante estos años no se reconocía como un elemento importante para la investigación en el laboratorio. El campo de la Bioinformática necesitaba un liderazgo y una financiación similar al que comenzaba a gestarse al mismo tiempo por los profesionales de la informática médica. Después, algunos investigadores mostraron que las computadoras podían acelerar dramáticamente la secuenciación y la determinación de estructuras de la proteína. Los métodos informatizados para la secuenciación del ADN empezaron a aparecer y los primeros bancos de datos de secuencias de proteínas se hicieron presentes [10].

Hacia finales de los años 80, comenzó a emplearse el término *Bioinformática*, aunque algunos pioneros habían aplicado las computadoras con éxito a los problemas de la biología molecular, incluso una década antes de que fuera posible la secuenciación del ADN. Entre estas aplicaciones, *Margaret Dayhoff* desarrolló los primeros programas para determinar la secuencia de aminoácidos de una proteína en 1965 y preparó el primer banco de datos de secuencias de proteínas que luego evolucionó para convertirse en PIR (Protein Information Resource) en 1983. Los programas de comparación de secuencias y de análisis filogenético fueron algunos de los primeros avances en este campo alrededor de los años 60.

El análisis estructural de las macromoléculas se inició por esos años, aunque limitado por las capacidades de la informática disponible en ese momento. A comienzos de los años 70, esos métodos se aplicaron al procesamiento de información sobre ácidos nucleicos. Se diseñaron programas para comparar secuencias. FASTA se desarrolló en 1985 aunque Genbank, el banco de datos de secuencias de ADN central se crea en 1980 y SwissProt, su homólogo para las proteínas empezó su actividad en 1987.

A finales de los años 80, se desarrollaron programas bioinformáticos en los centros académicos que rápidamente se convirtieron en productos comerciales, y se comenzaron a distribuir como paquetes integrados de herramientas para la administración de datos en el campo de la biología molecular.

Las mejoras en los sistemas computacionales permitieron el avance de las técnicas de aprendizaje automático con clara aplicabilidad en Bioinformática. Se aplicaron redes de neuronas artificiales, modelos de Markov ocultos o métodos de agrupamiento para analizar conjuntos de datos incompletos, con cierto ruido o error.

Desde el principio de los años 90, muchos laboratorios han estado analizando el genoma completo de varias especies tales como bacterias, levaduras, ratones y seres humanos. Durante estos esfuerzos de colaboración, se han generado cantidades enormes de datos los cuales se recogen y se almacenan en grandes bases de datos, la mayoría de las cuales son publicadas y accesibles.

Con el incremento en complejidad y capacidad tanto de las computadoras como de las técnicas de investigación, se necesitan *puentes* humanos que puedan entender ambas disciplinas y sean capaces de comunicarse con los expertos de los dos campos.

Hoy, algunos de los problemas más importantes de la biología moderna y la genómica son imposibles de resolver sin el poder de cálculo de las computadoras. Los programas de búsqueda y anotación de genes son muy importantes para completar el proyecto *Genoma humano* porque permiten, entre otras cosas, llevar a cabo la secuencia de datos estructurales a nivel molecular y obtener la infraestructura de secuenciación y computación, pilares para el desarrollo de un proyecto genómico.

El número de estructuras a nivel molecular se dobla cada dos años. Las técnicas como la comparación de pares de secuencias biológicas, alineación múltiple, análisis filogenético o búsquedas por similitud en bases de datos, facilitan el trabajo de los biólogos ocupados en tareas de identificación de genes o en la predicción de su estructura y función. De esto resulta que la Bioinformática suscite una atención creciente durante los últimos años [13].

2.2. Relación entre la Biología y la Informática

Existen tres subdisciplinas en las que se unen la biología y la informática, pero con objetivos y metodologías diferentes [16]:

Bioinformática o biología molecular computacional: investigación y desarrollo de la infraestructura y sistemas de información y comunicaciones que requiere la biología molecular y la genética (redes y bases de datos para el genoma, microarrays, informática aplicada a la biología molecular y la genética).

Biología computacional: computación que se aplica al entendimiento de cuestiones biológicas básicas, no necesariamente en el nivel molecular, mediante la modelización y simulación (ecosistemas, modelos fisiológicos, informática y matemáticas aplicadas a la biología).

Biocomputación: desarrollo y utilización de sistemas computacionales basados en modelos y materiales biológicos (biochips, biosensores, computación basada en ADN, redes de neuronas, algoritmos genéticos, biología aplicada a la computación).

Existen múltiples definiciones sobre la Bioinformática, la figura 2.1 muestra la intersección de ciencias como las matemáticas y estadística, las ciencias biomédicas, la biología molecular, las ciencias físicas y las ciencias de la computación para la creación de la Bioinformática.

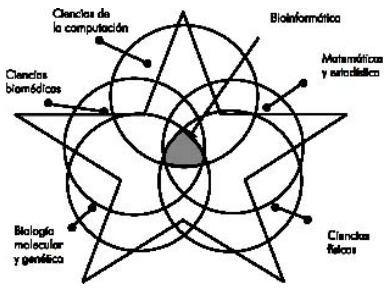


Figura 2.1: Creación de la Bioinformática

Una de las definiciones más completas es la que la refiere como una disciplina científica que se interesa por todos los aspectos relacionados con la adquisición, almacenamiento, procesamiento, distribución, análisis e interpretación de la información biológica, mediante la aplicación de técnicas y herramientas propias de las matemáticas, la biología y la informática, con el propósito de comprender el significado biológico de una gran variedad de datos. Se pueden distinguir además las siguientes definiciones [5]:

* Bioinformática es 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.

* Bioinformática es un área de investigación multidisciplinaria, la cual puede ser ampliamente definida como la interfaz entre dos ciencias: biología y computación y está impulsada por la incógnita del genoma humano y la promesa de una nueva era en la cual la investigación genómica puede ayudar dramáticamente a mejorar la condición y calidad de vida humana.

* Bioinformática es el uso de las matemáticas y de las técnicas informáticas para resolver problemas biológicos, normalmente creando o usando programas informáticos, modelos matemáticos o ambos. Una de las principales aplicaciones de la bioinformática es la simulación, la minería de datos (data mining) y el análisis de los datos obtenidos en los proyectos genoma o el proteoma.

2.3. Vinculación a la Minería de Datos

Desde hace algunos años han ocurrido avances espectaculares en las ciencias biomédicas como resultado del proyecto Genoma Humano. Las nuevas tecnologías, basadas en la genética molecular y la informática, son claves para este desarrollo, pues ellas suministran potentes instrumentos para la obtención y el análisis de la información genética. En los últimos años, la minería de datos (data mining) ha experimentado un auge como soporte para las filosofías de la gestión de la información y el conocimiento, así como para el descubrimiento del significado que poseen los datos almacenados

en grandes bancos. Esta permite explorar y analizar las bases de datos disponibles para ayudar a la toma de decisiones; además de facilitar la extracción de la información existente en los textos, así como crear sistemas inteligentes capaces de entenderlos, a esto se denomina comúnmente como minería de textos (text mining). Se describen sintéticamente los componentes básicos de la minería de datos y su aplicación en una emergente y trascendental actividad científica: *la Bioinformática*.

2.3.1. Aplicación de la Minería de Datos en Bioinformática

El aumento continuo de la disponibilidad de datos convierten en imprescindible el empleo de técnicas y herramientas que le den sentido y utilidad a la información existente. El surgimiento de técnicas como la minería de datos está asociado con la necesidad de procesar y analizar grandes volúmenes de datos, a fin de obtener información mediante la consolidación de los datos y conocimiento útil [12].

Algunas definiciones de Minería de Datos se muestran a continuación:

- Minería de datos, es descubrimiento eficiente de información valiosa no obvia de una gran colección de datos.
- Proceso que permite transformar información en conocimiento útil a través del descubrimiento y cuantificación de relaciones en una gran base de datos.
- Proceso de extracción de información y patrones de comportamiento que permanecen ocultos entre grandes cantidades de información.
- Minería de datos es la exploración y análisis de datos a través de medios automáticos y semiautomáticos con el fin de descubrir patrones y reglas significativos.

La bioinformática se encuentra en la intersección entre las ciencias de la vida y de la información, proporciona las herramientas y recursos necesarios para favorecer la investigación biomédica. Como campo interdisciplinario, comprende la investigación y el desarrollo de sistemas útiles para

entender el flujo de información desde los genes a las estructuras moleculares, su función bioquímica, su conducta biológica y, finalmente, su influencia en las enfermedades y en la salud [25].

Las motivaciones principales para el desarrollo de la bioinformática son:

- El enorme volúmen de datos generados por los distintos proyectos como el genoma humano y de otros organismos.
- Los nuevos enfoques experimentales, basados en biochips, que permiten obtener datos genéticos a gran velocidad, bien de genomas individuales (mutaciones, polimorfismos) o de enfoques celulares (expresión génica).
- El desarrollo de internet, que permite el acceso universal a las bases de datos de información biológica.

La bioinformática se ocupa de la utilización y almacenamiento de grandes cantidades de información biológica, es decir, trata del uso de las computadoras para el análisis de la información biológica, entendida ésta como la adquisición y consulta de datos, los análisis de correlación, la extracción y el procesamiento de la información. Muchos de los métodos de la computación y de las ciencias de la información sirven para estos fines, incluyendo el aprendizaje de las máquinas, las teorías de la información, la estadística, la teoría de los gráficos, los algoritmos, la inteligencia artificial, los métodos estocásticos, la simulación, la lógica, etc.

La magnitud de la información que generan las investigaciones realizadas sobre aspectos biológicos es tal que, probablemente, supera la generada por otras investigaciones en otras disciplinas científicas. Ante tal situación, uno de los retos de la bioinformática es el desarrollo de métodos que permitan integrar los datos genómicos (de secuencia, de expresión, de estructura, de interacciones, etc.). Dicha integración, sin embargo, no puede producirse sin considerar el conocimiento acumulado durante decenas de años, producto de la investigación de miles de científicos, recogido en millones de comunicaciones científicas.

El problema se encuentra en la carencia de herramientas bioinformáticas adecuadas para el análisis y gestión de los datos, precisamente por el enorme volumen de datos que se generan. Las técnicas de la minería de datos se emplean en el manejo de grandes volúmenes de información estructurada y almacenada en bases de datos, esto resalta la necesidad de emplear estas técnicas como la mejor forma de obtener conocimiento a partir de los resultados experimentales. El reto en la construcción de bases de datos es el establecimiento de una arquitectura que permita la realización de búsquedas inteligentes, la comunicación con otras bases de datos y la unión con herramientas de análisis y de minería de datos específicas, que permitan responder a problemas biológicos concretos.

La minería de datos es fundamental en la investigación científica y técnica, como una herramienta de análisis y descubrimiento de conocimiento a partir de la observación de datos o de resultados de múltiples experimentos. En este sentido, es necesario continuar elaborando herramientas computacionales apropiadas para su uso en varios proyectos y elevar con ello el nivel de conocimientos sobre su utilidad para los investigadores, el producir software que se ajuste a las necesidades de cada usuario implica la posibilidad de introducirse en un mercado con grandes perspectivas de futuro [15].

2.4. Técnicas de Minería de Datos y Bioinformática

En estos momentos, la mayoría de los proyectos que se desarrollan en el mundo en materia de genómica y proteómica, demandan la aplicación de técnicas de minería de datos para poder determinar que es realmente importante dentro del enorme volumen de información que se genera diariamente en el mundo.

Las técnicas de minería de datos se emplean para mejorar el rendimiento de procesos en los que se manejan grandes volúmenes de información estructurada y almacenada en bases de datos. Mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los datos.

Las técnicas de minería de datos más comúnes son las que se mencionan a continuación:

Agrupamiento, Asociación, Secuenciamiento, Reconocimiento de Patrones, Previsión, Simulación, Optimización, Clasificación, Normalización, Preprocesado, Métodos estadísticos, Métodos basados en árboles de Decisión, Reglas de Asociación, Redes Neuronales, Algoritmos Genéticos, Lógica Difusa, Series Temporales, Redes Bayesianas, Inducción de Reglas, Sistemas Basados en el Conocimiento y Sistemas Expertos, Algoritmos Matemáticos [9].

Aunque existen diversas técnicas de minería de datos, no todas son aplicables para resolver problemas de tipo biológico, las siguientes técnicas son aquellas que más se utilizan en aplicaciones de Bioinformática:

- Normalización.
  - Normalización LOWESS.
- Agrupamiento.
  - Supervisado.
  - No supervisado.
- Algoritmos de Distancia.
- Técnicas de Visualización
- Métodos de Proyección

Se describen brevemente estas técnicas con el propósito de entender la aplicación de éstas en el campo de la bioinformática.

Normalización

Hay muchas fuentes de variación sistemática en experimentos de microarrays, ellos pueden afectar las mediciones de los niveles de expresión genética. *Normalizacion* es el término usado para describir el proceso de remover tal variación, permite extraer la información biológica de los datos brutos. Es decir, se trata de remover el impacto del efecto de la tecnología de microarrays en datos de microarrays [18].

El proceso de normalización es que una vez obtenida la matriz de datos de la imágen escaneada, el próximo paso es aplicar técnicas de preprocesamiento para limpiar los datos (eliminar o substituir valores perdidos, identificar datos faltantes) y normalizar los datos. Después de este paso seguiría el análisis estadístico de los datos tal como lo muestra el esquema de la figura 2.2.

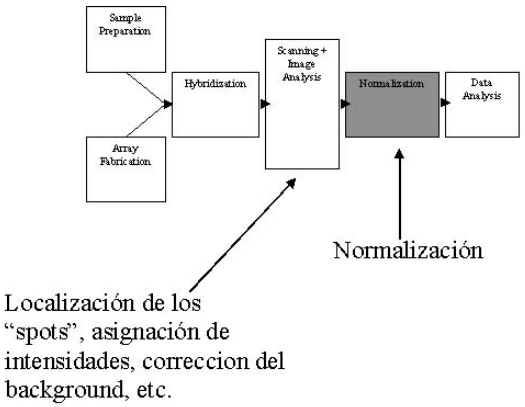


Figura 2.2: Ubicación del proceso de normalización en un experimento con microarrays

Métodos de Normalización

La mayoria de los métodos de normalización hacen uso de la siguiente suposición: El promedio de las razones C_y5/C_y3 es 1 o equivalentemente, el promedio del logaritmo de las razones es 0. Es decir, que el gen promedio no cambia su expresión bajo la condición que está siendo estudiada. Esto es debido a que sólo entre el 10 y 20 % de los genes son ex-

presados al mismo tiempo. O sea que en un plot de C_y5 versus C_y3 los puntos deben estar alrededor de una línea con pendiente 1. O equivalentemente los puntos en el plot de $M = log_2(C_y5/C_y3) = log_2(C_y5) - log_2(C_y3)$ versus $A = (log_2(C_y3) + log_2(C_y5))/2$ deben estar alrededor de una línea horizontal (M es por minus y A es por Add). Cualquier valor negativo de C_y3 o C_y5 producirá valores perdidos para M y A y el spot correspondiente será excluido de los análisis posteriores incluyendo normalización.

Dependiendo del experimento que se está llevando a cabo hay tres tipos distintos de normalización: normalización dentro del slide, normalización de slides pareados y normalización de multiples slides. En cada una de estas situaciones hay que tomar una decisión con respecto al conjunto de genes que se deben usar para la normalización. Hay tres caminos a seguir: el primer camino envuelve usar todos los genes del microarray, el segundo es usar solamente un subconjunto de genes con un nivel de expresión constante cuando es expuesto a distintas condiciones y la última alternativa es usando elementos de control.

Una vez que se ha elegido el conjunto de genes donde aplicar la normalización se calcula un valor de normalización o una función de normalización usando el conjunto elegido para luego aplicarlo a todo el microarray. Para ello existen algoritmos para calcular el factor o función de normalización y uno de ellos se conoce como normalización dependiente de la intensidad (LOWESS).

Normalización LOWESS

En este caso se genera una función de normalización usando los genes seleccionados. Esta función depende de la intensidad y generalmente es obtenida ajustando el suavizador no paramétrico LOWESS (regresión local ponderada) al plot del logaritmo de las razones (M) versus el promedio de los logaritmos de las intensidades (A) puede ser en forma global o en cada sector del grid. Luego se aplica esta normalización a todo el microarray [19].

Este algoritmo considera que la corrección que debe aplicarse a los datos brutos es función de la intensidad de las señales analizadas. El método consiste en ajustar los datos brutos a curvas de regresión locales que se establecen dentro de unas gráficas de correlación entre el ratio bruto de expresión génica (log_2R) y la media de las intensidades brutas en los dos canales ($A = 1/2[log_2Cy5 + log_2Cy3]$).

Agrupamiento

El término agrupamiento surge de los trabajos de Michalsky (1980) donde se propone encontrar a partir de una colección de datos, no sólo las clases en las que éstos se estructuran, sino además conformar las explicaciones de tales agrupamientos. El problema se muestra así compuesto por dos tareas fundamentales y relacionadas entre sí: el agrupamiento de entidades, en el que se determina e identifican subconjuntos útiles de una muestra de objetos, y la caracterización, la cual determina una explicación o concepto para cada subconjunto descubierto.

El objetivo de encontrar agrupamientos en un conjunto de datos es el poder describirlos en términos de clases o grupos de datos con fuertes semejanzas internas. Existen dos clases de agrupamientos:

- Agrupamiento supervisado: se basa en la idea de que para la clasificación de la mayoría de muestras ya existe información preliminar que puede utilizarse para la agrupación de nuevos datos en clusters.
- Agrupamiento no supervisado: conjunto de técnicas que agrupan los datos en función de una distancia sin utilizar ningún tipo de información externa para organizar los grupos. Dependiendo de la forma en que los datos son agrupados se pueden distinguir dos tipos:
    1. Jerárquicos: asumen que los datos se pueden agrupar de forma natural mediante una estructura de árbol (dendograma). Ejemplos: agrupamiento jerárquico aglomerativo y agrupamiento jerarquico divisivo.

2. No Jerárquicos: pretenden formar particiones naturales de los datos en un número de grupos (clases) posiblemente prefijado. Ejemplos: k-means, fuzzy c-means y SOM.

Algoritmos de Distancia

Cuando se trabaja con algoritmos de agrupamiento se necesita una medida de similaridad que permita evaluar las diferencias y similitudes entre los datos (genes). La estrategia más común consiste en medir las similitudes en términos de la distancia entre los pares de genes. La medida de similitud que se utiliza con mayor frecuencia es la distancia Euclidiana.

Distancia Euclidiana: Es un procedimiento de clasificación supervisada que utiliza esta distancia para asociar un gen a una determinada clase. En el entrenamiento supervisado, se definen los agrupamientos que representan las clases. En la clasificación, cada gen será incorporado a un grupo a través del análisis de la medida de similitud de la distancia Euclidiana, la cual esta dada por:

$$d(x, y) = \sqrt{\sum_{i=1}^{M} (x_i - y_i)^2} \tag{2.1}$$

Donde:
x = gen que está siendo probado.
y = media de un grupo.

El clasificador compara la distancia Euclidiana del gen a la media de cada grupo. El gen será incorporado al grupo que presenta la menor distancia Euclidiana [17].

Distancia de Manhattan: Conocida también como la distancia de Calles Urbanas entre dos objetos, es la suma de las diferencias absolutas en los valores para cada variable, está dada por:

$$d(x,y) = \sum_{i=1}^{M} |(x_i - y_i)| \tag{2.2}$$

Coeficiente de correlación de Pearson: El coeficiente de correlación de Pearson *(r)* es un índice estadístico que permite definir la relación entre dos variables, se mide en una escala de 0 a 1, tanto en dirección positiva como negativa. Un valor de 0 indica que no hay relación lineal entre las variables, un valor de 1 indica una correlación positiva perfecta y un valor -1 indica una correlación negativa perfecta entre dos variables. Normalmente, el valor se ubicará en alguna parte entre 0 y 1 o entre 0 y -1. Cuanto más cercanos al 0 sean los valores, indican una mayor debilidad de la relación o incluso ausencia de correlación entre las dos variables. Su cálculo se basa en la expresión:

$$\gamma = \frac{C(xy)}{\sigma_x \cdot \sigma_y} \tag{2.3}$$

Donde:
(x y) = covarianza de las dos variables.
$\sigma_x \cdot \sigma_y$ = producto de las desviaciones típicas de las dos variables.

Coeficiente de correlación de rangos de Spearman: El coeficiente de correlación de Spearman, ρ (rho), es una prueba no paramétrica que mide la asociación entre dos variables discretas. Para calcular ρ, los datos son ordenados y reemplazados por su respectivo orden. Se consideran, n objetos clasificados según dos variables o criterios. Por ejemplo, supongamos dos variables x e y que toman n valores emparejados $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. Se definen los rangos sobre cada una de las variables, de modo que se emparejan
$(\gamma_{x1}, \gamma_{y1}), (\gamma_{x2}, \gamma_{y2}), \cdots, (\gamma_{xn}, \gamma_{yn})$.
Se definen las diferencias $d_i = (\gamma_{xi} - \gamma_{yi})$, es decir, las diferencias de la posición del individuo i-ésimo según la clasificación (rango) dada por x y la clasificación (rango) dada por y.

El coeficiente de correlación viene dado por la expresión:

$$\gamma_s = 1\frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{2.4}$$

Donde:

d_i = diferencia entre los rangos de x e y.

n = número de parejas.

Técnicas de Visualización

La visualización se dedica a la transformación de datos científicos y abstractos en imágenes, es la generación de una imagen mental o una imagen real de algo abstracto o invisible. El propósito de la visualización es analizar, comprender y comunicar la información que viene contenida en datos. Muchas de las técnicas de visualización y sus procesos utilizan varios pasos en forma simultánea por lo que la automatización de la conversión y lectura transparente de datos en más que deseable. Estas técnicas son buenas para ubicar patrones en un conjunto de datos y pueden ser usadas al comienzo de un proceso de minería de datos para identificar la calidad del conjunto de datos.

Ejemplos de técnicas de visualización son el dibujo de diagramas para visualizar la relación que existe entre un grupo de datos o gráficos bidimensionales o tridimensionales que muestren el comportamiento de los datos y en algunos casos incluso se usan animaciones, en este trabajo se han usado dos tipos de visualización de datos, los dendogramas generados por las técnicas de agrupamiento jerarquico y k means y las gráficas bidimensionales generadas por el PCA.

Un problema con la visualización es el número de dimensiones de los datos. Puede suceder que se quiera visualizar más de cuatro características de un conjunto de datos, al menos se

tienen que visualizar cuatro dimensiones que muestren esas características y una imagen por lo regular sólo tiene dos dimensiones. La pregunta es como reducir este problema sin que se pierdan datos importantes.

Métodos de Proyección

Los métodos de agrupamiento reducen la cantidad de datos agrupándolos. Existen métodos que pueden ser usados para reducir la dimensionalidad de los datos y presentar estos datos en un sistema dimensional más bajo. Estos métodos tratan de encontrar el patrón total de los datos para proyectarlos y/o predecirlos a futuro.

Algunos de estos métodos son:

Escalamiento Muldidimensional (MDS)

El análisis de escalamiento multidimensional, MDS (Muitidimensional Scaling), consiste en un conjunto de modelos y métodos de análisis de datos cuya finalidad consiste en obtener la estructura subyacente de los datos, además de una representación geométrica de los mismos en un espacio de mínima dimensionalidad, de forma que sea accesible por simple inspección visual [1].

El análisis de escalamiento multidimensional tiene una doble finalidad:

1. Obtener la estructura subyacente de los datos.
2. Obtener una representación geométrica de los mismos en un espacio de mínima dimensionalidad, de forma que sea accesible por simple inspección visual.

Básicamente el análisis MDS consiste en lo siguiente: en un conjunto de n objetos se define una medida de proximidad que cuantifica el grado de (di)similaridad o (de)semejanza

entre cada par de objetos. Como medidas de proximidad suelen utilizarse la frecuencia o proporción de veces en que dos objetos aparecen conjuntamente, correlaciones, etc. A partir de esta medida de proximidad se obtiene una matriz de similaridades o disimilaridades. Se dice que esta matriz es de similaridad si a mayor semejanza entre dos objetos le corresponde mayor puntuación y, por el contrario, se dice que es de disimilaridad si a mayor semejanza le corresponde menor puntuación.

Cada objeto será representado como un punto en un espacio que es generalmente (pero no necesariamente) euclídeo, y las disimilaridades entre los objetos, d_{ij}, serán representadas por las distancias, d_i, entre los puntos que representan esos mismos objetos de forma que se preserve la ordinalidad de los datos.

Promedios Móviles

El método de promedios móviles es útil para suavizar ciertos factores estacionales, cíclicos o aleatorios, lo cual permite ver el patrón de tendencia de los datos. El principio detrás de los promedios móviles es calcular una media aritmética de los datos a partir de un grupo de periodos, y luego, calcular la siguiente media aritmética descartando los datos del periodo más antiguo y agregando los datos de un nuevo periodo, sólo sirve para pronosticar un sólo período: el siguiente. Se debe especificar el número de observaciones que se tomarán; se llama móvil porque siempre se toman las N últimas observaciones para hacer el pronóstico. Este método presenta desventajas como la pérdida de datos del inicio y del final de los grupos [14].

Se pueden considerar promedios móviles simples y promedios móviles lineales. En el primer caso se toman los N últimos datos y se calcula el promedio; en el segundo caso se construyen además promedios de los promedios y con ellos se establece una ecuación lineal que permite elaborar el pronóstico. Para el caso de los promedios móviles simples, algebráicamente se representa así:

$$F_{t+1} = S_t = (X_t + X_{t-1} + ... + X_{t-N+1})/N \tag{2.5}$$

Donde:

F_{t+1}= pronóstico para el tiempo t+1.

S_t= valor suavizado en el tiempo t.

X_t= valor actual en el tiempo i.

t= periodo de tiempo.

N= número de observaciones en el promedio.

Este método puede utilizarse cuando se sabe que los datos son estacionarios. La ventaja sobre el promedio total es que permite ajustar el valor de N para que responda al comportamiento de los datos.

Suavización Exponencial

El suavizamiento exponencial se refiere a una clase de métodos en los que el valor de una serie de tiempo en algún punto de ese tiempo es determinado por valores del pasado de la serie de tiempo. La importancia de los valores del pasado declinan exponencialmente cuando ellos envejecen. Este método es similar a los de movimientos de promedios sólo que, con el suavizamiento exponencial, los valores del pasado tienen diferente peso y todos los valores del pasado contribuyen de alguna manera al pronóstico.

Los métodos de suavizamiento exponencial son útiles para pronósticos a corto plazo. Ellos pueden producir a menudo buenos pronósticos para uno o dos periodos en el futuro. La ventaja del suavizamiento exponencial está en su aplicación relativamente simple para obtener pronósticos rápidamente, cuando se opera con un gran número de datos. Por esta razón, este método ha encontrado gran aplicación en el inventario de pronósticos. El suavizamiento exponencial no debe usarse para pronósticos a mediano o largo plazo. Tales pronósticos dependen

fundamentalmente de los datos más recientes, y por lo tanto, éste tiende a responder bien en el corto plazo y muy pobremente en el largo [26]. La ecuación para obtener el pronóstico es:

$$F_{t+1} = \alpha X_t + (1 - \alpha) F_t \tag{2.6}$$

Donde:

F_{t+1}= pronóstico para el periodo t+1.

α= primer factor de ponderación, $0 < \alpha < 1$.

X_t= valor real en el tiempo i.

$1 - \alpha$= segundo factor de ponderación, $0 < 1 - \alpha < 1$.

F_t= pronóstico para el periodo t.

t= periodo de tiempo.

Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales es una técnica estadística de síntesis de la información, o reducción de la dimensión (número de variables). Es decir, ante un banco de datos con muchas variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible. Se trata de encontrar la mejor representación bidimensional posible de los datos, es decir, aquella que es capaz de dar la mayor información de ellos. En algunos casos la representación puede llegar al 90 % mientras que en otros no pasa del 30 %, esto indica el grado de orden de los datos representados (la información que contiene).

El objetivo del PCA es hallar las direcciones que explican la máxima variabilidad de los datos y utilizarlas como nuevos ejes de cordenadas denominados componentes principales (CP). De esta forma se reduce la dimensionalidad de un espacio de k dimensiones a uno de a dimensiones ($a < k$), manteniendo intacta la información relevante del conjunto de datos.

Capítulo 3

Categorías de Análisis para la Evaluación

Antes de iniciar la evaluación de las técnicas de minería de datos en las herramientas de bioinformática, se debe establecer una secuencia sobre la manera en la que se efectuará la evaluación, esta secuencia debe incluir todos los aspectos o factores necesarios para que la evaluación se realize de manera completa. Por ejemplo, se debe elegir el tipo de dato biológico para evaluar las técnicas, se debe seleccionar una técnica que pueda ser aplicada para ese tipo de dato e incluso analizar la complejidad de cada técnica para evaluarlos. Por ello se han establecido tres categorías de análisis para la evaluación que nos permitirán tener una visión más clara de lo que se quiere obtener, estas categorías se han elegido de acuerdo al tipo de dato, a la técnica de minería de datos y a la complejidad de cada una de esas técnicas.

3.1. Datos

Las distintas áreas de la bioinformática determinan el tipo de dato utilizado por cada herramienta. En este trabajo se han utilizado datos de tipo biológico enfocados en el área de la biología molecular.

3.1.1. Datos biológicos

A pesar de que existen múltiples datos sobre los cuales se pueden obtener resultados de experimentos de tipo biológico, tales como proteínas, aminoácidos, moléculas, nucleótidos, etc., se han considerado aquellos basados en genes obtenidos de experimentos representados a través de microarrays [20]. Para entender mejor este tipo de datos, se presentan las siguientes definiciones:

ADN

ADN es la abreviatura del ácido desoxirribonucleico. Constituye el material genético de los organismos. Es el componente químico primario de los cromosomas y el material del cual los genes están formados.

Gen

Un gen es una secuencia lineal de nucleótidos de ADN o ARN que es esencial para una función específica, bien sea en el desarrollo o en el mantenimiento de una función fisiológica normal. Es considerado como la unidad de almacenamiento de información y unidad de herencia al transmitir esa información a la descendencia. La realización de esta función no requiere de la traducción del gen ni tan siquiera de su trascripción. Los genes están localizados en los cromosomas en el núcleo celular y se disponen en línea a lo largo de cada uno de los cromosomas. El conjunto de genes de una especie se denomina genoma.

Microarray

Un Microarray es una serie ordenada de ADN (10,000 genes apróximadamente), son conocidos como chips de ADN y se representan por medio de slides de vidrio o siliconas que contienen miles de moléculas distintas de ADN que pueden representar:

* Genes expresados por un tipo celular particular.
* Todos los genes del organismo.
* Genes seleccionados para investigar.

Un Microarray es una nueva manera de estudiar como interactúan entre sí un gran número de genes y como las redes regulatorias de la célula controlan enormes baterías de genes simultáneamente. Esta técnica crea las micromatrices utilizando un computador para aplicar con alta precisión, gotas minúsculas que contienen ADN de genes sobre un portaobjetos. Luego las placas se hibridizan con ADN complementario marcado en forma fluorescente o radiactiva y un computador mide la intensidad de cada punto fluorescente o radioactivo. Con esto se puede saber que tanto de un fragmento de ADN se encuentra presente y en ciertos tipos de micromatrices también es un indicador de la actividad de un gen específico.

3.1.2. Bases de datos biológicas

Una base de datos biológica es un volumen grande y consistente de datos, generalmente persistente; asociado a herramientas computacionales diseñadas para actualizar, consultar y devolver una parte o la totalidad de tales datos [27].

Una base de datos simple puede ser un archivo de texto que contiene varias entradas delimitadas por un formato específico. Por ejemplo, una entrada o registro en una base de datos de secuencias nucleotídicas puede incluir, además de la secuencia como tal, información acerca del organismo a partir del cual ésta fue aislada, del tipo de molécula y del grupo de investigación que obtuvo la secuencia.

Para un buen aprovechamiento de las bases de datos biológicas, las mismas deben cumplir dos requerimientos básicos:

1. La información debe ser fácilmente accesible y,
2. Debe estar implementado un método que permita extraer sólo la información requerida para responder una pregunta biológica específica.

En el ámbito biológico las bases de datos suelen clasificarse en primarias o secundarias. Las bases de datos primarias son aquellas que almacenan secuencias de ácidos nucleicos y proteínas,

estructuras o patrón de biomoléculas o de expresión de ADN y perfiles de expresión génica. Las bases de datos secundarias almacenan información derivada de las bases de datos primarias, osea, el resultado de la aplicación de diversas técnicas analíticas sobre estas fuentes primarias de datos [28].

En Internet existen numerosas bases de datos con información genética de diversos tipos. Algunas sólo almacenan información de secuencias de ADN. Otras bases de datos guardan información acerca de las mutaciones presentes en estas secuencias de ADN, la frecuencia con la que ocurren en distintas poblaciones, etc. La información acerca de la localización de estas secuencias de ADN en cromosomas, así como también la información sobre marcadores cercanos que pueden ser de utilidad para el diagnóstico se encuentra almacenada en otro tipo de bases de datos.

Algunas bases de datos están relacionadas, de manera que es fácil moverse de una a otra siguiendo enlaces, obteniendo siempre información sobre el gen, la enfermedad o la región del cromosoma que nos interesa. En muchos casos, varias bases de datos están alojadas en un mismo sitio, por ejemplo la base de datos del Centro Nacional de Información Biotecnológica (NCBI).

Para realizar la evaluación de las técnicas de minería de datos, se utilizaron cuatro bases de datos diferentes como datos de entrada, estas bases de datos corresponden a distintos géneros, dos de ellas se basan en estudios de cáncer, una se refiere al estudio de un parásito y la otra al estudio de una especie de hongo, más adelante se describe el estudio de cada una de ellas. Las cuatro bases de datos se refieren a genes de expresión representados en una estructura de microarrays.

La tabla 3.1 muestra la información general de estas bases de datos y a continuación se describe en qué se basa el estudio de cada una de ellas.

Base de Datos	Compañía	Tipo de Dato
Leucemia (Cáncer)	Affymetrix	Microarrays
Malaria (Parásito)	De Risi Lab Malaria Transcriptome Database	Microarrays
Saccharomyces Cerevisiae (Levadura)	Saccharomyces Gemone Database	Microarrays
Diffuse Large B-Cell (Linfoma)	Eisen Lab	Microarrays

Tabla 3.1: Bases de Datos de Bioinformática

Leucemia

La leucemia aguda humana es un tipo de cáncer maligno, que implica un incremento incontrolado de leucocitos. Esta base de datos fué tomada de los resultados de un estudio basado en la clasificación molecular del cáncer a través del descubrimiento y predicción de clases. En este estudio se propone un método genérico para la clasificación del cáncer, basado en el monitoreo de genes de expresión usando microarrays de ADN. Se basa en el descubrimiento de una clase haciendo la distinción entre la leucemia aguda myeloid (AML) y la leucemia aguda lymphoblastic (ALL) sin un conocimiento previo de estas clases, los resultados muestran la viabilidad de clasificar el cáncer basándose solamente en genes de expresión y sugiere una estrategia general para descubrir y predecir clases de cáncer de otros tipos independientemente del conocimiento biólogico previo de las clases [6].

Malaria (Plasmodium falciparum)

La malaria es una enfermedad ocasionada por un parásito del género Plasmodium, siendo el Falciparum el más agresivo. Esta base de datos fué tomada de los resultados de un estudio que muestra el análisis del ciclo del desarrollo asexual del Intraerythrocytic Transcriptome, estos datos demuestran que este parásito ha desarrollado un modo extremadamente especializado de regulación transcripcional, esto produce una cascada continua de genes de expresión, empezando con genes correspondientes al proceso celular general tales como síntesis

de proteínas y finalizando con funcionalidades específicas del Plasmodium tales como genes implicados en la invasión del Erythrocyte. Los datos revelan que los genes contiguos a lo largo del cromosoma rara vez son corregulados y proporcionan la primera vista comprensiva de la sincronización de la transcripción a través del desarrollo del Plasmodium y proporcionan un recurso para la identificación de nuevos candidatos quimioterapéuticos y nuevas vacunas [23].

Levadura (Saccharomyces cerevisiae)

La levadura es un eucariote unicelular que se encuentra en plantas, animales y hongos, es conocida comúnmente como levadura de cerveza o de pan. Esta base de datos fué tomada de un estudio que muestra información acerca de la biología molecular y genética de la levadura basada en genes de expresión y secuencias de genomas. En los datos de este estudio se usaron muestras de microarrays de ADN de levaduras sincronizadas y se aplicaron algoritmos de periocidad y correlación para identificar genes que cumplían con mínimos criterios para la regulación del ciclo de las células, se encontró que varios elementos conocidos contenían información predictiva de la regulación del ciclo de las células [4].

Linfoma (Difusse large b-cell)

El linfoma es un tipo de cáncer maligno del sistema linfático causado por linfocitos maduros. Esta base de datos fué tomada de los resultados de un estudio obtenido de microarrays de ADN que han conducido a la caracterización sistemática de genes de expresión en anomalías de la célula B (B-cell). Estos datos muestran que existe una diversidad de genes de expresión entre los tumores generados por los diferentes tipos del linfoma B-cell que reflejan la variación de la proliferación de tumores. La clasificación molecular de los tumores basados en genes de expresión puede identificar previamente subtipos de cáncer que no son detectados clínicamente, el propósito es que la información a nivel molecular pueda ser usada para refinar el diagnóstico del linfoma [24].

3.2. Técnicas

Se refiere al tipo de técnica de minería de datos aplicada para analizar datos relacionados con la bioinformática.

Cuando se trabaja con conjuntos de datos que contienen cientos de miles de elementos puede ser muy difícil encontrar información útil sin hacer cierta clase de agrupamiento; el objetivo principal del agrupamiento es identificar dentro de una base de datos, estructuras o subclases de los datos que tengan algún sentido, esto se hace particionando un conjunto de datos en un conjunto de subclases significativas llamadas grupos (clusters) de tal manera que cada miembro de un grupo esté lo más cercano posible a otro, y grupos diferentes estén lo más lejos posible uno del otro, esto garantiza la obtención de aquellos grupos de datos con información significativa. Es por ello que en este trabajo se ha optado por utilizar técnicas de agrupamiento para evaluar los datos.

¿Por qué usar técnicas de agrupamiento?

Entre las razones que pueden decirse para justificar el uso de estas técnicas destacan las siguientes:

- Cuando no existe un conocimiento suficiente acerca de las clases en que se pueden distribuir los objetos de interés.
- Cuando existe un conocimiento completo de las clases y/o se desea comprobar la validez del conjunto de entrenamiento.

Ahora bien, el resultado de una técnica de agrupamiento depende de diversos factores:

- El algoritmo concreto empleado para encontrar los agrupamientos, la variedad de algoritmos hace que la elección de una estrategia de agrupamiento, e incluso entre diferentes implementaciones de un mismo algoritmo proporcione resultados diferentes.
- El valor de los parámetros del algoritmo.

- Los patrones utilizados y en algunas ocasiones, hasta el orden en que se procesan.
- La medida de similaridad adoptada.

Aunque existen diversas técnicas para agrupar datos procedentes de experimentos de tipo biológico, la idea básica es agrupar elementos con alto grado de similitud, un problema con la mayoría de las técnicas de agrupamiento es que algunos datos de entrada son colocados en agrupamientos en donde en realidad no comparten ninguna semejanza, así que puede ser importante analizar si el dato exhibe alguna tendencia de agrupamiento.

Dos de las técnicas de agrupamiento que se han seleccionado para realizar el análisis de las herramientas de bioinformática y que son presentadas a continuación pertenecen al grupo de técnicas conocidas como de análisis no supervisado y la medida de similaridad elegida para evaluarlos es la distancia Euclidiana, descrita en el capítulo 2 sección 2.4.

Las técnicas de análisis no supervisado indican que el análisis está hecho sin ningún conocimiento a priori de los datos de entrada. Las técnicas de análisis no supervisado se dividen en jerárquicas y no jerárquicas, se ha elegido una de cada una simplemente para establecer una diferencia entre ellas. Para el caso de las técnicas jerárquicas se ha seleccionado la técnica del agrupamiento jerárquico aglomerativo y para las no jerárquicas el k-means.

3.2.1. Agrupamiento jerárquico aglomerativo

Es un método determinístico basado en una matriz de distancias. Establece pequeños grupos de genes/condiciones que tienen un patrón de expresión común y posteriormente construye un árbol de forma secuencial. Este árbol establece una relación ordenada de los grupos previamente definidos y la longitud de sus ramas es representación de la distancia entre los distintos nodos del mismo [11].

La estrategia general de esta técnica es separar cada dato de entrada, en este caso, cada gen, en un nodo o cluster diferente, calcula la distancia entre los dos genes más próximos determinada por

una medidad de similaridad y los junta en un sólo cluster, entonces se vuelve a calcular la matriz de distancia sustituyendo los dos patrones que se han unido por el promedio de ambos hasta que todos los elementos se encuentren en el mismo cluster. En cada paso, el algoritmo es capaz de juntar los genes no sólo de dos en dos, sino de muchos más a la vez.

El resultado que produce el agrupamiento jerárquico es un árbol llamado dendograma, que es una representación gráfica de un grupo de relaciones basadas en la cercanía o similitud entre los datos, las hojas del árbol representan a los elementos de entrada y el nodo raíz es el resultado final del agrupamiento que contiene a todas las hojas, una rama es un punto del árbol donde dos clusters se han combinado. La figura 3.1 muestra un ejemplo de un dendograma simple.

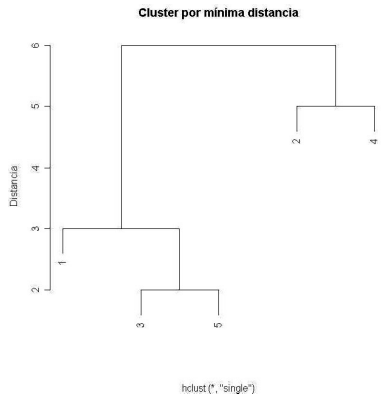


Figura 3.1: Ejemplo de dendograma obtenido por mínima distancia

Esta técnica puede diferenciarse de otras en la forma en que calcula la distancia del nuevo cluster formando al resto de los elementos de la matriz. La técnica de agrupamiento jerárquico admite algunas variantes para el cálculo de la similitud o distancia que debe existir entre cada cluster, estas variantes son conocidas como métodos de enlace (linkage methods). Los métodos de enlace más conocidos son el enlace simple, el enlace promedio y el enlace completo. Las figuras 3.2, 3.3 y 3.4 muestran las definiciones de estos métodos.

Métodos de Enlace

* Enlace Simple: La distancia entre cada miembro de un cluster con respecto a cada miembro de otro cluster es la mímima. $d(C_{(ij)}, C_k) = min\{d(C_i, C_k), d(C_j, C_k)\}$

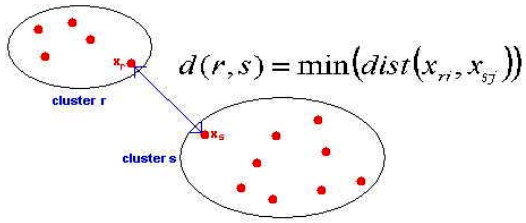


Figura 3.2: Enlace Simple

* Enlace Promedio: Toma la media entre todos los miembros de un cluster con los miembros de otro cluster. $d(C_{(ij)}, C_k) = \frac{d(C_i,C_k)+d(C_j,C_k)}{2}$

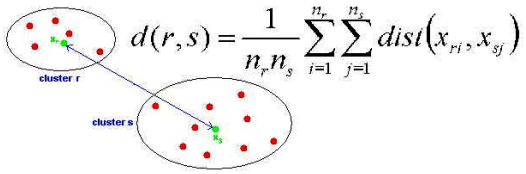


Figura 3.3: Enlace Promedio

* Enlace Completo: La distancia entre cada miembro de un cluster con respecto a cada miembro de otro cluster es la máxima. $d(C_{(ij)}, C_k) = max\{d(C_i, C_k), d(C_j, C_k)\}$

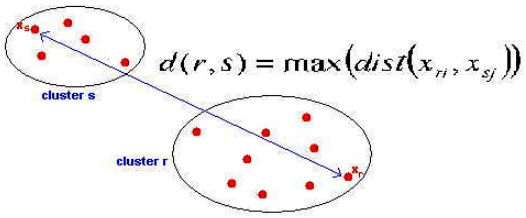


Figura 3.4: Enlace Completo

Donde:

$d(r, s)$ = distancia entre los clusters r y s.

x_{ri} = objeto i en el cluster r.

x_{sj} = objeto j en el cluster s.

3.2.2. K-means

El K-means es una técnica de agrupamiento muy popular, también es conocida como el algoritmo de las medias móviles porque en cada iteración se recalculan los centros de los agrupamientos. Por esta razón se incorpora el índice t a la notación que se emplea, de manera que con $S_i(t)$ indicamos el conjunto de patrones asociados al agrupamiento S_i en la iteración t y mediante $Z_i(t)$ indicamos el valor de su centro en esa iteración [11].

Este algoritmo requiere un único parámetro, K, que es el número de agrupamientos que debe encontrar. Se puede plantear en tres pasos:

Inicialización: Consiste en inicializar arbitrariamente los centros de los K grupos.

Asignación y actualización de los centros: En este paso se asigna cada patrón al grupo más cercano y se recalculan los centros en base a esta asignación. Los centros de cada grupo se denominan centroides.

Finalización: En el paso anterior algunos patrones pueden cambiar de agrupamiento y en consecuencia, los centros de éstos. Si esto ocurre, se trata de repetir el paso 2 hasta que no se modifiquen los centros. Cuando no haya modificaciones se considera que se ha encontrado una buena partición y se termina el agrupamiento.

Normalmente se utiliza una medida de similaridad basada en el error cuadrático:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{3.1}$$

Donde p representa al objeto y m_i a la media del cluster C_i (ambos son objetos multidimensionales).

El mapa de clusters que ofrece este algoritmo carece de topología. K-means es susceptible a valores extremos porque distorsionan la distribución de los datos. También se pueden utilizar las modas (K-modes) para agrupar objetos categóricos. Otra posibilidad es usar medianas (K-medoids) para agrupar en base al objeto más representativo del cluster.

Otra variante es hacer un K-means jerárquico, en donde se empieza con K=2 y se continúan formando clusters sucesivos en cada rama. Si queremos escalarlo a grandes bases de datos, podemos tomar únicamente muestras de los datos.

3.2.3. PCA

Otra técnica utilizada en este trabajo para la evaluación de datos biológicos se conoce como un método de proyección llamado Análisis de Componentes Principales (PCA).

El análisis de componentes principales es una técnica de análisis multivariado que se emplea en la reducción de la dimensionalidad de un conjunto de datos. Tranforma un conjunto de p variables correlacionadas (miden información común) en un conjunto de variables no correlacionadas (que no tengan repetición o redundancia en la información) que contienen la información principal del sistema [2].

Estas nuevas variables llamadas *componentes principales* serán una combinación lineal de las variables originales y además serán independientes entre sí, son obtenidas en orden decreciente de importancia, de modo que las primeras componentes principales resuman la mayor cantidad posible de la variabilidad de los datos originales.

Se parte de una matriz X, donde cada fila es un dato y cada columna una variable, la dimensio-

nalidad de esta matriz será de m x k, donde m es el número de datos y k el número de variables. Busca las direcciones ortogonales que explican la máxima variabilidad de los datos y los utiliza como nuevos ejes de coordenadas llamados componentes principales para representarlas. El primer componente principal es la dirección que explica la máxima variabilidad de los datos; el segundo se escoge de tal forma que sea perpendicular al primero y que explique la máxima variabilidad una vez extraída la explicada por el primer componente principal y así sucesivamente.

Para definir estos ejes se utilizan sus *loadings* que son los cosenos de los ángulos que forman con los ejes antiguos y los *scores* que son las coordenadas de los datos en estos nuevos ejes. Matemáticamente la matriz de datos X se descompone en el producto de dos matrices, T (matriz de *scores*) y P (matriz de *loadings*) más una matriz E de residuales de X.

$$X = TP^T + E \tag{3.2}$$

Los diferentes componentes principales no contienen la misma información; los primeros describen la fuente de variación más importante de los datos, que se puede asociar a la información más relevante. El análisis de componentes principales se puede emplear en el análisis de expresiones genéticas en combinación con otras técnicas como el k-means por ejemplo.

3.3. Complejidad

La complejidad de un algoritmo, no tiene que ver con dificultad, sino con rendimiento, permite medir de alguna forma el coste (en tiempo y recursos) que consume un algoritmo en su ejecución, es decir, el tiempo que tarda para ejecutarse. La gran mayoría de estudios de complejidad están orientados hacia el desempeño de los algoritmos en función del tiempo, por lo que de aquí en adelante, al hablar de complejidad, se asume que se refiere al tiempo [8].

Las técnicas expuestas en la sección 3.1 podrían haber sido estudiadas para establecer a que orden de complejidad pertenecen y decir con esto cual de ellas debería ser la que consumiría menos

tiempo y recursos, pero esto no garantizaría que el resultado que se obtuviera después de evaluar las técnicas coincidiera con ese criterio. Para que sea relevante el definir la complejidad de un algoritmo, ésta es expresada en términos del tamaño de los datos de entrada. Al referir la complejidad al tamaño de la entrada, esta medida es más representativa y útil.

Por ello al evaluar las distintas técnicas sobre cada base de datos en las diferentes herramientas se toma en cuenta el tiempo que tarda cada técnica en obtener el resultado, este resultado se será influenciado por el tamaño de los datos de entrada. Los datos de entrada usados para la evaluación son tomados de las bases de datos descritas en la sección 3.1.2.

Estas bases de datos son relativamente pequeñas en comparación con otras bases de datos de tipo biólogico que involucran variedad de experimentos, sin embargo, para ser datos tomados para realizar pruebas, podrían considerarse como grandes, por ejemplo, para la base de datos leucemia se tienen 4630 filas por 41 columnas, lo que da un total de 189830 genes, para la base de datos malaria se tienen 5494 filas por 11 columnas osea 60434 genes, para la base de datos levadura se tienen 6222 filas por 82 columnas, es decir 510204 genes y para la base de datos linfoma se tienen 13412 filas por 40 columnas y en total 536480 genes.

Tener bases de datos que provienen de estudios y dimensiones diferentes ofrece la posibilidad de comparar las distintas técnicas que resuelven un mismo problema. Es importante considerar que el tamaño de los datos de entrada son un factor importante en la obtención rápida de los resultados, no necesariamente depende de la técnica que se utilize para evaluarlos.

Capítulo 4

Herramientas de Bioinformática

A través de Internet se puede acceder a diversas herramientas de bioinformática. Muchas de éstas intentan resolver el mismo problema biológico pero con aproximaciones diferentes. La elección de la mejor herramienta depende del contexto del problema a resolver y los recursos a los que uno tiene acceso.

Existen varios parámetros que hay que tener en cuenta a la hora de seleccionar una herramienta independientemente del problema biológico a resolver: el procesador y el sistema operativo donde va a instalarse y/o ejecutarse la herramienta, la facilidad de adaptación del usuario a la interfase, la capacidad de interactuar con otros programas y el precio, en el caso de las herramientas comerciales, que muchas veces ofrecen más alternativas que las herramientas disponibles de manera pública. Es común encontrar las mejores herramientas para usuarios con poco conocimiento de computación en sistemas operativos como Windows, MacOS y OSx.

Las características que se deben considerar para seleccionar una herramienta deben ser las siguientes:

- Nombre de la Herramienta.
- Entidad responsable de la herramienta (NCBI, EBI, TIGR, etc.)

- Modalidades de uso (Página Web, Programa local, Linea de comandos, etc.)
- Sistema operativo en que funciona (Windows, MacOS, UNIX/Linux, etc.)
- Dirección Web del programa.
- Problema biológico que resuelve (Homología entre secuencias, Búsqueda de dominios conservados, etc.)
- Problema computacional (Alineamiento, Búsqueda de patrones, etc.)
- Algoritmos usados (Agrupamiento, K-means, etc.).

4.1. Descripción de Herramientas de Bioinformática

La tabla 4.1 contiene la información más importante de las herramientas seleccionadas para realizar la evaluación. Se describen las características generales de la herramienta, como son, el nombre, la compañía o entidad por la que fué creada, la plataforma en la que se encuentra disponible, el tipo de método que utiliza y el tipo de dato con el que trabaja. En seguida se hace una breve descripción de cada una de las herramientas con el propósito de tener un conocimiento básico de su aplicación en la Bioinformática.

Software	Compañía	Plataforma	Método	Tipo de Dato
J Express Pro v2.7	Molmine	Win, Linux	Agrupamiento jerárquico, K-means, PCA	Genes
Cluster & Tree View v3.0	Michael Eisens Lab, Lawrence Berkeley National Lab (LBNL)	Win, Linux, Unix, MacOS	Agrupamiento jerárquico, K-means, PCA	Genes
MEV v3.1	The Institute of Gemonic Research (TIGR)	Win, Linux, Unix, MacOSx	Agrupamiento jerárquico, K-means, PCA	Genes
GEPAS v3.0	National Spanish Cancer Center (CNIO)	Web(en línea)	Agrupamiento jerárquico, K-means	Genes
Expresion Profiler (EPCLUST) v0.9.23 beta	European Bioinformatics Institute (EBI)	Web(en línea)	Agrupamiento jerárquico, K-means	Genes

Tabla 4.1: Herramientas de Bioinformática

4.1.1. J Express Pro

J Express Pro es una herramienta portable y comprensible para el análisis y visualización de datos de microarrays. El programa brinda un acceso de escalamiento multidimensional, agrupamiento y visualización de una manera integrada y flexible. Permite al usuario cargar conjuntos de datos que son resultado de un conjunto de experimentos de microarrays, aplicar técnicas de análisis, visualizar los resultados y producir figuras de calidad.

J Express Pro es compatible con datos de GenPix, Affymetrix, Agilent, Scanalyze y Array Express, tiene una interfaz flexible para construir archivos de programas de análisis de otros sistemas. Importa fácilmente archivos tradicionales en formato tabular. Las técnicas de análisis incluyen: agrupamiento (jerárquico, k-means y self-organizing maps (SOM)), métodos de proyección (PCA) y análisis de correspondencia. J Express posee un sistema cliente/servidor por lo que permite que múltiples usuarios trabajen en un sólo proyecto simultáneamente. La figura 4.1 muestra la ventana principal de J Express Pro.

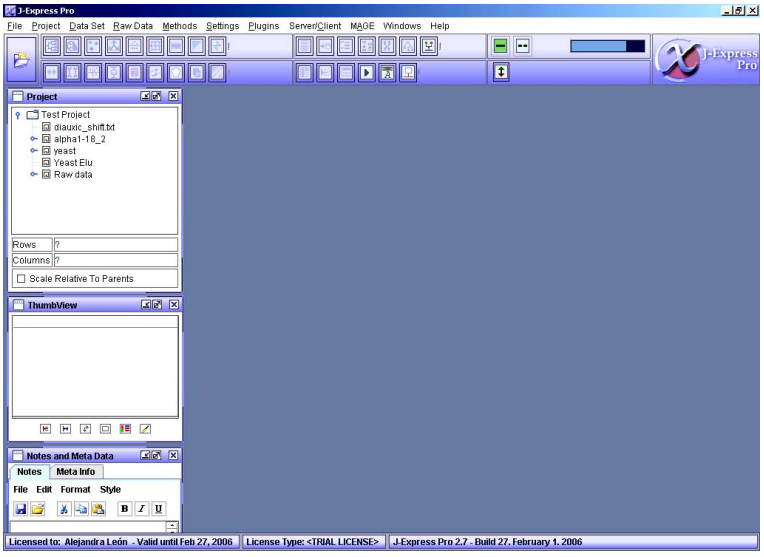


Figura 4.1: Ventana principal de J Express Pro

4.1.2. Cluster & TreeView

Cluster & TreeView son programas que proveen un ambiente computacional y gráfico para el análisis de datos de experimentos de microarrays u otros conjuntos de datos genómicos.

El software Cluster puede organizar y analizar los datos de diferente manera, implementa las técnicas de agrupamiento más utilizadas para el análisis de datos de expresión genética. Provee una interfaz gráfica para accesar a las rutinas de las técnicas de agrupamiento. El software Tree View permite que los datos ya organizados sean visualizados.

Las técnicas de análisis incluyen: agrupamiento (jerárquico, k-means y self-organization maps)y métodos de proyección (PCA). La figura 4.2 muestra la ventana principal de Cluster & TreeView.

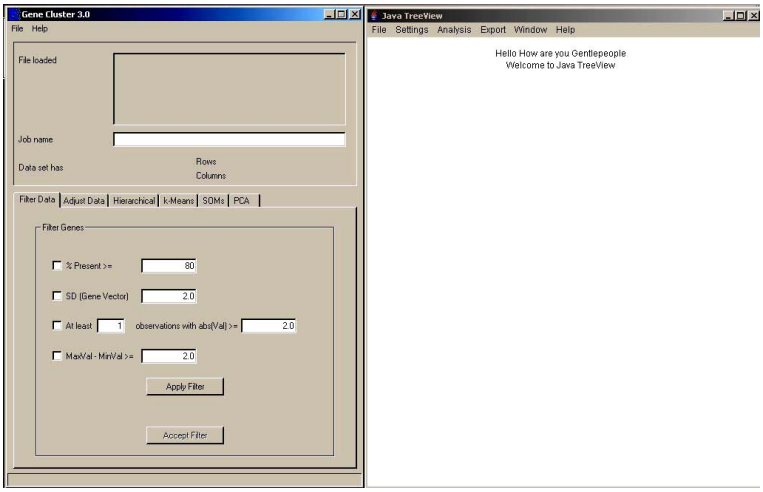


Figura 4.2: Ventana principal de Cluster & Tree View

4.1.3. Mev (Multiexperiment Viewer)

MEV es un software desarrollado por el Instituto de Investigación Genómica (TIGR) conocido también como TM4. Es una aplicación que permite la identificación de genes y patrones de expresión de interés, contiene una variedad de algoritmos de normalización y técnicas de agrupamiento que permite al usuario la flexibilidad de crear vistas significativas de los datos.

MEV es una herramienta versátil para el análisis de datos de microarray, permite analizar archivos de expresión que han sido normalizados y filtrados, incorpora algoritmos sofisticados de agrupamiento, visualización, clasificación y análisis estadístico. MEV maneja varios formatos de archivos de entrada, éstos incluyen el .mev y el .tav generados por TIGR Spotfinder y TIGR MIDAS, archivos de Affymetrix (.txt) y de Genepix (.gpr).

Las técnicas de análisis incluyen: agrupamiento (jerárquico, k-means, bootstrapping, self-organizing trees, self-organization maps), métodos de proyección (PCA), métodos estadísticos (ANOVA, T-tests), entre otros. La figura 4.3 muestra la ventana principal de MEV.

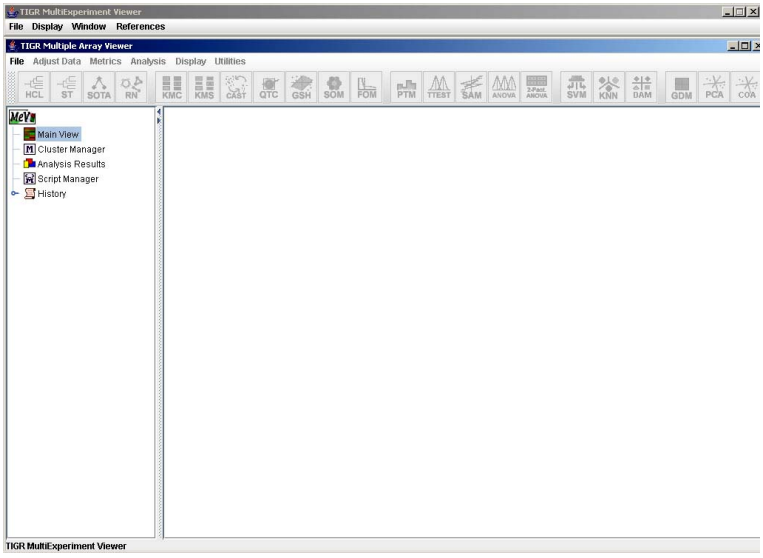


Figura 4.3: Ventana principal de MEV

4.1.4. GEPAS (Gene Expression Pattern Analysis Suite)

GEPAS es una herramienta en línea (web-based) para el análisis de patrones de expresión genética. Incluye técnicas de normalización, agrupamiento, predicción, clasificación, anotación funcional y expresión diferencial de genes.

Los métodos de análisis incluyen normalización, preprocesamiento, técnicas de agrupamiento (jerárquico, kmeans, SOM, SOM-Tree, SOTA), métodos de clasificación (SVM Support Vector Machines)y algunas técnicas de análisis de minería de datos (FatiGO, FatiScan, Tissues Mining Tool, etc). La figura 4.4 muestra la ventana principal de GEPAS.

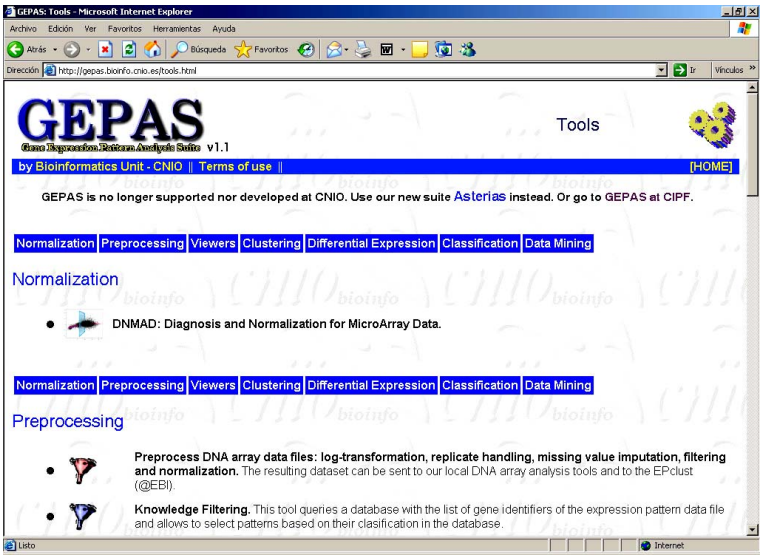


Figura 4.4: Ventana principal de GEPAS

4.1.5. Expression Profiler(EPCLUST)

Expression Profiler es un conjunto de herramientas de agrupamiento, análisis y visualización de genes de expresión y otros datos genómicos. Estas herramientas permiten realizar análisis de clusters, descubrimiento de patrones, visualización del patrón, extraer secuencias reguladoras, estudiar interacciones de las proteínas así como exportar los resultados del análisis a bases de datos y herramientas externas. Expression Profiler contiene varios módulos de análisis, entre ellos se encuentra EPCLUST.

EPCLUST (Expression profile data clustering and analysis) es una herramienta de análisis, visualización y agrupamiento de datos, funciona para datos númericos (datos de expresión genética) así como para datos de secuencias.

Las técnicas de análisis de EPCLUST incluyen el agrupamiento jerárquico y el k-means. La figura 4.5 muestra la ventana principal de EPCLUST.

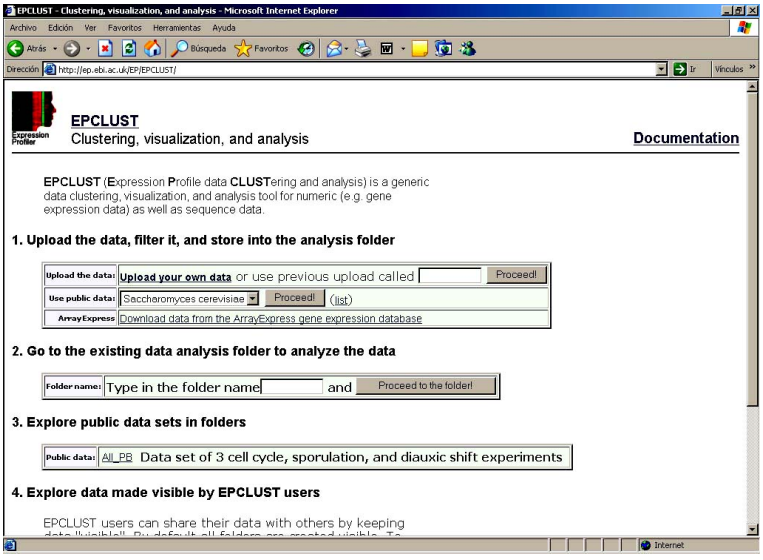


Figura 4.5: Ventana principal de EPCLUST

4.2. Parámetros de evaluación

Se ha explicado brevemente la aplicación que tiene cada herramienta dentro del ámbito de la bioinformática, se han mencionado las técnicas de minería de datos utilizadas para evaluar los datos en las distintas herramientas, ahora se definen los parámetros utilizados para realizar la evaluación. Definimos los parámetros de acuerdo al orden establecido en las categorías de análisis para la evaluación de las herramientas explicado en el capítulo 3.

4.2.1. Parámetros de los datos

En el capítulo 3 en la sección 3.1 se han mencionado tanto el tipo de dato, como las bases de datos usadas en la evaluación, en el caso de los datos no podemos hablar de parámetros como tal, sino más bien, del formato de los datos que no necesariamente es el mismo. La tabla 4.2 muestra el formato de los datos de entrada que utiliza cada herramienta.

Sotware	Tipo texto (.txt)	Tipo raw (.gpr)
J Express Pro	√	√
Cluster & Tree View	√	
MEV	√	√
GEPAS	√	
EPCLUST	√	

Tabla 4.2: Formato de los datos de entrada para cada herramienta

Nótese que sólo se trabaja con dos tipos de datos: los datos de tipo tabular o tipo texto (tab delimited .txt), que son aquellos en forma de tabla donde las filas suelen corresponder a genes y las columnas a las medidas individuales de cada gen, en este tipo de datos las columnas suelen estar separadas por espacios en blanco, por tabulaciones o comas; y los datos conocidos como datos en crudo (raw data .gpr), que son datos que han sido colectados pero no procesados, comúnmente se representan por imagenes obtenidas de análisis de experimentos, mayormente de microarrays y que tienen asociado a la imagen, un archivo que contiene información del gen representado por

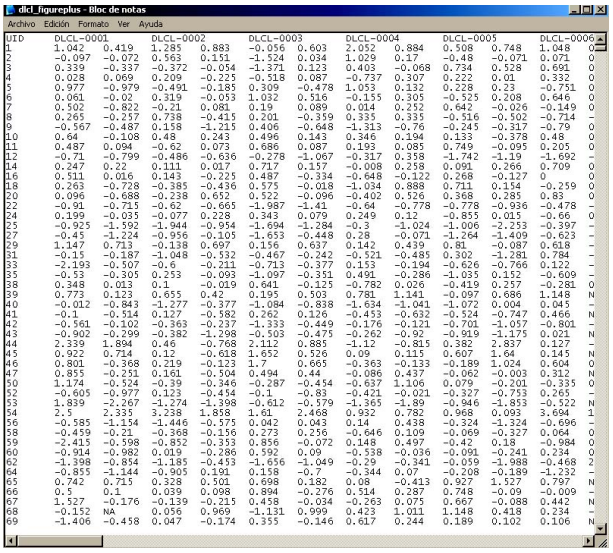
esa imagen, estos datos se convierten en información una vez que se extraen, organizan y analizan para su presentación.

La tabla 4.3 muestra el formato de los datos en los que se encuentra cada base de datos.

Base de Datos	Formato de los Datos
Leukemia (cáncer)	Tipo texto (.txt)
Malaria (parásito)	Raw data (.gpr)
Saccharomyces Cerevisiae (levadura)	Tipo texto (.txt)
Diffuse Large B-Cell (Linfoma cáncer)	Tipo texto (.txt)

Tabla 4.3: Formato de las bases de datos

La figura 4.6 representa los datos de tipo tabular tomados de la base de datos Linfoma, donde las filas representan a los individuos, que son los genes de ADN, y las columnas representan las componentes o medidas inviduales de esos genes.


dlcl_figureplus - Bloc de notas

Archivo Edición Formato Ver Ayuda

UID	DLCL-0001		DLCL-0002		DLCL-0003		DLCL-0004		DLCL-0005		DLCL-0006	
1	1.042	0.419	1.285	0.883	-0.056	0.603	2.052	0.884	0.508	0.748	1.048	0
2	-0.097	-0.072	0.563	0.151	-1.524	0.034	1.029	0.17	-0.48	-0.071	0.071	0
3	0.339	-0.337	-0.372	-0.054	-1.371	0.123	0.403	-0.068	0.734	0.528	0.691	0
4	0.028	0.069	0.209	-0.225	-0.518	0.087	-0.737	0.307	0.222	0.01	0.332	0
5	0.977	-0.979	-0.491	-0.185	0.309	-0.478	1.053	0.132	0.228	0.23	-0.751	0
6	0.061	-0.02	0.319	-0.053	1.032	0.516	-0.155	0.305	-0.525	0.208	0.646	0
7	0.502	-0.822	-0.21	0.081	0.19	0.089	0.014	0.232	0.642	-0.026	-0.149	0
8	0.265	-0.257	0.738	-0.415	0.201	-0.359	0.335	0.335	-0.516	-0.502	-0.714	-
9	-0.567	-0.487	0.158	-1.215	0.406	-0.648	-1.313	-0.76	-0.245	-0.317	-0.79	0
10	0.64	-0.108	0.48	0.243	0.496	0.143	0.346	0.194	0.133	-0.378	0.48	0
11	0.487	0.094	-0.62	0.073	0.686	0.087	0.193	0.085	0.749	-0.095	0.205	0
12	-0.71	-0.799	-0.486	-0.636	-0.278	-1.067	-0.317	0.358	-1.742	-1.19	-1.692	-
14	0.247	0.22	0.111	0.017	0.717	0.157	-0.008	0.258	0.091	0.266	0.709	0
16	0.511	0.016	0.143	-0.225	0.487	-0.334	-0.648	-0.122	0.268	-0.127	0	0
18	0.263	-0.728	-0.385	-0.436	0.575	-0.018	-1.034	0.888	0.711	0.154	-0.259	0
20	0.096	-0.688	-0.238	0.652	0.522	-0.096	-0.402	0.526	0.368	0.285	0.83	0
22	-0.91	-0.715	-0.62	-0.665	-1.987	-1.41	-0.64	-0.778	-0.778	-0.936	-0.478	-
24	0.199	-0.035	-0.077	0.228	0.343	0.079	0.249	0.12	-0.855	0.015	-0.66	0
25	-0.925	-1.592	-1.944	-0.954	-1.694	-1.284	-0.3	-1.024	-1.006	-2.253	-0.397	-
27	-0.45	-1.224	-0.956	-0.105	-1.653	-0.448	0.28	-0.071	-1.264	-1.409	-0.623	-
29	1.147	0.713	-0.138	0.697	0.156	0.637	0.142	0.439	0.81	-0.087	0.618	-
31	-0.15	-0.187	-1.048	-0.532	-0.467	-0.242	-0.521	-0.485	0.302	-1.281	0.784	-
33	-2.198	-0.507	-0.6	-0.211	-0.713	-0.377	0.153	-0.194	-0.626	-0.766	0.122	-
35	-0.53	-0.305	0.253	-0.093	-1.097	-0.351	0.491	-0.286	-1.035	0.152	-0.609	-
38	0.348	0.013	0.1	-0.019	0.641	-0.125	-0.782	0.026	-0.419	0.257	-0.281	0
39	0.773	0.123	0.655	0.42	0.195	0.503	0.781	1.141	-0.097	0.686	1.148	N
40	-0.012	-0.843	-1.277	-0.377	-1.084	-0.838	-1.634	-1.041	-1.072	0.004	0.045	-
41	-0.1	-0.514	0.127	-0.582	0.262	0.126	-0.453	-0.632	-0.524	-0.747	0.466	N
42	-0.561	-0.102	-0.363	-0.237	-1.333	-0.449	-0.176	-0.121	-0.701	-1.057	-0.801	-
43	-0.902	-0.299	-0.382	-1.298	-0.503	-0.475	-0.262	-0.92	-0.919	-1.175	0.021	N
44	2.339	1.894	0.46	-0.768	2.112	0.885	-1.12	-0.815	0.382	2.837	0.127	-
45	0.922	0.714	0.12	-0.618	1.652	0.526	0.09	0.115	0.607	1.64	0.145	N
46	0.801	-0.368	0.219	-0.123	1.7	0.665	-0.363	-0.133	-0.189	1.024	0.604	0
47	0.855	-0.251	0.161	-0.504	0.494	0.44	-0.086	0.437	-0.062	-0.003	0.312	N
50	1.174	-0.524	-0.39	-0.346	-0.287	-0.454	-0.637	1.106	0.079	-0.201	-0.335	0
52	-0.605	-0.977	0.123	-0.454	-0.1	-0.83	-0.421	-0.021	-0.327	-0.753	0.265	-
53	1.839	-2.267	-1.274	-1.398	-0.612	-0.579	-1.365	-1.89	-0.946	-1.853	-0.522	N
54	2.5	2.335	3.238	1.858	1.61	2.468	0.932	0.782	0.968	0.093	3.694	1
56	-0.585	-1.154	-1.446	-0.575	0.042	0.043	0.14	0.438	-0.324	-1.324	-0.696	-
58	-0.459	-0.21	-0.368	-0.156	0.273	0.256	-0.646	0.109	-0.069	-0.327	0.064	0
59	-2.415	-0.598	-0.852	-0.353	0.856	-0.072	0.148	0.497	-0.42	0.18	-0.984	0
60	-0.914	-0.982	0.019	-0.286	0.592	0.09	-0.538	-0.036	-0.091	-0.241	0.234	0
62	-1.398	-0.854	-1.185	-0.453	-1.656	-1.049	-0.29	-0.341	-0.059	-1.988	-0.468	2
64	-0.855	-1.144	-0.905	0.191	0.158	-0.7	-0.344	0.07	-0.208	-0.189	-1.232	-
65	0.742	0.715	0.328	0.501	0.698	0.182	0.08	-0.413	0.927	1.527	0.797	N
66	0.5	0.1	0.039	0.098	0.894	-0.276	0.514	0.287	0.748	-0.09	-0.009	-
67	1.527	-0.176	-0.139	-0.215	0.458	-0.034	-0.263	0.075	0.667	-0.088	0.442	N
68	-0.152	NA	0.056	0.969	-1.131	0.999	0.423	1.011	1.148	0.418	0.234	-
69	-1.406	-0.458	0.047	-0.174	0.355	-0.146	0.617	0.244	0.189	0.102	0.106	N

Figura 4.6: Ejemplo de datos de tipo tabular

La figura 4.7 representa los datos de tipo raw que son los datos en crudo, tomados de la base de datos Malaria, se observa una especie de fotografía de los datos, cada celda representa un gen que contiene información del experimento realizado sobre éste, esta representación de los datos es típica de experimentos basados en microarrays.

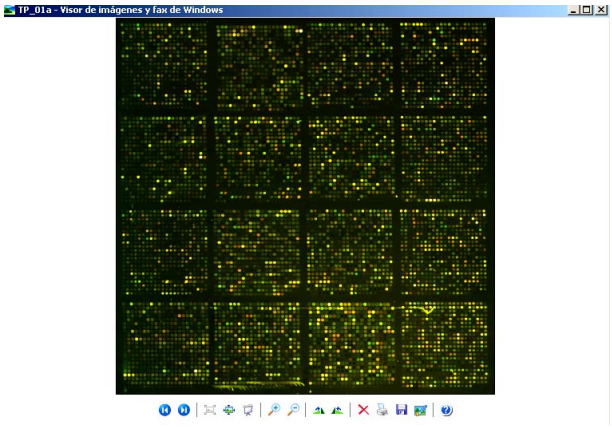


Figura 4.7: Ejemplo de datos de tipo raw data

4.2.2. Parámetros de las técnicas

Como ya se mencionó, las técnicas elegidas para evaluar los datos en las distintas herramientas son el Agrupamiento jerárquico aglomerativo, el K-means y el PCA. Para que estos métodos funcionen correctamente es necesario definir ciertos parámetros de entrada que determinan la manera en la que se inicia y concluye la evaluación.

Cada herramienta es y funciona de manera diferente a pesar de que ejecutan las mismas técnicas, y aunque todas coinciden en los parámetros de entrada, la manera de introducirlos suele variar; algunas de estas herramientas ofrecen algún parámetro que en otras no se encuentra, esto no es significativo para realizar el análisis esto se comenta para hacer incapié en la funcionalidad y variabilidad de cada herramienta, a continuación se definen los parámetros usados para el análisis.

Agrupamiento jerárquico aglomerativo

* Distancia: Euclideana.
* Método de enlace: Enlace simple.

En todas las herramientas se usó como método de enlace el enlace simple, excepto en la herramienta MEV en la que se utilizó el enlace promedio porque por razones desconocidas no se generaba ningún resultado visible con el enlace simple.

K-means

* Distancia: Euclideana
* Número de clusters: 3
* Iteraciones máximas: 200

No existen procedimientos óptimos para determinar a priori el número k de clusters, en este caso se eligió k=3, basándose en el hecho de que tres son las fases principales que caracterizan el estudio de la mayoría de las bases de datos seleccionadas, por ejemplo, en la base de datos Leucemia se tienen tres fases de estudio: ALL b-cell, ALL t-cell y AML [Cap. 3, Sec. 3.1.2].

PCA

Para aplicar un PCA no se necesita especificar ningún parámetro, esta técnica trabaja de manera interna, es decir, hace un cálculo interno de la matriz de la covarianza y presentan los resultados a través de una gráfica bidimensional.

La única variante de esta técnica ocurre en la herramienta MEV, en donde además se debe especificar el número de vecinos para el método de imputación KNN (K-Nearest Neighbor), los métodos de imputación se pueden definir simplemente como promedios o selecciones provenientes de una distribución de predicción de los valores faltantes que se basa en los valores observados [7]. Se han creado métodos que hacen uso de métricas para medir distancias entre

unidades basadas en los valores de las variables asociadas dentro del mismo conjunto de datos.

Luego se calculan tales distancias y se procede a imputar los valores faltantes utilizando las unidades del conjunto de unidades completas más cercanas. El algoritmo KNN imputa valores de esta forma utilizando como métrica la distancia euclideana entre las unidades. El valor elegido para la imputación KNN fué de 10, que es el valor que aparece por default.

4.2.3. Parámetros para la complejidad

Recordemos que medir la complejidad significa medir el costo en tiempo y recursos que consume cada técnica en obtener el resultado, por ello el parámetro de evaluación empleado aquí es el tiempo. Éste fué tomado en minutos y medido para cada técnica en su evaluación para cada herramienta.

La tabla 4.4 muestra el resultado de medir el tiempo que consume el agrupamiento jerárquico, la tabla 4.5 muestra el resultado para el k-means y la tabla 4.6 muestra el resultado para el PCA; la obtención de estas tablas ofrece la posibilidad de comparar el tiempo que consume cada técnica en la ejecución realizada en las distintas herramientas.

Jerárquico

Dato	J Express	Cluster	MEV	GEPAS	EPCLUST
Leucemia	6	0.03	1	2	0.05
Malaria	4	NA	2	NA	NA
Levadura	16	0.05	5	5	0.1
Linfoma	21	1	14	39	0.62

Tabla 4.4: Tiempo de ejecución en min. para el agrupamiento jerárquico

Si observamos la tabla 4.4 podemos concluir que la técnica de agrupamiento jerárquico aplicada en todas las bases de datos consume mayor tiempo cuando es evaluada en J Express, seguida de la

herramienta MEV aunque para la herramienta GEPAS la evaluación del agrupamiento jerárquico para la base de datos linfoma consume mayor tiempo, esta técnica evaluada en las herramientas Cluster y EPCLUST consume muy poco tiempo, todos los resultados fueron obtenidos en menos de 1 minuto, excepto para la base de datos linfoma en la herramienta Cluster en donde se obtuvo exactament el minuto; la palabra NA significa que esa base de datos no aplica en la herramienta señalada por no cumplir con el tipo de formato exigido por la misma.

K-means

Dato	J Express	Cluster	MEV	GEPAS	EPCLUST
Leucemia	0.02	1	0.02	9	0.02
Malaria	0.01	NA	0.03	NA	NA
Levadura	0.03	2	0.11	22	0.05
Linfoma	0.8	3	0.14	26	0.13

Tabla 4.5: Tiempo de ejecución en min. para el k-means

La tabla 4.5 muestra que en general el k-means consume pocos recursos, la mayoría de los resultados se obtuvieron en menos de un minuto, excepto en la herramienta Cluster y en GEPAS, cuyo tiempo varía, en Cluster se sobrepasa el minuto por muy poco, pero en GEPAS sí existe una diferencia grande en comparación con las demás herramientas, podemos decir que el k-means aplicado en esta herramienta consume un tiempo significativo en la obtención de resultados y por lo tanto no es recomendable de aplicar.

PCA

Dato	J Express	Cluster	MEV	GEPAS	EPCLUST
Leucemia	0.05	0.02	0.03	NA	NA
Malaria	0.03	NA	0.55	NA	NA
Levadura	0.08	0.03	3.51	NA	NA
Linfoma	0.10	0.05	12	NA	NA

Tabla 4.6: Tiempo de ejecución en min. para el PCA

De la tabla 4.6 sólo se pueden hacer comparaciones entre tres herramientas porque tanto en GEPAS como en EPCLUST no existe la opción de ejecutar el PCA, entre estas herramientas el tiempo de ejecucuión del PCA es mínimo, en todas el resultado se obtiene en menos de un minuto, excepto en MEV en donde hay una variante para las bases de datos levadura y linfoma; la ejecución del PCA aplicado sobre la base de datos levadura pasa los 3 minutos y para la base de datos linfoma el tiempo es de 12 minutos.

Del resultado obtenido de medir el tiempo de ejecución que consume cada técnica y mediante la visualización de este resultado expresado en estas tres tablas podemos observar que cualquiera de las técnicas aplicadas en cualquier herramienta sobre la base de datos linfoma es la que mayor tiempo consume en comparación con el tiempo consumido por cada técnica evaluada en cualquier herramienta para el resto de las bases de datos. En general no se puede decir con exactitud si una técnica es más rápida que otra, esto viene fuertemente ligado a la magnitud de la base de datos con la que se esté trabajando.

La dimensión del conjunto de datos de entrada representan un factor importante a considerar porque afecta la rapidez en la obtención del resultado que se quiere, independientemete de la técnica de minería de datos que se utilize para obtenerlo.

Capítulo 5

Resultados

En este capítulo se presentan los resultados visuales obtenidos después de la evaluación de las técnicas de minería de datos sobre las bases de datos aplicadas en cada herramienta. Se incluyen sólo los resultados más significativos que permitan hacer un análisis más profundo de la aplicación de estas técnicas.

5.1. Entrada de Información

Una parte importante es la manera en la que se cargan o introducen los datos dentro de la herramienta, ya se ha mencionado en el capítulo 4 sección 4.2.1, el tipo de formato de los datos con el que trabaja cada herramienta y el tipo de formato de los datos en el que se encuentra cada base de datos.

La figura 5.1 corresponde a la herramienta J Express, muestra como se visualizan los datos de tipo tabular después de cargarlos, se observa la separación de los datos distribuidos en filas y columnas, de lado izquierdo se muestra la descripción de cada gen y de lado derecho las medidas individuales de esos genes; esta herramienta es una de las más completas en cuanto a la entrada de información ya que permite trabajar con ambos tipos de formato de los datos, el tipo tabular y el tipo raw data (.txt y .gpr).

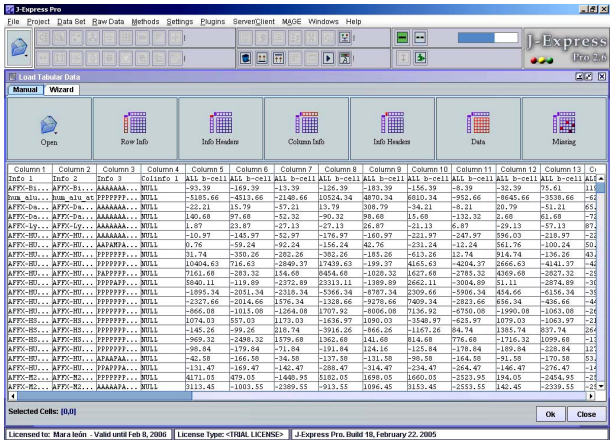


Figura 5.1: Ejemplo de entrada de datos de tipo tabular en J Express

La figura 5.2 muestra como se visualizan los datos de tipo raw data después de cargarlos, de lado izquierdo se observan los datos y de lado derecho las características de esos datos, la particularidad de este tipo de dato es que antes de cargarlo no puede ser entendible a simple vista, sólo se muestra una especie de fotografía de los datos, esta herramienta convierte esa fotografía asociándola con el archivo asignado a cada una y carga los datos de manera que puedan ser usados para el análisis.

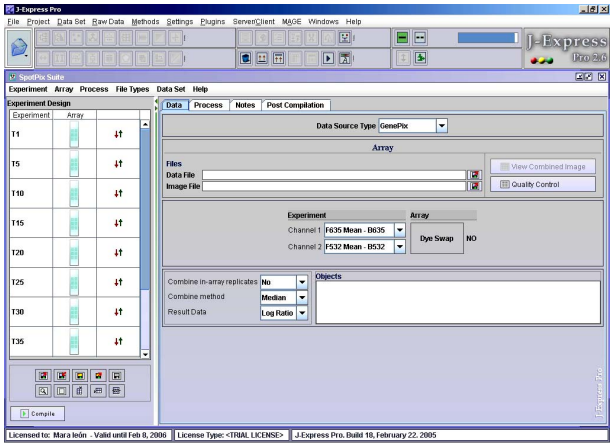


Figura 5.2: Ejemplo de visualización de datos de tipo raw data en J Express

La importancia de esta herramienta radica en la facilidad de poder interpretar este tipo de datos para analizarlos sin necesidad de entender el dato en bruto. Para la mayoría de los usuarios entender un tipo de dato tabular no resulta en mayor complicación, se sabe que las filas y columnas representan información que puede ser leída con mayor facilidad, en cambio los datos en crudo no poseen una fácil interpretación, la ventaja de esta herramienta es que se pueden trabajar ambos tipos de dato sin necesidad de inrepretarlos antes de ser procesados.

Otro aspecto interesante de esta herramienta es que permite visualizar los datos antes de aplicarles cualquier técnica, lo que brinda la oportunidad, en el caso de que alguna persona quisiera cambiar o limitar el tamaño de la base de datos, de trabajar con una cantidad menor de datos, tomando sólo aquellos que sean más significativos para realizar el análisis y reduciendo con ello el tiempo y gasto en recursos.

La manera en la que funciona esta herramienta es simple, para los datos de tipo tabular no existe mayor problema que especificar el formato del dato con el que se va a trabajar para que éstos automáticamente se carguen y se ordenen de manera que parece que se está trabajando con una hoja de cálculo. Después de esto se tiene la opción de delimitar la base de datos o trabajar con ella de manera completa.

Para los datos conocidos como raw data, además de especificar el dato con el que se va a trabajar, se deben vincular las imagenes de los microarrayas y asociarlos a sus respectivos archivos (.gpr) para que puedan ser cargados, note que la representación visual de este tipo de datos no es la misma que la de los datos de tipo tabular; una vez que los datos han sido cargados, la herramienta se encarga internamente de filtrarlos y normalizarlos para generar una matriz de datos que pueda ser utilizada para el análisis, esta matriz de datos no se muestra, se sabe que internamente la herramienta hizo esta conversión, esto tiene sentido si se sabe que para aplicar cualquier método de evaluación los datos deben tener un formato de tipo númerico que permita realizar diversas operaciones sobre ellos.

Otro aspecto interesante en la carga de datos ocurre en la herramienta MEV, esta herramienta es parecida a J Express porque también tiene la particularidad de aceptar ambos formatos en el tipo de datos, lo que la distingue es la visualización de los mismos.

En el caso de los datos de tipo tabular la forma de cargarlos es muy parecida al de J Express, se elige el tipo de formato, y en seguida éstos se presentan en una tabla, pero una vez que se cargan aparece una representación gráfica de los mismos como en una especie de dendograma con colores diferentes que representan la variabilidad de los datos.

Para el caso de los datos de tipo raw data, la idea es la misma, se selecciona el tipo de formato, pero en este caso al seleccionar el formato del dato, en lugar de que éstos aparezcan como en una tabla, sólo aparece una lista de los archivos .gpr, pero al momento de cargarlos la representación es la misma que como en los datos tabulares, la salida es un dendograma. La figura 5.3 muestra una representación visual de la manera en la que se muestran los datos ya cargados en esta herramienta.

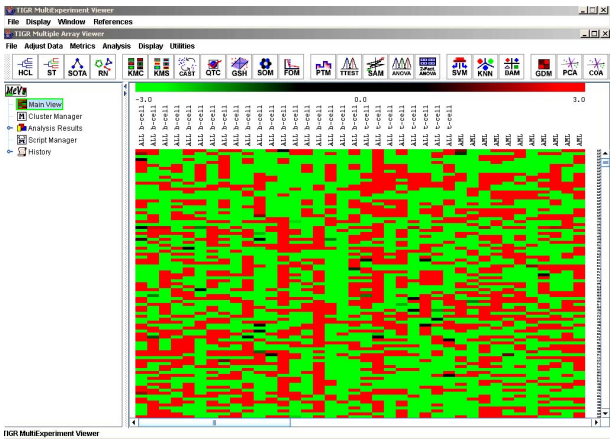


Figura 5.3: Ejemplo de entrada de datos de cualquier tipo en MEV

Una ventaja de MEV con respecto a J Express, está en la visualización, para un experto en el área será mucho más fácil interpretar los datos si éstos estan representados en un dendograma, ya que los colores mostrados en el, seguramente dicen mucho más que si sólo se presentan en una hoja tabulada.

En general podemos decir que estas dos herramientas ofrecen la ventaja de que además de trabajar con ambos tipos de datos y algunos otros que no se han utilizado en este trabajo, ofrecen la oportunidad de poder visualizar los datos con los que se está trabajando antes de aplicarles cualquier evaluación, lo que permite si se quiere, reducir la dimensión de los datos. Estas dos herramientas son las únicas que permiten visualizar los datos, Cluster, GEPAS y EPCLUST no ofrecen esta opción.

5.2. Resultados del Agrupamiento jerárquico

La idea de realizar un agrupamiento jerárquico es la de agrupar objetos en clases tales que los de una misma clase presenten un alto grado de asociación natural entre sí, mientras que las clases sean relativamente distintas unas de otras. Dicho de otra manera, se trata de encontrar agrupamientos naturales en un conjunto de objetos, de forma que la descripción de éstos se realice en términos de clases o grupos de objetos con fuertes semejanzas internas.

Como se ha mencionado, se aplicó el agrupamiento jerárquico aglomerativo, donde la idea es que los datos se agrupen de manera natural mediante esta jerarquía que indica que el modo de agrupamiento será, empezando desde n clusters hasta llegar a obtener un sólo cluster, la figura 5.4, muestra esta idea, este resultado se obtuvo de aplicar el agrupamiento jerárquico para la base de datos Leucemia utilizando la herramienta GEPAS.

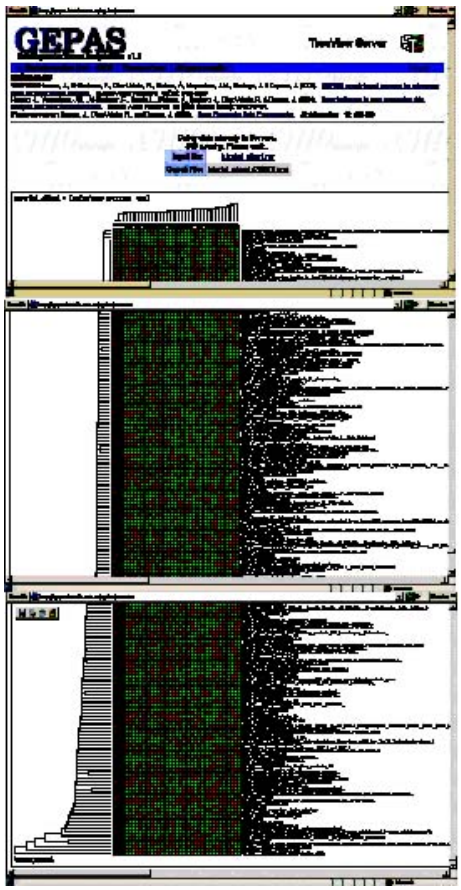


Figura 5.4: Resultado de la evaluación del agrupamiento jerárquico en GEPAS

Como se puede observar en la figura 5.4, el resultado del agrupamiento jerárquico es un árbol llamado dendograma que muestra la relación que existe entre el conjunto de datos de entrada agrupados en clusters. Las hojas que se observan en el dendograma representan a los datos de entrada y el nodo raíz es el resultado final del agrupamiento que contiene a todas las hojas, una rama en el dendograma es un punto en donde dos clusters se han unido.

Por lo general en este tipo de dendogramas cada color representa los valores de los datos de entrada, es decir, el color rojo generalmente indica una similaridad o correlación alta y de valor positivo, el color verde una similaridad baja y de valor negativo y los colores más obscuros como el negro o café suelen representar valores cerca del 0.

En este sentido todas las herramientas funcionan de la misma manera, todas hacen básicamente la misma operación y el resultado es similar, no existen grandes diferencias, excepto en las características físicas de cada dendograma, como por ejemplo, el color, la forma, etc.

El evaluar datos de diferentes tipos ya supone una salida distinta, la representación visual varía debido a la magnitud y naturaleza de los datos, sin embargo, el resultado que se obtuvo de la evaluación de cada herramienta con esta técnica es muy parecido.

La figura 5.5, muestra el resultado de la evaluación del agrupamiento jerárquico en MEV aplicado en la base de datos Leucemia, que es la misma utilizada en la herramienta anterior. Nuevamente se observa como la salida es un dendograma que cumple con características similares al dendograma obtenido por la herramienta GEPAS, la gama de colores es la misma, predominan el rojo que indica una correlación alta positiva y el verde que indica una correlación alta negativa, en algunos puntos se observa el negro que indica los valores cercanos a 0.

La forma de las ramas y del árbol, presentan características distintas, en el árbol obtenido en GEPAS la mayor concentración de conglomerados se encuentra hacia abajo y en el resultado obtenido en MEV es lo contrario, se encuentra hacia arriba, pero el resultado es similar.

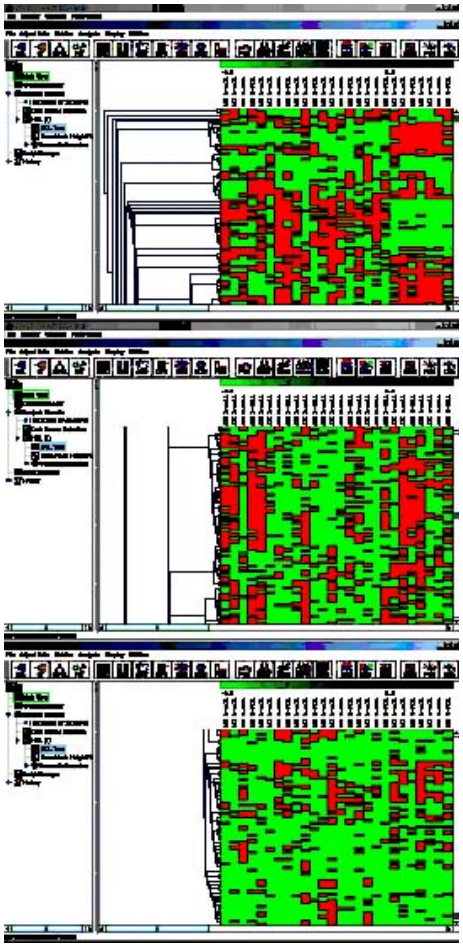


Figura 5.5: Resultado de la evaluación del agrupamiento jerárquico en MEV

5.3. Resultados del K-means

De la evaluación del K-means, se han obtenido resultados significativos que permiten comprender de mejor manera el funcionamiento de este método, independientemente de la herramienta seleccionada para su análisis.

La figura 5.6 muestra el resultado de aplicar la evaluación del K-means en la herramienta EPCLUST sobre la base de datos Levadura, este resultado consta de cinco gráficas, las dos primeras

representan los seeds y los centros finales, las otras tres muestran el resultado del agrupamiento en tres clusters, la primera gráfica de lado izquierdo muestra que para el primer cluster se agruparon 6208 genes, en el segundo se agruparon 10 y en el tercero solamente 4.

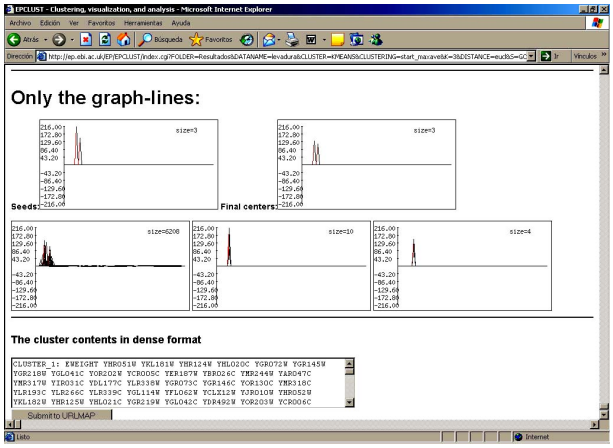


Figura 5.6: Resultado de la evaluación del K means en EPCLUST

¿Porqué existe una diferencia significativa en la cantidad de datos agrupados en cada uno de los clusters?, teóricamente sabemos que el algoritmo K-means empieza con una muestra de K datos elegidos al azar, cada uno de ellos se utiliza como centroide inicial de los K clusters que se van a formar, entonces la matriz de distancias se calcula desde dicho centroide hasta cada uno de los datos y cada uno de ellos será asignado al centroide más cercano, así la matriz de distancias se recalcula reemplazando cada centroide por la media de los datos asignados a él y el algoritmo repite el proceso anterior hasta no cambiar de medias, entonces debido a esto sería de suponer que la cantidad de datos en cada cluster debería ser más o menos equitativa.

Una de las razones por las que esto no ocurra así podría ser por la distancia que hay desde el centroide hasta los datos, es decir, si la distancia es demasiado grande, es probable que la mayor parte de los datos queden en un mismo cluster o por el contrario, si la distancia es demasiado pequeña se pueden obtener mayor cantidad de datos repartidos en varios clusters.

Sin embargo, analizando los resultados, se ha concluido que la razón por la que puede pasar este tipo de agrupamientos está dada por la inicialización de los centros, es decir, la manera en la

que el algoritmo decide como selecionar los centros. Por ejemplo, EPCLUST trabaja de manera interna, no se sabe nada de la inicialización, únicamente se específica el número K de clusters en los que se quiere que se realize el agrupamiento y se especifica la medida de distancia deseada, pero se desconoce la forma en la que inicia los centros.

En todas las herramientas evaluadas sucede lo mismo, excepto en J Express que es la única herramienta que ofrece la opción de especificar la inicialización de los centros, si bien es conocido que el resultado final depende del valor de K, el mayor problema esta relacionado con la inicialización de los centros.

El algoritmo K-means en J Express presenta una variación con respecto al algoritmo usado comúnmente, que es el que suponemos utilizan todas las demás herramientas, la variante consiste en que este algoritmo se ejecuta múltiples veces con valores diferentes de K y al final concluye con el K óptimo, es decir, fija los centros y distribuye, recalcula los centros, vuelve a distribuir y al final conluye con el mejor de los resultados de todas las pruebas realizadas, esto se logra gracias a que el algoritmo ofrece diferentes acercamientos para la inicialización que permite elegir como localizar o establecer los centros:

Inicialización de centros en J Express

1. *Random:* divide la entrada en particiones de k clusters al azar. Este es el método común.
2. *Forgy:* elige k entradas al azar como centros y asigna el resto de la entrada al centroide más cercano.
3. *Macqueen:* elige k entradas al azar como centros y asigna el resto de la entrada al centroide más cercano, siguiendo el orden de la instancia. Recalcula los centroides después de cada asignación.
4. *Kaufman:* el agrupamiento inicial es obtenido por la selección sucesiva de la entrada representativa hasta que se han encontrado los centros iniciales de k. El primer representante es

el punto más central de la entrada. El resto de los representantes se selecciona según la regla heurística de elegir los casos que prometen tener alrededor de ellos un número más alto del resto de los casos, y tiene una distancia relativamente grande de representantes ya elegidos.

Para evaluar J Express se utilizó el acercamiento Forgy, únicamente por hacer una variación con respecto a la inicialización usada por default que es la Random, y para comprobar si existen cambios significativos al establecer la inicialización. El resultado obtenido de evaluar el K-means para la base de datos Levadura usada en la evaluación de la herramienta EPCLUST se muestra en la figura 5.7.

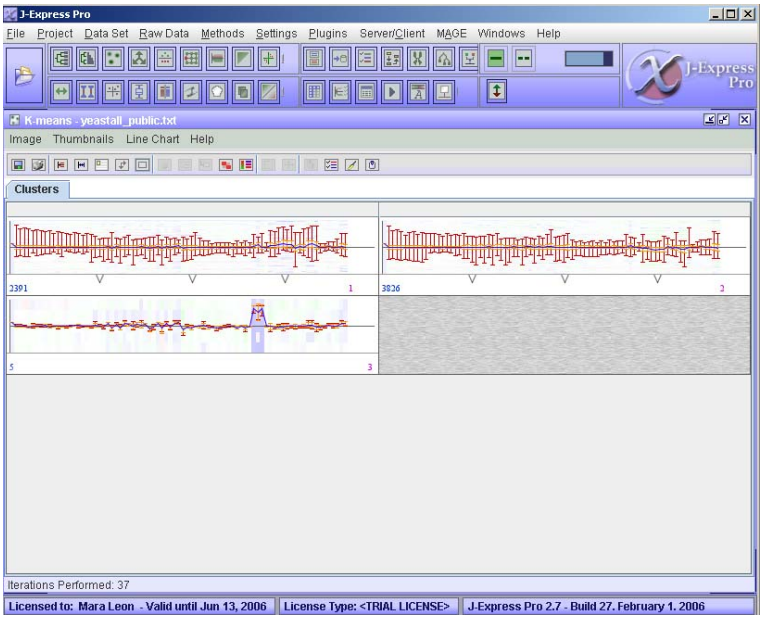


Figura 5.7: Resultado del K-means en J Express usando inicialización Forgy

Como se puede observar, existe una diferencia significativa en el número de genes agrupados por cada cluster, mientras que con la herramienta EPCLUST tenemos 6208 genes agrupados para el primer cluster en J Express tenemos 2391 genes, para el segundo cluster tenemos 10 genes contra 3826 y para el tercer cluster se tienen únicamente 4 genes contra 5. Este último valor no varía mucho, pero la diferencia entre los otros dos es muy grande.

Ahora bien, estamos asumiendo que la inicialización aleatoria es la que se usa comúnmente en todas las demás herramientas, pudiera ser que no es así y que esta inicialización para J Express cambia de alguna manera puesto que el resultado de evaluar ésta en J Express para la misma base de datos es el que se muestra en la figura 5.8.

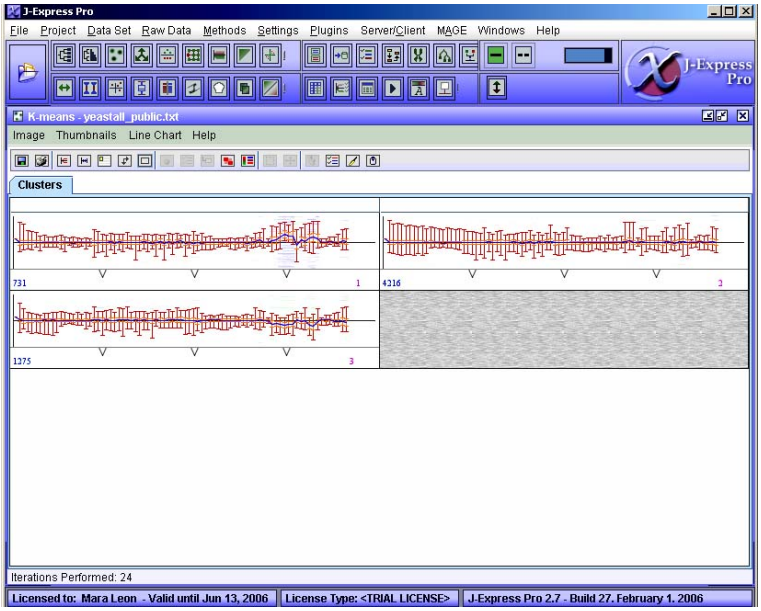


Figura 5.8: Resultado del K-means en J Express usando inicialización Random

Se observa una diferencia significativa tanto para el resultado de la evaluación con la misma herramienta como para el resultado de la evaluación con la herramienta EPCLUST, aquí cambia por completo la cantidad de datos agrupados por cada cluster, para el primer cluster se tienen 731 genes, para el segundo 4216 genes y para el tercero 1275 genes.

De esto resulta que se hicieran las demás evaluaciones correspondientes a cada opción de inicialización, en el caso de la inicialización Macqueen el resultado del agrupamiento es el mismo que el obtenido por la inicialización Random, pero en el caso de Kaufman el resultado presenta una variación que se muestra en la figura 5.9.

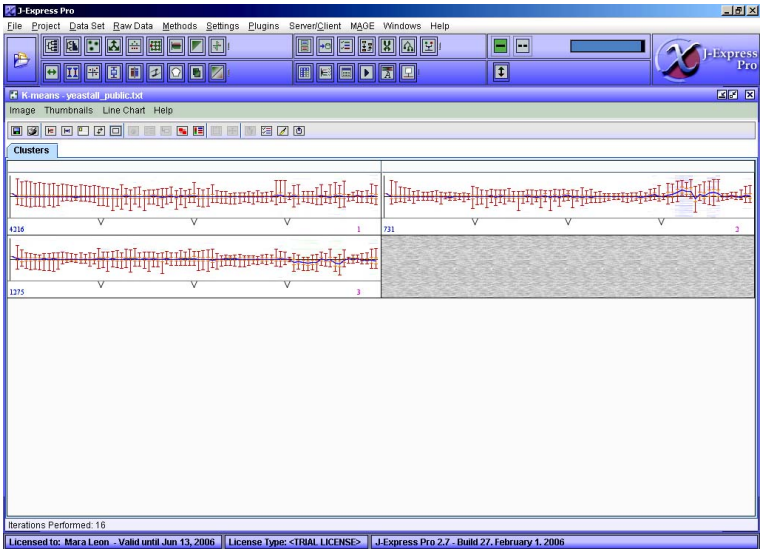


Figura 5.9: Resultado del K-means en J Express usando inicialización Kaufman

Se observa que el agrupamiento ha cambiado una vez más, ahora se tiene que para el primer cluster la cantidad de datos es de 4216 genes, para el segundo es de 731 genes y para el tercero es de 1275 genes. Al parecer en los tres acercamientos de inicialización sólo se invierte el orden en que los datos son agrupados en cada cluster, excepto en la inicialización Forgy en donde el número de datos para cada cluster es diferente.

Analizando esta situación podemos decir que la manera de inicializar los centros es determinante para obtener buenos resultados, y esto lo podemos observar a partir de estos resultados, en donde utilizando la misma base de datos, la misma herramienta y sólo variando la inicialización los resultados han sido diferentes, en este caso el establecer que inicialización se debe seguir dependerá del tipo de resultado que se está buscandoy del tipo de estudio que se esté realizando.

La figura 5.10 muestra otro tipo de resultado que es completamente diferente a los presentados anteriormente, este resultado corresponde al obtenido por la herramienta MEV, con la base de datos Levadura usada hasta ahora para explicar el K-means.

Observando este resultado, notamos que éste viene dado en forma de dendograma, que es una

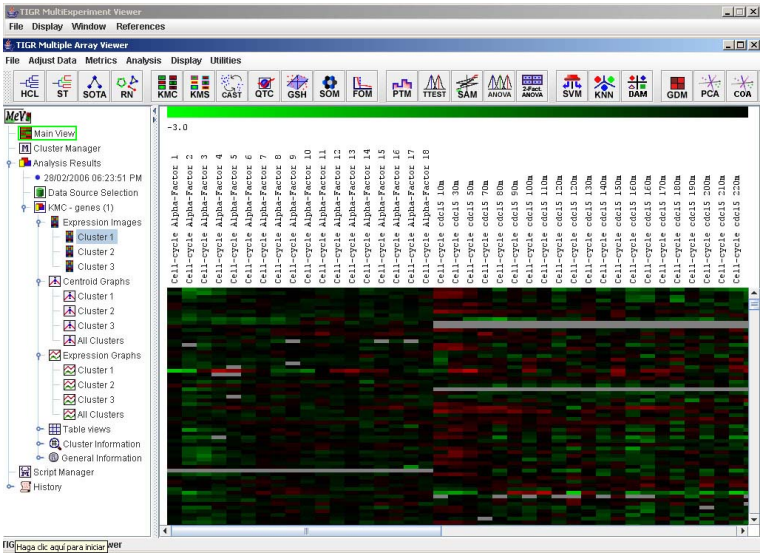


Figura 5.10: Resultado de la evaluación del K-means en MEV

representación gráfica del conjunto de datos, estos dendogramas vienen representados por una serie de colores que describen los valores de cada dato, por lo general los colores más obscuros indican valores bajos y los colores brillosos o más claros indican valores altos, es decir que comúnmente los colores obscuros indican valores negativos y los colores claros valores positivos.

Una diferencia con respecto a la visualización de los datos es que en esta herramienta no se pueden ver los tres clusters al mismo tiempo sino uno por uno, además de que tampoco se tiene conocimiento de la cantidad de datos agrupados por cada cluster, esta figura representa sólo el resultado del primer cluster, lo cual podría ser en cierta forma una desventaja con respecto a las otras herramientas, ya que aunque esta manera de visualización pueda ser más fácil de entender para un biólogo o experto en el área el no poder conocer la cantidad de genes agrupados quizá pueda representar algún problema.

Aunque contrarrestando esta desventaja, existe una opción que minimiza este problema, y es que si bien esta herramienta presenta el resultado de cada cluster por separado también ofrece la opción de ver las gráficas de los centros, y en este caso se puede conocer la cantidad de genes

agrupados por cada cluster. La figura 5.11 muestra la gráfica de centros obtenida para este conjunto de datos.

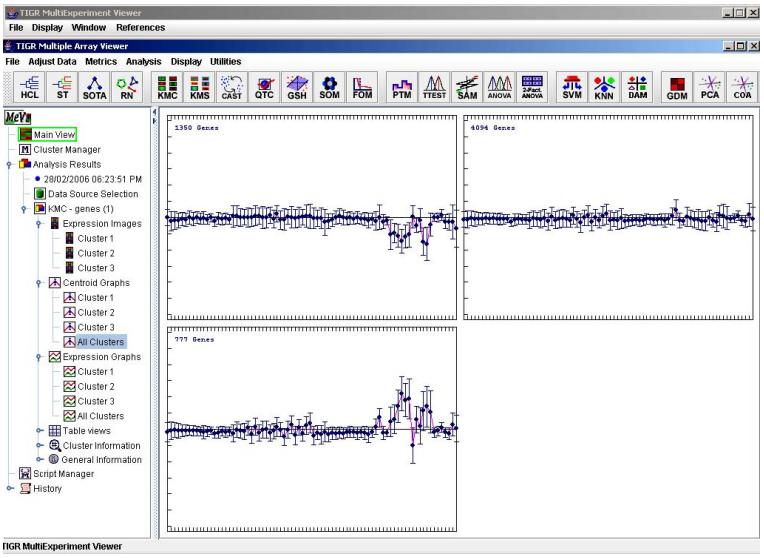


Figura 5.11: Gráfica de centroides obtenida de la evaluación del K-means en MEV

En esta figura podemos observar como han sido distribuidos los datos para cada cluster, tenemos que para el primer cluster se tienen 1350 genes, para el segundo 4094 y para el tercero 777. Si comparamos este resultado con el obtenido de las herramientas anteriores vemos que la diferencia entre ellos sigue siendo significativa, hasta ahora ninguno de ellos ha sido similar excepto los obtenidos en J Express en donde en algunos casos sólo varía el orden del agrupamiento.

Entonces, ¿qué significa todo esto?, ¿porqué existe una variación tan drástica en el agrupamiento de los datos?, podría pensarse que la herramienta no es lo suficientemente buena para realizar tal agrupamiento, o que el tipo de dato utilizado para la evaluación no es el adecuado ni para la herramienta ni para el método, pero esto no es así, independientemente del tipo de dato y de la herramienta, la diferencia radica en la inicialización.

Se sabe que el método funciona y que las herramientas no tienen ningún problema, el problema está, en que en la mayoría de ellas se desconoce la inicialización del algoritmo, aunque los centros

iniciales pueden seleccionarse aleatoriamente, si estos se seleccionaran de manera que se dispersaran uniformemente, el resultado final sería más preciso; como no sucede así, lo importante es no centrarse sólo en el algoritmo, sino más bien, enfocarse a la manera que tiene cada herramienta de visualizar los resultados, es decir, preguntarse de que forma un biólogo o experto en el área puede entender de mejor manera el resultado que se le está presentando.

Se podrían hacer múltiples comparaciones entre los resultados del k-means para cada herramienta, pero se han elegido los más significativos porque estos representan el concepto general de la técnica, los resultados de la herramienta GEPAS Y Cluster son parecidos a MEV, presentan los resultados cluster por cluster de forma sinilar y con el mismo tipo de visualización que es una especie de dendograma, una desventaja de estas dos herramientas es que no se tiene la opción de ver las gráficas de los centros, lo que provoca que no se tenga conocimiento de la cantidad de genes agrupados por cada cluster.

5.4. Resultados del PCA

La idea de aplicar un PCA es la de reducir la dimensionalidad de un conjunto de datos. Esta técnica se puede complementar con otras como el K-means, de hecho se recomienda aplicar primero un PCA antes de aplicar el K-means con el objetivo de que los datos resultantes del PCA al ser reducidos sean los más significativos y el resultado obtenido del K-means sea el más óptimo.

La figura 5.12 muestra el resultado de aplicar el PCA sobre la base de datos Linfoma usando la herramienta J Express, se observa este resultado representado por una gráfica bidimensional que muestra los componentes principales que contienen la varianza más alta, se observa que el primer componente principal contiene el 35.63 % de la varianza, mientras que el segundo componente principal contiene el 7.181 %, la varianza total corresponde al 42.8 %.

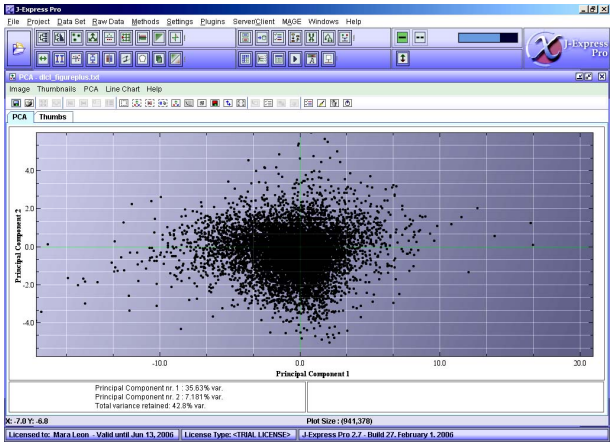


Figura 5.12: Resultado de la evaluación del PCA en J Express

A diferencia de J Express, la herramienta Cluster no ofrece la opción de ver los resultados de manera gráfica, la función de esta herramienta consiste en encontrar los valores propios (eigenvectores) calculando la descomposición del valor singular de la matriz de los datos. La salida es muy simple, solamente se generan dos archivos, uno que contiene los componentes principales y el otro que contiene la información de cada gen en los componentes principales, la figura 5.13 muestra la matriz de datos formada por los valores propios.

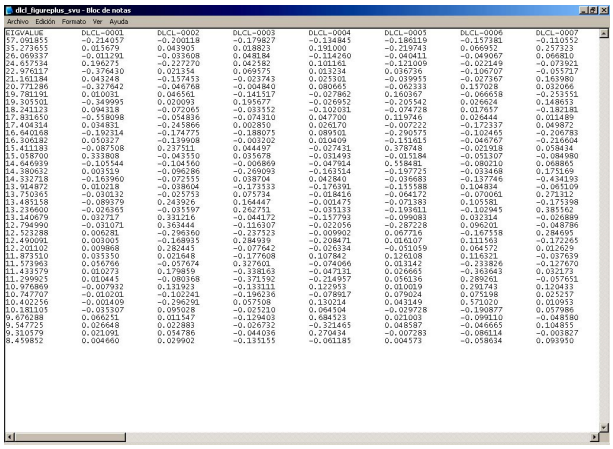


Figura 5.13: Resultado de la evaluación del PCA en Cluster

De este resultado no se puede concluir mucho, al no tener un resultado gráfico no es tan sencillo determinar como están representados los componentes principales ni el porcentaje que ocupan con respecto a la varianza, esto no quiere decir que la herramienta sea mala, sino simplemente es mucho más simple y por lo mismo más sencilla de usar que cualquiera de las otras.

Otra herramienta que presenta resultados con respecto al PCA es MEV, la diferencia de esta con respecto a J Express y a Cluster es que de cierta manera MEV es una combinación de las otras dos, no sólo presenta los resultados de manera gráfica sino que también es posible analizar la información de cada componente.

La figura 5.14 muestra el resultado gráfico del PCA en MEV, una desventaja de esta herramienta en la visualización del resultado gráfico en comparación con J express es que esta no muestra ninguna información con respecto a la varianza, sólo se observa la gráfica que muestra la dispersión de los datos en donde el eje de las Y corresponde al primer componente principal y el eje de las Y al segundo componente principal, pero no se conoce que porcentaje de la varianza le corresponde a cada componente.

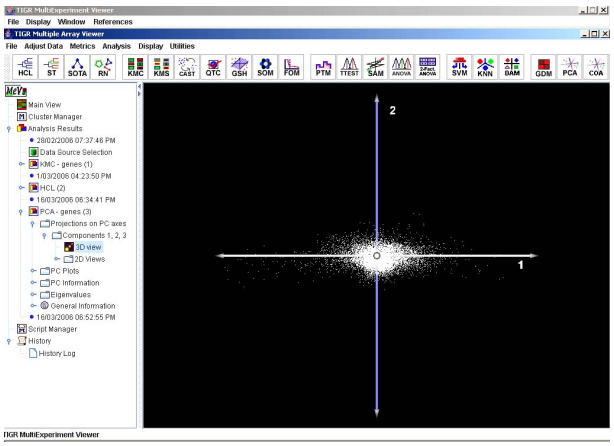


Figura 5.14: Resultado de la evaluación del PCA en MEV

La figura 5.15 muestra la información correspondiente a cada componente principal y la varianza correspondiente a cada uno, esta opción no la ofrece J express, y aunque Cluster muestra la matriz

de datos de los valores propios tampoco menciona que porcentaje de la varianza corresponde a cada componente.

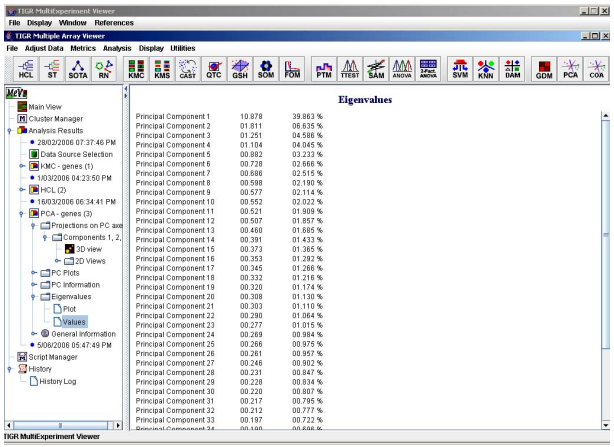


Figura 5.15: Resultado del porcentaje de la varianza en MEV

Esta es una de las herramientas más completas con respecto a la evaluación del PCA, ya que no sólo muestra la gráfica generada con los componentes principales, y el porcentaje de la varianza por cada componente, sino también ofrece la opción de visualizar una representación gráfica de la combinación del primer componente con el segundo y el tercero, y el segundo componente con el tercero, además es posible analizar los valores propios generados por cada componente así como cada una de sus gráficas y una información general sobre el tiempo de ejecución, el número total de componentes, etc.

Aunque todas las variables entran en la composición de cada componente principal, algunas son más importantes que otras, y éstas son las que determinan la naturaleza de cada componente, es por eso que los datos de la varianza son importantes para decidir que componentes principales se deben utilizar en el análisis de los datos. Como las variables son ordenadas de acuerdo a la magnitud de su varianza, es posible concluir que las primeras componentes principales bastan para describir en alto porcentaje la variabilidad total de las variables originales. Con frecuencia sucede que las primeras 2 o 3 componentes principales son las más significativas.

Capítulo 6

Análisis de Resultados

Las comparaciones entre las diferentes técnicas de agrupamiento suelen hacerse en base a los resultados obtenidos sobre cada uno de ellos para los diferentes problemas, principalmente por la diferencia que existe entre cada técnica para realizar el agrupamiento, algunos de los resultados que pueden compararse son, entre otros, la manera de visualización de los resultados, es decir, la facilidad de interpretación que éste represente para el usuario final, la similitud en la forma de agrupar los datos, el tiempo que tardan en realizar los agrupamientos, etc.

No obstante, algunas medidas pueden ayudar en el proceso de agrupamiento para seleccionar los parámetros más adecuados o para dirigir el algoritmo a diferentes situaciones. Una de estas medidas es la distancia entre los centros, tal como lo vimos con el k-means, en donde para la herramienta J Express, se tuvo la opción de usar las primeras k bservaciones, elegir aleatoriamente las k observaciones o tomar cualquier partición al azar en k clusters y calcular sus centroides. Esto ya representa diferencia significativa en la obtención de los resultados.

Hacer una comparación clara entre el agrupamiento jerárquico, el k-means y el PCA no es simple, se tiene que tomar en cuenta que cada algoritmo realiza sus agrupamientos de manera diferente, la aplicación de cada uno depende del resultado que se quiera obtener y del grupo de datos con el que se esté trabajando.

Algunas ventajas y desventajas de un algoritmo con respecto al otro son:

Agrupamiento jerárquico aglomerativo

1. Ventajas

   - los dendogramas son una representación más fácil de interpretar.
   - no se tiene que decidir acerca del número de clusters.

2. Desventajas

   - la decisión inicial influye mucho en el resultado.
   - puede ser muy lento el obtener el resultado, consume más tiempo de ejecución.
   - es dificil corregir lo que se hizo por la rígidez que le da la estructura de árbol.
   - existen problemas cuando existen datos faltantes o datos que contienen un alto nivel de error.
   - el dendograma correspondiente a un agrupamiento jerárquico no es único pues en cada junte de clusters no se conoce que sub-árbol va a la derecha y cuál a la izquierda.

K-means

1. Ventajas

   - es computacionalmente más rápido y fiable.
   - puede trabajar bien con datos faltantes (missing values) o con datos que contienen altos niveles de error.

2. Desventajas

   - necesita un valor inicial del número de clusters.
   - puede llegar a tomar mucho tiempo obtener los clusters, pero esto depende de la magnitud de los datos.

- si se tiene la opción de elegir la inicialización de los centros, el resultado para cada una es diferente, por lo que la decisión inicial influye mucho en el resultado.

PCA

1. Ventajas

   - reduce la dimensión de un conjunto grande de datos en un nuevo conjunto (más pequeño) sin perder una parte significativa de la información original.
   - incrementa la eficiencia computacional de la clasificación porque reduce la dimensionalidad de los datos.
   - como entrega resultados jerarquizados se aprecia que los primeros componentes concentran una proporción importante de la variabilidad de los datos.
   - el gráfico de dispersión generado por el PCA demuestra particularidades en la información porque comienzan marcando los diferentes puntos de información por lo que es fácil entender el resultado.
   - puede trabajar en conjunción con las técnicas de agrupamiento para obtener mejores resultados.

2. Desventajas

   - para dar una buena interpretación al resultado se debe tener conocimientos de estadística descriptiva y correlación lineal múltiple.
   - las variables deben ser correlacionadas y de varianzas parecidas.
   - en algunos casos se necesita específicar el número de vecinos para la imputación KNN [Cap. 4, Sec. 4.2.2].

6.1. Análisis de las Técnicas de Agrupamiento

El siguiente análisis permite comparar los resultados obtenidos después de aplicar las técnicas del agrupamiento jerárquico, el k- means y el PCA sobre los datos de entrada; como sabemos, cada técnica utiliza un algoritmo de agrupamiento diferente, el objetivo es demostrar que el agrupamiento realizado por cada una de estas técnicas es similar, independientemente de la forma en que lo realize.

Para iniciar el análisis se tomó como dato de entrada la base de datos levadura, para realizar la evaluación se usó la herramienta J Express y se aplicaron las técnicas del agrupamiento jerárquico y el k-means. En el caso del agrupamiento jerárquico se obtuvo un dendograma y para el k-mens un conjunto de 3 clusters. Como el objetivo es demostrar que el agrupamiento realizado por estas técnicas es similar, la idea es comparar los conjuntos de genes agrupados por el k-means y por el agrupamiento jerárquico y buscar que sean similares.

Para hacer esta comparación, se eligió un gen que representara al conjunto de datos agrupados, este gen es el *808 YCR014C POL4 DNA repair DNA polymerase IV S0000607* y tiene por características un identificador, un nombre y una descripción.

Seleccionado el gen YCR014C se buscó en los resultados obtenidos de las agrupaciones de ambas técnicas. La figura 6.1 muestra la ubicación del gen YCR014C en el resultado obtenido por el agrupamiento jerárquico aglomerativo, la figura muestra una ampliación del árbol que permitió encontrar este gen.

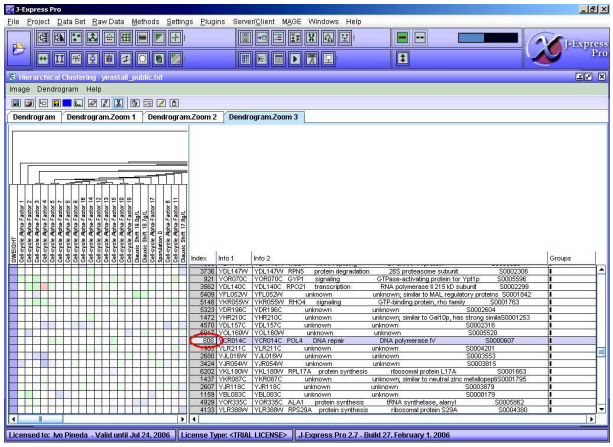


Figura 6.1: Análisis del gen YCR014C en el agrupamiento jerárquico

La figura 6.2 muestra la ubicación del gen YCR014C en el resultado obtenido por el k-means en el cluster 1.

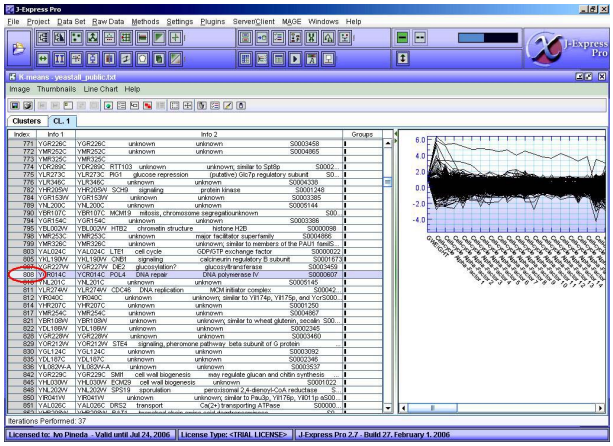


Figura 6.2: Análisis del gen YCR014C en el k-means

Observe que en ambas figuras la distribución del gen YCR014C con respecto a los demás es diferente, los genes que se encuentran por encima y por debajo del gen YCR014C no son los mismos en ambos resultados, por ejemplo, note que en el resultado del agrupamiento jerárquico el gen que está por encima del YCR014C es el YOL160W y el que está por debajo es el YLR211C, en el resultado del k-means el gen que está por encima del YCR014C es el YGR227W y el que está por debajo es el YNL201C, esto es debido a la representación que tiene cada técnica en la obtención del resultado, y no al agrupamiento como tal.

Esta diferencia no indica que el agrupamiento realizado por ambas técnicas es distinto, la diferencia consiste en que, para el caso del agrupamiento jerárquico los genes se encuentran distribuidos de acuerdo a la manera en la que se fué construyendo el árbol, es decir, que los genes se encuentran ubicados en el nivel del árbol correspondiente; en el caso del k-means los genes se encuentran distribuidos de manera ordenada, este orden se ha realizado de manera ascendente de acuerdo al número de identificación de cada gen.

Para demostrar que la representación visual no muestra el resultado real del agrupamiento, se ha usado el archivo de texto generado en la obtención del resultado de los agrupamientos en ambas técnicas. Este archivo de texto es conocido como la representación textual de los resultados del agrupamiento para ambas técnicas, contiene el nombre y la descripción de cada gen, similar al resultado visual de los agrupamientos sólo que estos se muestran exactamente en la manera en que fueron agrupados.

La figura 6.3 muestra la representación textual del resultado del agrupamiento jerárquico, el gen YCR014C se encuentra resaltado en el recuadro azul.

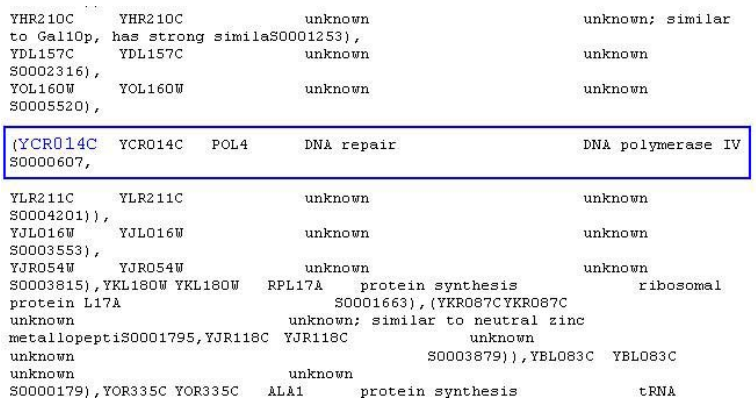


Figura 6.3: Resultado de la representación textual del gen YCR014C en el agrupamiento jerárquico

La figura 6.4 muestra la representación textual del resultado del k-means, el gen YCR014C se encuentra resaltado en el recuadro rojo.

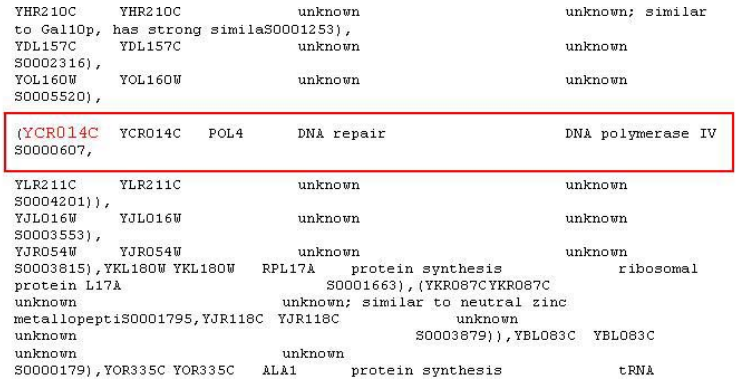


Figura 6.4: Resultado de la representación textual del gen YCR014C en el k-means

De estos resultados se observa que el conjunto de genes agrupados contienen características genéticas similares. En ambos casos el resultado de la representación textual es muy parecido, tanto para el k-means como para el agrupamiento jerárquico el gen YCR014C se encuentra exactamente en la misma posición, y los genes que se encuentran por encima y por debajo de éste también son los mismos. Existe además una coincidencia con el resultado mostrado en la figura 6.1, en donde los genes que se encuentra por encima y por debajo del gen YCR014C son los mismos que los que se encuentran en las representaciones textuales de ambas técnicas.

Con esto comprobamos efectivamente que la funcionalidad de estas técnicas es similar y que a pesar de que el agrupamiento jerárquico y el k-means son técnicas que realizan el agrupamiento de los genes de manera distinta, el resultado final del agrupamiento es similar. En este sentido no importa cual de las dos técnicas se utilize para agrupar datos, se sabe que el resultado será muy similar. Decidir que técnica utilizar dependerá del resultado que se espere, porque aunque el resultado del agrupamiento es similar, la representación visual de éste no lo es, por ello es importante analizar el problema que se esté resolviendo; en algunos casos la representación visual que ofrece el agrupamiento jerárquico puede ser más conveniente que la ofrecida por el k-means o viceversa.

Sin embargo, si la representación de los resultados no es una característica a considerar, entonces se sugiere utilizar la técnica más rápida, si se observa la tabla 4.5 del capítulo 4 sección 4.2.3, observamos que el k-means es más rápido que el agrupamiento jerárquico y por lo tanto más óptimo de utilizar.

Sólo se han comparado el agrupamiento jerárquico y el k-means, la idea principal era hacer una comparación entre el K-means y el PCA y el agrupamiento jerárquico y el PCA, pero no fué posible hacer estas comparaciones, se sabe que el PCA es una técnica eficiente cuando se trata de reducir la dimensionalidad de un conjunto de datos, hubiera sido un buen análisis el compararlo con el k-means y el jerárquico y ver si los datos obtenidos en cada componente principal coincidían con los datos de los agrupamientos, pero desafortunadamente ninguna herramienta proporciona información útil sobre el PCA como para hacer una comparación de este tipo.

Es importante mencionar que si las herramientas que realizan el PCA ofrecieran la opción de obtener una matriz de scores entonces se podría hacer una comparación, la matriz de scores contiene las proyecciones de los datos sobre el nuevo espacio de representación constituido por los componentes principales, esta matriz contiene un número menor de columnas en relación a la matriz original, y estas columnas representan las características más importantes de los datos.

Como esta matriz contendría los genes más significativos de todo el conjunto de datos, se podrían comparar los genes contenidos en esta matriz con los genes obtenidos del agrupamiento jerárquico y el k-means y ver si el gen YCR014C se encuentra ahí, pero como no es posible obtener la matriz de scores, no es posible hacer esta comparación.

6.2. Análisis de Herramientas

Se ha demostrado que las técnicas de agrupamiento obtienen resultados similares, ahora se hace un análisis para comprobar si la manera en la que funcionan estas técnicas es la misma en las distintas herramientas. El objetivo es analizar si tanto para el agrupamiento jerárquico como para el k-means se obtienen agrupamientos similares en cada una de las herramientas a pesar de que cada una de ellas tiene alguna variación en la funcionalidad de cada técnica.

Para iniciar este análisis usamos como dato de entrada la base de datos leucemia, y analizamos únicamente los resultados obtenidos de aplicar el k-means, sólo se usa el k-means porque de acuerdo al resultado obtenido del análisis anterior en la sección 6.1 sabemos que tanto para el k-means como para el agrupamiento jerárquico se obtiene resultados similares, no es necesario evaluarlos en los dos, y porque de acuerdo a la visualización de la mayoría de las herramientas es más sencillo analizar el resultado de los agrupamientos a través de clusters que hacerlo a través de dendogramas.

Tomamos como base el resultado de los agrupamientos en el cluster 1 obtenidos por la herramienta J Express, veremos si los datos contenidos en el cluster 1 por J Express coinciden con los datos obtenidos en el cluster 1 de las demás herramientas.

La figura 6.5 muestra el resultado obtenido por el k-means en la herramienta J Express, analizamos el cluster 1 , se observa que los genes se muestran ordenados de manera ascendente de

acuerdo a su número de identificación, se observa el nombre y la descripción de cada gen.

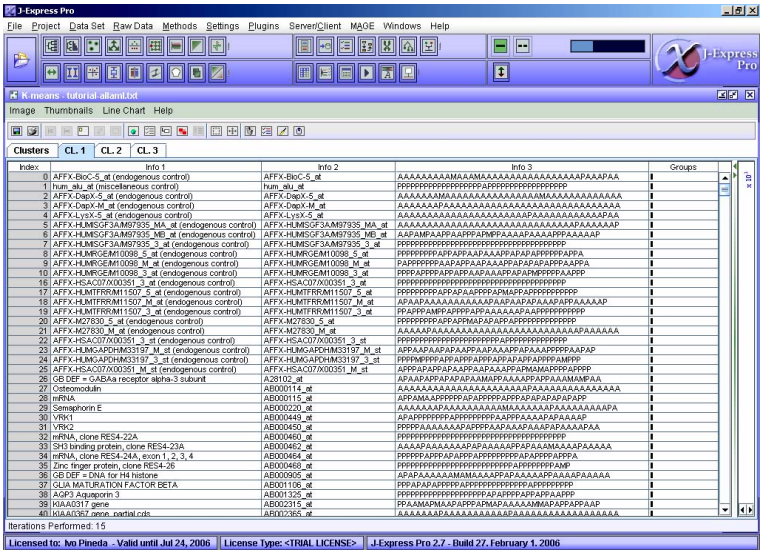


Figura 6.5: Resultado del agrupamiento del k-means en J Express

Analizando el resultado del agrupamiento obtenido por el k-means en la herramienta MEV, notamos que los datos agrupados en el cluster 1 por J Express no son los mismos que los agrupados por MEV en ese mismo cluster, estos se encuentran agrupados en el cluster 3.

La figura 6.6 muestra este resultado, los genes resaltados en azul son algunos de los genes que coinciden con los genes obtenidos del agrupamiento en J Express en el cluster 1. Si se observa la figura, estos datos coinciden, la diferencia radica en que el número de cluster en el que se agruparon los datos es diferente y en que los datos agrupados en MEV se encuentran dispersos en comparación con los datos agrupados por J Express.

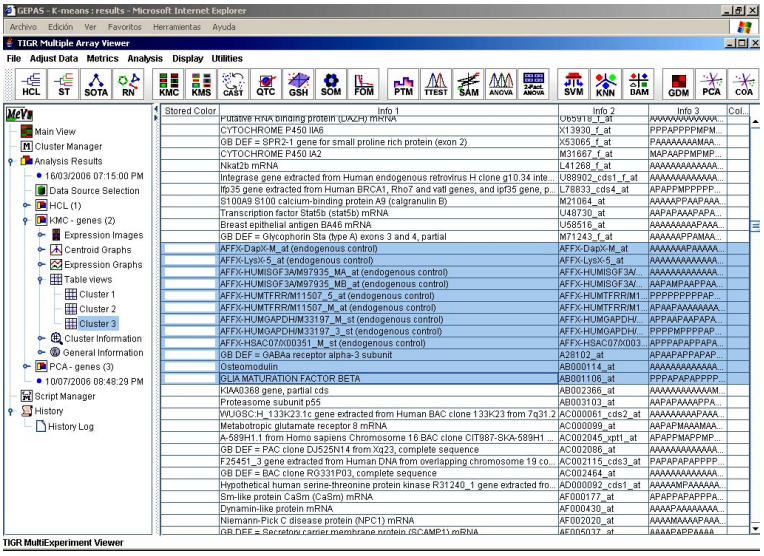


Figura 6.6: Búsqueda de genes como resultado del agrupamiento del k-means en MEV

Analizando el resultado del agrupamiento obtenido por el k-means para la herramienta Cluster & Tree View, vemos que en primer lugar no es tan sencillo como en J Express y en MEV identificar los números de los clusters, esta herramienta presenta los resultados agrupados en una especie de dendograma y la única separación que indica en donde termina y empieza un nuevo cluster es una línea blanca horizontal que divide el dendograma.

En la figura 6.7 se muestra el resultado del agrupamiento obtenido en la herramienta Cluster & Tree View, se observa la línea blanca que indica la separación entre el cluster 1 y el cluster 2, la descripción de los genes que aparece a la derecha indica que éstos se encuentran agrupados en el cluster 2. Si se observan estos genes, nos damos cuenta que elresultado del agrupamiento coincide con el obtenido por J Express, en ambos casos los genes obtenidos son los mismos y también son los mismos que los obtenidos por MEV, sólo varía el orden en el que se encuentran.

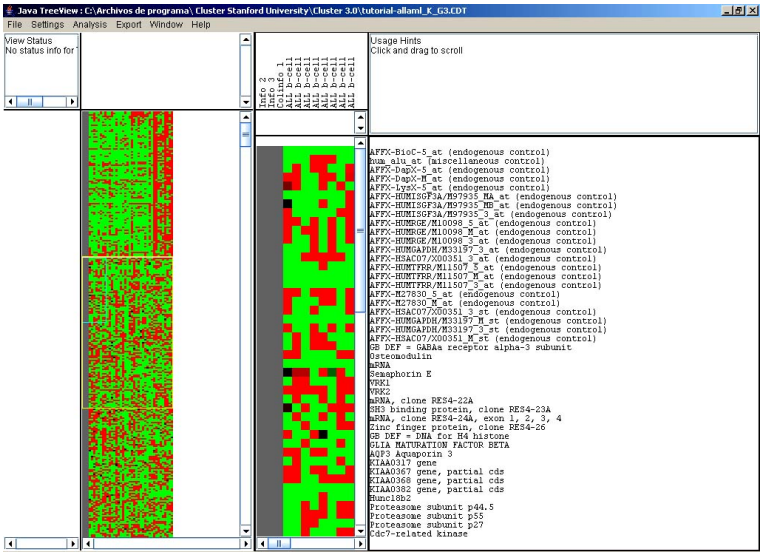


Figura 6.7: Resultado del agrupamiento del k-means en Cluster & Tree View

Esto hace suponer que Cluster & Tree View ordena sus datos obtenidos en el agrupamiento al igual que lo hace J Express, la diferencia radica en el número de clusters en el que fueron agrupados, mientras que en J Express se agruparon en el cluster 1 en Cluster & Tree View se agruparon en el cluster 2. Nuevamente sucede lo mismo que con MEV, aunque los datos coinciden en el mismo grupo han sido colocados en clusters diferentes.

Si analizamos ahora que sucede con el resultado del k-means obtenido por la herramienta GEPAS vemos que en primer lugar la representación de este resultado es completamente diferente al obtenido por las tres herramientas anteriores, GEPAS genera un archivo independiente que contiene las agrupaciones de cada cluster, el archivo contiene el nombre del gen seguido de su descripción y adelante de él un número, este número indica el cluster al que pertenece, si pertenece al cluster 1 se indica con el número 0, si pertenece al cluster 2 se indica con el número 1 y si pertenece al cluster 3 se indica con el número 2. La figura 6.8 muestra este resultado.

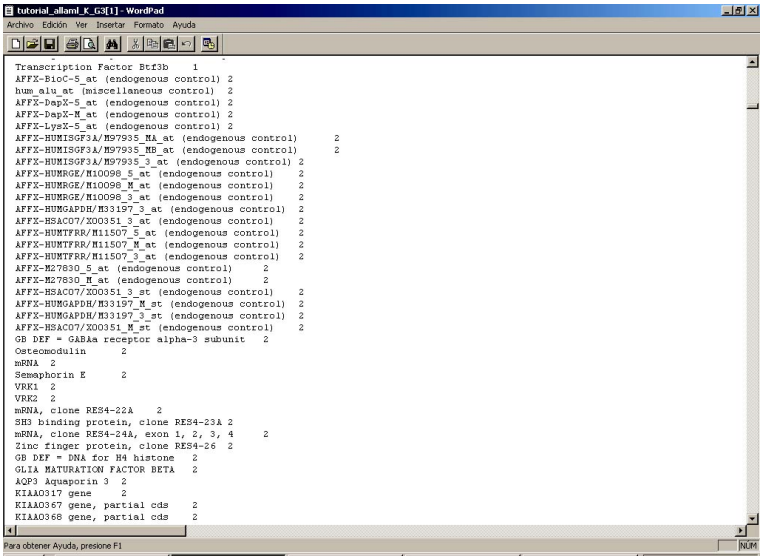


Figura 6.8: Resultado del agrupamiento del k-means en GEPAS

Analizando la figura 6.8, vemos que adelante de la descripción de cada gen se encuentra el número 2, esto quiere decir que los genes se encuentran agrupados en el cluster 3. Si observamos este resultado, vemos que los genes agrupados coinciden con los genes agrupados por J Express, MEV y Cluster & Tree View; al parecer GEPAS también ordena sus datos como lo hace J Express y Cluster & Tree View, mientras que como ya sabemos MEV no. Nuevamente tenemos la diferencia en el número de clusters en el que fueron agrupados los genes, GEPAS tiene sus genes agrupados en el cluster 3 al igual que MEV.

Por último, analizamos el resultado obtenido por el k-means en la herramienta EPCLUST, la representación de los clusters es parecida a la de GEPAS, sólo que en EPCLUST sí se indica claramente en que número de cluster se encuentran los genes. La figura 6.9 muestra el resultado del agrupamiento en EPCLUST.

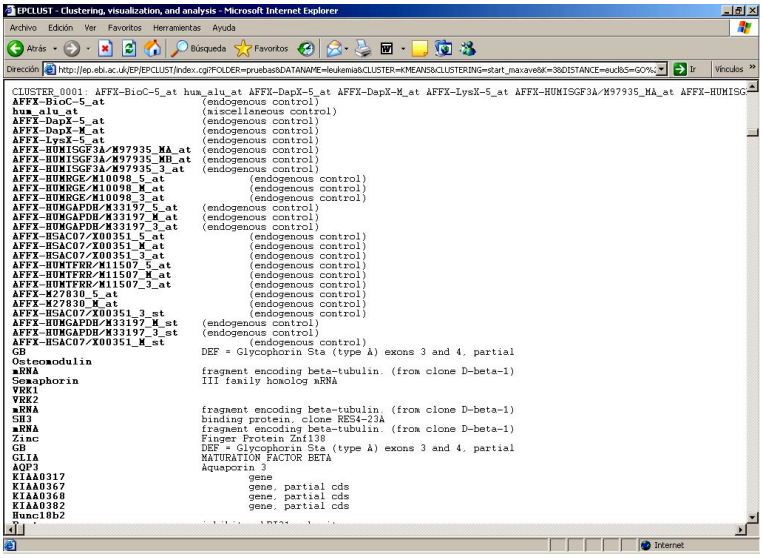


Figura 6.9: Resultado del agrupamiento del k-means en EPCLUST

Como se puede observar en la figura 6.9 los genes se encuentran agrupados en el cluster 1 al igual que en J Express. Este cluster coincide con los clusters obtenidos en J Express, MEV, Cluster & Tree View y GEPAS no sólo en el contenido sino en la forma ordenada de agrupamiento, por supuesto también coinciden con el obtenido por MEV, pero como hemos mencionado MEV no ordena los datos dentro del cluster.

Del análisis realizado en cada herramienta podemos decir que existen diferencias en la forma de visualizar los resultados y en la forma en que estos son ubicados, la diferencia en la ubicación de éstos radica en dos puntos:

- El orden en que son presentados los datos dentro del cluster. J Express, Cluster & Tree View, GEPAS y EPCLUST presentan los genes ordenados de acuerdo a su identificador y MEV presenta los genes desordenados.
- La distribución de los genes en los clusters.

En la tabla 6.1 vemos la distribución de los genes hecha por k-means en cada cluster para las diferentes herramientas, la agrupación de los datos en cuanto al número de cluster coincide para

el cluster 1 en J Express y EPCLUST, para el cluster 2 sólo se obtienen de Cluster & Tree View y para el cluster 3 coinciden MEv y GEPAS. Aunque los genes hayan sido agrupados en diferentes cluster, el contenido de cada uno es similar.

Herramienta	Número de Cluster
J Express	C_1
MEV	C_3
Cluster & Tree View	C_2
GEPAS	C_3
EPCLUST	C_1

Tabla 6.1: Distribución de los datos en clusters

Con el análisis completo de estos resultados, podemos comprobar que el agrupamiento obtenido por el k-means en cada herramienta es muy similar a pesar de que existan algunas variantes en la funcionalidad de la técnica en cada herramienta. Esto quiere decir que no importa que herramienta se elija, el resultado de las técnicas de agrupamiento será similar y es correcto.

Si bien hemos comprobado que las cinco herramientas estudiadas proporcionan resultados similares en cuanto a las técnicas de agrupamiento que utilizan y que este es un factor que hace que posean algo en común, existen otros factores que las hacen muy diferentes las unas de las otras, estos factores permiten decidir que herramienta utilizar en determinada situación para resolver cierto problema. Se mencionan algunas ventajas y/o desventajas entre ellas:

Con respecto al tipo de dato, se sabe que la mayoría de las herramientas soportan varios formatos del tipo de dato de entrada además de los mencionados aquí, tal es el caso de J Express que además de trabajar con datos de tipo tabular (.txt) y tipo raw data (.gpr), es compatible con datos de GenPix, Affymetrix, Agilent, Scanalyze y Array Express, lo mismo sucede con MEV que además de trabajar con formatos .txt y .gpr maneja formatos como el .mev y el .tav generados por TIGR Spotfinder y TIGR MIDAS, y también archivos de Affymetrix y de Genepix. El que J Express y MEV trabajen con herramientas externas les permite hacer análisis más completos, caso

contrario ocurre con Cluster, GEPAS y EPCLUST que tienen la limitante de trabajar únicamente con datos de tipo tabular y que además no permiten interactuar con otros sistemas.

Por otro lado podemos hablar de la interfaz, que hace que una herramienta sea sencilla y fácil de usar. J Express posee una interfaz técnica, lo que complica a simple vista entender su funcionamiento, de debe recurrir al manual de usuario para saber por donde empezar, lo mismo sucede con MEV, son herramientas más científicas, destinadas a un tipo de usuario específico que posee un conocimiento sólido de lo que se está trabajando o solicitando; Cluster por ejemplo, tiene una interfaz muy sencilla y fácil de entender para cualquier usuario, pero necesita trabajar en unión con el software Tree View para poder visualizar los resultados ya que sin este es imposible verlos, en cuanto a GEPAS y EPCLUST, son herramientas en línea que siguen un proceso de análisis, en cada paso se va especificando lo que se necesita para encontrar el resultado.

Un punto a considerar es la instalación del software, se debe cumplir con ciertos requerimientos de hardware para que las herramientas puedan funcionar, como la mayoría de las herramientas usadas son software libre no existen problemas en la instalación del software, la única herramienta que no es del todo libre ya que sólo se puede obtener una versión de prueba es J Express, esta requiere una licencia para poder instalar el software, esta licencia puede obtenerse del sitio donde se descarga el software y sólo esta disponible por un tiempo determinado, una vez vencido el tiempo no es posible continuar trabajando con la herramienta. En el caso de MEV y Cluster & Tree View, sólo se descarga el archivo de instalación del sitio donde se obtiene la herramienta y los pasos a seguir son guíados por el asistente de instalación, para aquellos usuarios que no estén interesadops en instalar un software o que no cumplan con los requerimientos específicos de hardware, existe la opción de trabajar con GEPAS y EPCLUST que son herramientas que trabajan en línea y que por ende no necesitan instalarse, la diferencia entre éstas y aquellas que pueden instalarse es que éstas ofrecen la opción de almacenar los resultados en servidores propios de la herramienta para después accederlos por internet y descargarlos.

Si hablamos de rapidez o consumo de recursos, la herramienta que más tiempo consume en la ejecución de los algoritmos es J Express seguida de MEV, aunque esto es relativo, depende mucho de la técnica que se aplique y de la magnitud de la base de datos sobre la cual se hace el estudio. Si se observan las tablas obtenidas del resultado de la ejecución de cada técnica en las diferentes herramientas, mostradas en el capítulo 4 sección 4.2.3, se tiene que este resultado varía y que en efecto el resultado final en cuanto al tiempo de ejecución depende mucho del tamaño de los datos y de la técnica elegida para analizarlos y no tanto de la herramienta en sí.

Analizando estas ventajas y desventajas, podemos decir que para elegir una herramienta se deben analizar y tomar en cuenta diversos factores, en primer lugar se debe tener un conocimiento pleno de los datos con los que se va a trabajar y tener claro lo que se pretende obtener de ellos. Las herramientas que se estudiaron ofrecen la opción de trabajar con datos de diferentes tipos, habría entonces que decidir que herramienta se acopla más al tipo de datos que se tiene; posteriormente decidir como se desea obtener la visualización de los resultados, como se explicó, la manera en la que funcionan las técnicas en las distintas herramientas es muy similar, por lo que esto no representa una mayor diferencia, pero sí la representa la forma de la visualización, el usuario debe preguntarse de que forma puede darle una mayor y mejor interpretación a los resultados; otro punto a considerar es el tiempo que consume cada técnica en las diferentes herramientas, aunque este punto es importante, no es lo suficiente si se intercambia precisión y entendimiento por rapidez. La decisión radica en el propio usuario, en su manera de trabajar y entender lo que está realizando.

6.3. Análisis de Medidas de Distancia

Se ha analizado y demostrado que las técnicas de agrupamiento jerárquico y el k-means agrupan los datos de manera similar, también se ha analizado y demostrado que estas técnicas evaluadas en diferentes herramientas agrupan los datos de manera parecida. En el capítulo 3 sección 3.2, se mencionó que la medida o métrica de similaridad usada para evaluar los datos es la distancia Euclideana, el análisis que se presenta ahora, pretende demostrar la importancia de utilizar dife-

rentes medidas de distancia para obtener resultados diferentes, el objetivo es probar si el cambiar la medida de similaridad afecta los resultados obtenidos en los dos análisis anteriores con respecto al agrupamiento.

La tabla 6.2 muestra otras medidas de distancia seleccionadas para realizar este análisis, la tabla indica que medida de distancia se encuentra disponible en las diferentes herramientas, en el capítulo 2 sección 2.4 se describe cada una de ellas.

Medida	J Express	MEV	Cluster & Tree View	GEPAS	EPCLUST
Manhattan	√	√	X	X	X
Correlación de Pearson	√	√	√	√	X
Correlación de rangos de Spearman	√	√	√	√	X

Tabla 6.2: Medidas de Distancia

Para iniciar el análisis tomamos como dato de entrada la base de datos leucemia y la técnica del k-means para agrupar los datos. Se probaron las diferentes medidas de distancia en cada herramienta con estos datos y con esta técnica (se excluye la herramienta GEPAS de este análisis por encontrarse fuera de servicio temporalemnte), con el objetivo de analizar si los resultados obtenidos de aplicar diferentes distancias son similares a los obtenidos en la sección 6.1 y 6.2, se plantean dos puntos a considerar:

1. Que los resultados de la distribución de los genes en los clusters dado por el agrupamiento realizado por el k-means en la misma herramienta y con las tres distancias sean diferentes.

2. Que los resultados obtenidos por la misma distancia evaluada en las diferentes herramientas, sea similar, por jemplo, que el resultado del agrupamiento obtenido de evaluar la distancia de Manhattan en J Express y MEV sea parecido.

Con el punto 1 se pretende establecer que el seleccionar medidas de distancia diferentes para evaluar los datos resulta en la obtención de agrupamientos y distribución de genes diferentes.

Con el punto 2 se pretende establecer que a pesar de que las herramientas seleccionadas para el análisis son diferentes, la funcionalidad de las medidas de distancia entre ellas es similar.

Para analizar el punto 1 se muestra la tabla 6.3 que describe el resultado de la distribución de los genes en cada cluster obtenido por la evaluación de las tres medidas de distancia mediante la técnica del k-means en la herramienta J Express.

Medida de distancia	Cluster 1	Cluster 2	Cluster 3
Manhattan	4261	295	74
Correlación de Pearson	1096	2180	1354
Correlación de Spearman	1219	1800	1611

Tabla 6.3: Distribución de clusters en J Express

En la tabla 6.3 podemos ver que la cantidad de genes distribudos como resultado del agrupamiento evaluado para el k-means y aplicado en cada medida de distancia es diferente, la cantidad de genes agrupados en cada cluster varía considerablemente, además se analizaron los grupos de genes asignados a cada cluster y se obtuvo que, por ejemplo, los genes que se encuentran agrupados en el cluster 1 resultado de aplicar la distancia de Manhattan son diferentes a los obtenidos en el cluster 1 evaluados por la distancia de Pearson y lo mismo sucede con los genes agrupados en el mismo cluster por Spearman.

La tabla 6.4 describe el resultado de la distribución de los genes en cada cluster obtenido por la evaluación de las tres medidas de distancia mediante el k-means en la herramienta MEV.

Medida de distancia	Cluster 1	Cluster 2	Cluster 3
Manhattan	4627	0	0
Correlación de Pearson	1071	2162	1394
Correlación de Spearman	1868	1668	1091

Tabla 6.4: Distribución de clusters en MEV

Nuevamente, la tabla 6.4 muestra la diferencia que existe en la cantidad de genes distribuidos en cada cluster, de la misma manera analizamos los genes distribuidos en cada cluster y notamos que la distribucipon de éstos varía de acuerdo a la medida de distancia utilizada, es decir, los genes agrupados en el cluster 2 resultado de la evalución de la distancia de Manhattan no son los mismos que los obtenidos por la distancia de Pearson o Spearman en el mismo cluster, ni siquiera en los resultados obtenidos por esas medidas en otros clusters. El resultado del agrupamiento obtenido por el k-means evaluado en cada medida de distancia es muy diferente.

A pesar de que en la herramienta Cluster & Tree View se pudo obtener resultados de la aplicación de las distancias de Pearson y Spearman, no se obtuvieron resultados con respecto a la distribución de los genes en cada cluster, esta herramienta no proporciona tal información, se sabe cuales son los genes asignados a cada cluster pero no la cantidad de genes que posee cada uno, por lo que en este sentido no pudimos comparar estos resultados para este análisis.

De la comparación de los resultados obtenidos se puede decir que efectivamente la aplicación de medidas de distancia diferentes en la evaluación de los datos afecta significativamente el resultado, no sólo porque la cantidad de genes distribuidos en cada cluster es diferente sino porque los genes asignados a cada uno de ellos también es diferente a pesar de que usan la misma técnica de agrupamiento; con esto queda demostrado el punto 1.

Ahora analizamos el punto 2 que trata de establecer que las mismas medidas de distancia aplicadas en herramientas diferentes ofrecen resultados similares con respecto al agrupamiento de genes en cada cluster, es decir, por ejemplo, que la aplicación de la distancia de Manhattan evaluada en J Express y MEV obtienen resultados de agrupamiento de los genes de manera similar.

Para realizar esto se compararon los resultados de la distribución de los genes en cada cluster obtenidos por cada medida de distancia en las diferentes herramientas y se obtuvo que: En el caso de la distancia de Manhattan aplicada en J Express y MEV el resultado fué que el agrupamiento

obtenido por el k-means en estas dos herramientas es similar, es decir, los genes agrupados en el cluster 1 en J Express son los mismos que los genes agrupados en el cluster 1 en MEV. El resultado visual de estos agrupamientos muestra que en el caso de J Express los datos son ordenados de manera ascendente dado su número de identificación y en MEV se encuentran desordenados, pero analizado la representación textual de ambos resultados y tomando como base alguno de los genes se observa que los genes que se encuentran por encima y por debajo de este son similares en ambas herramientas, por lo que el agrupamiento realizado es muy similar.

Al usar la distancia de Pearson en J Express, MEV y Cluster & Tree View el resultado del agrupamiento muestra que los genes agrupados en el cluster 1 en J Express se encuentran agrupados en el cluster 1 en MEV y también en el cluster 1 en Cluster & Tree View, nuevamente los genes agrupados en MEV se encuentran visualmente desordenados y en J Express y Cluster & Tree View se encuentran ordenados, sin ambargo aunque el resultado visual varíe, la representación textual del agrupamiento realizado por cada herramienta para la distancia de Pearson es muy similar, los genes se encuentran ordenados de la misma manera.

Al considerar la distancia de Spearman el resultado de los agrupamientos muestra que los genes agrupados en el cluster 1 por J Express también se encuentran agrupados en el cluster 1 por Cluster & Tree View pero que en el caso de MEV se encuentran agrupados en el cluster 3, a pesar de que existe una diferencia en el número de cluster en el que han sido agrupados los genes esto no tiene que ver con el agrupamiento como tal, la representación textal muestra que los genes agrupados en el cluster 3 por MEV corresponden de manera similar a los genes agrupados en el cluster 1 por J Express y Cluster & Tree View, por lo que el resultado del agrupamiento es parecido; al igual que los resultados obtenidos de la aplicación de las medidas de Manhattan y Pearson el resultado visual suele variar, como se sabe J Express y Cluster & Tree View muestran sus resultados ordenados, a diferencia de MEV que los muestra desordenados, pero esto es sólo la parte visual del agrupamiento de cada cluster.

Mediante el análisis de estos resultados se obtuvo que el agrupamiento realizado por cada una de las distancias evaluado en las diferentes herramientas es muy similar. No importa si el resultado visual es diferente o si los genes se encuentran distribuidos en un número de cluster diferente, la representación textual de cada resultado muestra que los genes se encuentra agrupados de manera muy similar; si se toma como base un sólo gen, los genes que se encuentran por debajo y por encima de este son muy parecidos por lo que afirmamos que el resultado del agrupamiento para una misma distancia evaluada en diferentes herramientas es similar.

6.4. Metodología

Finalmente, después de las pruebas, obtención y análisis de resultados, podemos definir los pasos utilizados para realizar la evaluación de las técnicas de minería de datos aplicadas a datos biológicos.

1. Se evaluaron 4 tipos de bases de datos diferentes: Leucemia, Malaria, Levadura y Linfoma.
2. La representación de los datos se obtuvo a través de Microarrays de genes.
3. Se usaron 5 versiones de Software de Bioinformática: J Express, MEV, Cluster & Tree View, GEPAS y EPCLUST. Todas ellas evaluadas bajo plataforma Windows.
4. Se consideraron 3 técnicas de minería de datos: Agrupamiento jerárquico aglomerativo, K-means y PCA.
5. Se aplicaron diferentes medidas de distancia para evaluar las técnicas: Euclideana, Manhattan, Coeficiente de correlación de Pearson y Coeficiente de Correlación de rangos de Spearman con el propósito de demostrar la importancia de utilizar diferentes medidas de distancia para obtener resultados diferentes.
6. Se hicierón un total de 60 ejecuciones para obtener los resultados. Para una base de datos se evaluaron 3 técnicas diferentes en una herramienta, como se tienen 5 herramientas diferentes

se hicieron un total de 15 ejecuciones, como son 4 bases de datos diferentes, se obtiene el total de las 60 ejecuciones.

7. El objetivo de las ejecuciones fué buscar una consistencia en los resultados obtenidos por cada herramienta, es decir, buscar que el resultado en cada una se mantuviera. La consistencia de los resultados entre herramientas fué de un 90 %. La razón por la que no se obtuvo el 10 % restante para obtener el 100 % de la consistencia se debió a que los parámetros utilizados en las técnicas de minería de datos varía entre cada herramienta y no es posible realizar un cambio en ellos, por lo que no fué posible refinar el método.

Teniendo en cuenta los pasos utilizados en la evaluación de las técnicas de minería de datos, podemos establecer la metodología a seguir para realizar la evaluación de técnicas de minería de datos sobre datos biológicos.

Metodología

Paso 1 Elegir el conjunto de datos sobre el cual se aplicará la evaluación y que aporte información necesaria para obtener resultados.

Se debe considerar la representación de los datos (microarrays u otro tipo) y el tipo de dato (texto o raw data).

De acuerdo al estudio realizado se recomienda seleccionar datos que sean representados a través de estructuras de microarrays y cuyo formato sea tipo texto porque son los que se utilizan con mayor frecuencia y que mejor se adaptan a cualquier herramienta de Bioinformática

Paso 2 Definir la técnica(s) de minería de datos que se utilizará para evaluar los datos.

La elección de la técnica de minería de datos se deja a consideración del usuario, ya que depende del estudio que se realize y del resultado que se espere obtener.

De acuerdo a los resultados obtenidos es este estudio proponemos al K-means como una técnica adecuada para la obtención de resultados computacionalmente más rápidos, precisos y fiables.

Paso 3 Seleccionar la herramienta (Sw) de Bioinformática que mejor se adapte al contexto del problema que se quiere resolver y que proporcione los recursos necesarios para realizar la evaluación.

Estos recursos pueden ser, que la herramienta sea accesible al tipo de dato con el que se dispone y que proporcione la técnica de minería de datos elegida para evaluarlos.

De acuerdo al estudio realizado, se propone la herramienta J Express por ser la más completa, permite utilizar varios tipos de datos ya que es compatible con datos de otros sistemas (GenPix, Affymetrix, etc), ofrece diversas técnicas para evaluar los datos y una variedad de medidas de distancia que permiten obtener resultados diferentes para un mismo conjunto de datos, la representación visual de los resultados es más sencilla y fácil de interpretar y proporciona información textual del resultado de la aplicación de la técnicas.

Paso 4 Después de seleccionar la herramienta que mejor se adecué a las necesidades del usuario, se debe elegir la plataforma sobre la que trabajará la herramienta, esta elección depende completamente de las necesidades del usuario.

Paso 5 Realizar la evaluación de la técnica(s) seleccionada aplicada sobre el conjunto de datos biológicos elegidos.

Paso 6 Analizar e interpretar los resultados obtenidos de acuerdo al problema planteado al inicio.

Capítulo 7

Conclusiones

Con la presentación de las herramientas de Bioinformática descritas en este trabajo se ha comprobado que la Bioinformática es un campo de investigación amplio que tiene diversas aplicaciones que, combinadas con las matemáticas, la biología, la estadística, etc., permiten resolver y entender problemas concretos, principalmente de biología.

Considerando que este campo es relativamente nuevo, se estudiaron varias técnicas de minería de datos para evaluar bases de datos de tipo biológicas. En el estudio de las técnicas existentes aplicadas a la Bioinformática se mencionaron las características, fortalezas y debilidades de cada técnica, así como su actuación dentro de cada herramienta.

A través de la evaluación de las herramientas de Bioinformática hemos visto que es posible obtener resultados visuales diferentes, esto facilita la tarea para un experto en el área de interpretar de mejor manera los resultados e incluso de elegir el tipo de visualización que mejor se acople a sus necesidades.

El análisis realizado con respecto a las medidas de distancia permitió concluir que éstas son un factor importante a considerar en la obtención de resultados porque afectan significativamente el resultado de los agrupamientos, estas medidas comparadas entre ellas mismas en una sola

herramienta, ofrecen resultados muy diferentes de agrupamiento con respecto a la distribución y asignación de genes, pero comparadas entre ellas y aplicadas en distintas herramientas ofrecen resultados similares de agrupamiento.

Al inicio del trabajo no se conocía una metología a seguir para realizar la evaluación de técnicas de minería de datos aplicadas a datos biológicos, el análisis realizado en la obtención de evaluar diferentes bases de datos biólogicas, en herramientas distintas y para cada una de las técnicas, permitió establecer la metodología a seguir para evaluarlas, esta metodología se resume en lo siguiente:

1. Elección del tipo de dato.
2. Elección de la técnica de minería de datos.
3. Elección de la herramienta de Bioinformática y de la plataforma sobre la que trabajará la herramienta.
4. Aplicación de la evaluación.
5. Obtención de resultados para su análisis e interpretación.

Esta metodología permitió tener un panorama más amplio acerca de cuales podrían ser las futuras direcciones y sugerencias para hacer investigación en este campo. Existe una variedad de tópicos no explorados y problemas reales que hacen atractivo este campo de investigación.

Se espera que este trabajo haya aportado los conocimientos básicos necesarios para impulsar la investigación hacia esta área, la idea es despertar la inquietud de aquellos que deseen conocer un poco más de estas aplicaciones y desarrollar con ello en un futuro programas bioinformáticos que satisfagan las enormes necesidades que surgen cada vez con más ímpetu en la bifurcación de las ciencias biológicas y la computación.

Se espera que este trabajo sea la puerta de entrada para seguir indagando y buscando respuestas concretas a tantos problemas biológicos que pueden ser solucionados sólo con el interés y participación conjunta de biólogos y computólogos. El deseo es que este trabajo no sólo sea un documento didáctico más, sino una herramienta de impulso para continuar investigando y para que en el futuro pueda aplicarse en cualquier organización en nuestro país, ya sea de tipo académico, científico o de negocios.

Bibliografía

[1] Conchillo J. Angela, R.Gallego-Largo Trinidad, *Escalamiento Multidimensional: Una metodología de análisis en el campo de los factores humanos*, Facultad de Psicología, Universidad Complutense de Madrid. Agosto 1993.

[2] K.I. Diamantaras, S.Y. Kung, *Principal Component Neural Networks Theory and Applications*, John Wiley & Sons, INC Publication, Capítulo 3, pp 44-73, 1996.

[3] Blattner F.R.; Plunkett III G.; Bloch C.A.; Perna N.T.; Burland V.; Riley M.; Collado-Vides J.; Glasner J.D.; Rode C.K.; Mayhew G.; Gregor J.; Davis N.W.; Kirkpatrick H.; Goeden M.; Mau R.; Shao Y., *The complete genome sequence of Escherichia coli*, Science, volumen 277, pp. 313-319, 1453-1462 , septiembre 5 de 1997.

[4] Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick o. Brown, David Botstein, Bruce Futcher, *Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization*, The American Society for Cell Biology Vol. 9, 15 de octubre de 1998.

[5] Martín Sánchez F, López Campos G, Maojo García V., *Bioinformática y Salud: impactos de la aplicación de las nuevas tecnologías para el tratamiento de la información genética en la investigación biomédica y la práctica clínica*, Informática y Salud 1999

[6] T.R Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, *Molecular Classification of*

Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science Vol. 286, 15 de Octubre 1999.

[7] López Vázquez Victor, *Comparación de los Métodos de Imputación con respecto al poder de Separación del Modelo de Regresión Logística*, Universidad de Puerto Rico, 2000.

[8] R. Guerequeta y A. Vallecillo. *Técnicas de Diseño de Algoritmos*, Universidad de Málaga, Segunda Edición Mayo 2000. ISBN: 84-7496-666-3.

[9] Han, J., Kamber, M., *Data Mining: Concepts and Techniques*, Morgan Kaufman Publishers, Capítulo 7, 2000.

[10] Unidad de Coordinación de Informática Sanitaria (BIOTIC), *¿Qué es la Bioinformática?*, 13 de Noviembre de 2001.
Disponible en http://biotic.isciii.es/informacion/bioinfo/definicion/queesbioinfo.htm.

[11] Pierre Baldi, Soren Brunak, *BIOINFORMATICS, the machine learning approach*, Segunda Edición, pp 84-87, 2001.

[12] Joaquín Dopazo, Alfonso Valencia, *Bioinformática y Genómica*, BIOINFORMATICS Centro Nacional de Investigaciones Oncológicas, 2001.

[13] Consultoría BIOMUNDI, *Estado del arte en Bioinformática*, La Habana: Consultoría BIOMUNDI, 2001.

[14] Vélez Ignacio, *Apuntes de Problemas de Estadística para Ingeniería y Administración*, Facultad de Ingeniería Industrial, Bogotá, Colombia, Octubre, 2002.

[15] Juan Pedro Febles Rodríguez, Abel González Pérez, *Aplicación de la Minería de Datos en la Bioinformática*, ACIMED 02 2002.

[16] Martín Sánchez F, Maojo García V., *La convergencia entre la Bioinformática y la Informática Médica*, (38):25-31, I+S 2002.

[17] Basilevsky A., *A Estatistical Factor Analysis and Related Methods*, Ed. John Wiley and Sons Inc., 2002.

[18] John Quackenbush, *Microarray Data Normalization and Transformation*, volumen 32, pp 496-498, Diciembre 2002.

[19] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP, *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation*, Nucleic Acids Res., 2002 Feb 15;30(4):e15.

[20] S. Draghici, *Data Analysis Tools for DNA Microarrays* , Capitulo 12, 2003.

[21] Smyth, G.K. y Speed, T.P, *Normalization of cDNA microarray data*, 2003.
Disponible en: www.stasci.org/smyth/pubs/mormalize.pdf

[22] A. M. Campbell, L. J. Heyer, *Discovering genomics, proteomics and bioinformatics*, Editorial Benjamin Cummings y Cold Spring Harbor Laboratory Press, 2003.

[23] Zbynek Bozdech, Manuel Llinás, Brian Lee Pulliam, Edith D. Wong, Jingchun Zhu, Joseph L. DeRisi, *The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum*, Plos Biology Vol. 1, 25 de Julio 2003.

[24] Ash A. Alizadeth, Michael B. Eisen, R. Eric Davis, Chi Ma, Izidore s. Lossos, Andreas Rosenwald, Jennifer C. Boldrick, Xin Yu, James Hudson, Dennis D. Weisenburger, John C. Byrd, David Botstein, Patrick o. Brown, Louis M. Staudt, *Distinc Types of Diffuse Large B-cell lymphoma identified by Gene Expression Profiling*, Macmillan Magazines Ltd Vol. 403, 3 de febrero de 2003.

[25] Rubén Cañedo Andalia, Ricardo Arencibia Jorge, *Bioinformática: en busca de los secretos moleculares de la vida*, Red Telemática de Salud en Cuba (Infomed), Diciembre 2004.

[26] Balbi Eduardo, *Modelización Estadística: de las series de tiempo a la simulación*, Buenos Aires, Argentina, ISBN:987-98351-1-5, 2004.

[27] Llanes Mazón Alejandro, *Bases de datos biológicas*, Grupo de informática, Facultad de Biología, Universidad de La Habana, Abril 2005.

[28] Joan Valdivia Alain, Febles Rodríguez Juan Pedro, *Bioinformática: reflexiones y perspectivas*, ACIMED v.12, n.4, Septiembre 2005.

Ligas de interés:

Herramientas

1. J Express, http://www.molmine.com/frameset/fmr_jexpress2.asp
2. Cluster and & Tree View, http://rana.lbl.gov/EisenSoftware.htm
3. MEV, http://www.tm4.org/mev.html
4. GEPAS, http://gepas.bioinfo.cnio.es
5. Expression Profiler, http://ep.ebi.ac.uk/EP

Bases de datos

1. Leukemia, http://www.molmine.com, J Express pro-learning
2. Malaria, http://malaria.ucsf.edu
3. Levadura, http://www.yeastgenome.org
4. Lymphoma, http://rana.lbl.gov/EisenData.htm

Temas varios

1. http://bioinformatica.el.sitio.net, *Portal de Bioinformática.*
2. http://ccc.inaoep.mx/bioinformatica/bio2005, *Taller de Bioinformática y Biología Computacional.*
3. http://tigr.org, *The institute for genomic research.*

4. http://www.lamolina.edu.pe/institutos/ibt/bioinformatica, *Bioinformática: Conceptos generales.*

5. http://www.cgb/cl/application.asp, *Centro de Genómica y Bioinformática.*

6. http://bioinfo.ochoa.fib.es/local/cursos/normalizacion/index.html, *Procesado de Datos de Microarrays.*

7. http://www.elet.pilomi.it/upload/matteucc/Clustering/tutorial_html, *A tutorial on Clustering Algorithms.*

8. http://strix.ciens.ucv.ve/ iartific/Material/tecnicasclasificacion.doc *Técnicas de Clasificación.*

9. http://www.embnet.cl/bio/CLASE-URZUA/microarray-software.pdf, *Microarray Analysis Software Packages.*

10. http://www.gene-chips.com/, *DNA Microarray (Genome Chip).*

11. http://fbio.uh.cu/bioinfo/, *Portal de Bioinformática de la Facultad de Biología, UH.*

12. http://www.nslij-genetics.org/microarray/soft.html, *Public domain programs for Microarray Data Analysis.*

13. http://infobiochip.isciii.es/Textos/GuiaRec/software.htm, *Guía de Recursos de Empresas y Distribuidores de Software* .

14. http://www.statsci.org/micrarra/analysis.html, *Statistical Analysis Software.*

15. http://ihome/cukh.edu.hk/arraysoft_mining_comprehensive.html, *Microarray Data Mining Software.*

16. http://www.es.embnet.org/Services/databases.es.html, *Bases de Datos en EMBnet/CNB.*

17. http://www.geocities.com/plantmolbiol/acidosnucleicos.html, *EBMP: Bases de Datos, Secuencias.*