UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

AGENTES INTELIGENTES PARA EL ANÁLISIS DE EXPRESIÓN DE GENES ASOCIADOS AL CÁNCER DE CUELLO UTERINO

T E S I S

QUE PARA OBTENER EL GRADO DE:

DOCTORA EN CIENCIAS (COMPUTACIÓN)

P R E S E N T A:

MARÍA DEL CARMEN EDNA MÁRQUEZ MÁRQUEZ

DIRECTOR DE TESIS:
DR. JESÚS SAVAGE CARMONA

CD. DE MEXICO **JUNIO 2016**

RESUMEN

El interés despertado por el estudio del genoma de diversas especies y del propio humano a partir de la década de los 90s se ve reflejado en los grandes bancos de información genómica con que se cuenta, tanto de dominio público como privado. Es ahora cuando se requiere el apoyo de las herramientas de la información que permitan extraer la mayor cantidad de conocimiento, que aplicado puede redundar en grandes beneficios para los seres humanos. Uno de los objetivos de las ciencias de la computación es desarrollar la tecnología que nos permita el procesamiento de datos, en este caso biológicos. La Inteligencia Artificial contribuido con importantes logros en este campo.

Este trabajo se encuentra enmarcado en la nueva disciplina que se encarga del procesamiento de información de tipo biológica para extraer nuevos conocimientos con aplicaciones prácticas importantes, la bioinformática. Para las ciencias biológicas, en especial para la medicina y la farmacéutica, es muy importante la identificación de los genes que se expresan en respuesta a diferentes estímulos, pues la expresión de genes, principalmente en grupos, interfiere directamente con el desarrollo de las enfermedades. Los datos correspondientes a miles de genes de un organismo pueden someterse a diversos experimentos de forma conjuntautilizando microarreglos. Lo anterior, como resultado una matriz de miles de datos numéricos cuyo análisis y procesamiento requiere sistemas de computación específicos.

La investigación realizada para este trabajo da como resultado una herramienta inteligente basada en la tecnología de agentes que permiten el procesamiento distribuido de las tareas necesarias para clasificación de muestras e identificación de genes asociados a enfermedades como el cáncer. El caso de estudio tomado corresponde al cáncer de cuello uterino, pero el alcance de este proyecto puede ser cualquier otra enfermedad, e incluso datos no biológicos, debidoa que lacantidad de datos que representa el manejo de microarreglos de expresión de genes pertenece a lo que hoy llamamos BigData.

ABSTRACT

The interest in the study of the genomes of different species of living beings, mainly of the humans, from the decade of the 90s is reflected in the huge repositories of genomic data that are available, public and private domain. It is now when support tools that allow extract knowledge form the great amount of information, its application can result in great benefits to all. One of the goals of computer science is to develop the information technology necessary for processing the biological data banks, and specifically Artificial Intelligence have made important achievements in this field.

This work is framed in the new discipline that deals with the biological information processing type to extract new knowledge with important practical applications, bioinformatics. In life sciences, especially to medical and pharmaceuticals, is very important to identify the genes expressed in certain circumstances, these genes interferes directly with the development of diseases. Data for thousands of genes in an organism can be subjected to various experiments together using microarrays, a matrix of thousands of numerical data which requires nontrivial computer systems to their analysis and processing.

The contribution of this thesis is an intelligent tool based on agent technology enabling distributed processing tasks required for classification of samples and identification of genes associated with diseases, such as cancer. The case study corresponds to cervical cancer, but the scope of this project can be any other disease and even data non biological, if we think that the amount of data in the microarray could belong to what today called Big Data.

Agradecimientos

Agradezco a los miembros de mi comité tutoral, Dr. Jesús Savage Carmona, Dr.Christian Lemaitre León y Jaime Berumen Campos por sus conocimientos y recomendaciones para mi proyecto de investigación.

Gracias al Depto. De Medicina Genómica del Hospital General por haberme recibido y permitido colaborar con ellos.

Gracias al Posgrado en Ciencia e Ingeniería de la Computación y al CONACYT por permitirme llegar hasta aquí.

Gracias a todos los que me brindaron su tiempo, sus comentarios y consejos desde el inicio hasta el final del proyecto, entre ellos las Doctoras Ana Espinosa, Ana Lilia Laureano y Katya Rodríguez.

Gracias a mi esposo e hijos por estar siempre a mi lado y por su cariño.

Dedico esta tesis, especialmente a mi Padre y a mi Madre querida.

Índice general

Índice de fíguras

con listas de genes seleccionados

Índice de tablas

CAPÍTULO 1
Introducción

1.1 Motivación

En la actualidad una de las aplicaciones más relevantes de la tecnología desarrollada por las ciencias de la computación es aquella que se encuentra enmarcada en la bioinformática. La bioinformática es un área donde convergen los conocimientos tecnológicos y las necesidades de interpretación de una gran cantidad de datos biológicos. A partir del siglo pasado, cuando se inició el estudio del genoma de diversas especies, entre ellas del hombre, ha surgido un enorme interés por aprovechar toda esa información biológica. Un ejemplo de lo anterior, es la medicina genómica que con el conocimiento generado por la bioinformática trata de aplicar tratamientos personalizados, dependientes de las características de poblaciones específicas o incluso de una persona, y mejor aún, de contribuir a la prevención de enfermedades.

Es así como una de las tareas de las ciencias de la computación es crear la tecnología de la información necesaria para extraer información de bancos de datos biológico. Específicamente lasmetodologíasaportadas por el área de Inteligencia Artificial ha dado importantes logros en este campo.

Este proyecto se encuentra enmarcado en esta disciplina que utiliza las herramientas computacionales para el procesamiento de información biológica con el fin de extraer nuevos conocimientos con aplicaciones prácticas interesantes, como es el caso de la identificación de los genes que se expresan en determinadas circunstancias, donde su expresión, principalmente en grupos, interfiere directamente con el desarrollo de las enfermedades. Los datos correspondientes a miles de genes de manera conjunta pueden someterse a diversos experimentos mediante el uso de microarreglos, lo que vierte como resultado una matriz de miles de datos numéricos cuyo análisis y procesamiento requiere sistemas de computación no triviales. Los datos que se trabajaron en este proyecto provienen de mujeres con cáncer de

cuello uterino e infectadas del virus de papiloma humano 16 además de controles provenientes de mujeres que no tienen el virus.

1.2 Objetivo general

Desarrollar un sistema que integre en una sola herramienta,basada en agentes, el análisis de expresión de genes para la identificación de genes humanos asociados al proceso de invasión en el cáncer de cuello uterinoy para la clasificación de tipos de cáncer, combinando métodos de filtrado, aprendizaje de máquina junto con el empleo de la información genómica publicada en la web.

1.3 Objetivos específicos

1) Automatizar el análisis de los datos obtenidos de los microarreglos que contienen miles de genes.
2) Diseñar los agentes inteligentes necesarios que operen los módulos de pre-procesamiento de datos, identificación de genes asociados, clasificación de muestras y caracterización precisa de los genes seleccionados.
3) Diseñar el agente que se encargue de la relación con las bases de datos genómicas externas.
4) Implementar métodos estadísticos y de procesamiento de señales para el pre-procesamiento de datos que serán utilizados durante el aprendizaje de máquina.
5) Implementar las herramientas de aprendizaje de máquina que realicen la identificación de los genes y clasificación de muestras.
6) Establecer los mecanismos de comunicación necesarios entre los diferentes agentes del sistema para obtener las soluciones correctas a los problemas de análisis de expresión genética planteados por el usuario.

El sistema multi-agente pretende facilitar la solución a las preguntas: ¿Qué genes son afectados en su expresión debido al cáncer? y ¿Cómo pueden clasificarse los tumores de acuerdo a la expresión de los genes?, ver figura 1.

1.4 Contribución y relevancia

El cáncer de cuello uterino, también denominadocáncer cérvico-uterino(CaCu)es uno de los más comunes en la población femenina. La incidencia del CaCuen México (50 por cada 100,000 habitantes por año) es de las más elevadas en todo el mundo. La alta frecuencia de los virus papiloma humano tipo 16 (HPV16) Asiático-Americanos (AA) detectados en México (23% de todos los cánceres del cuello uterino), pudiera contribuir importantemente a la alta incidencia de este cáncer en México [Berumen 2003]. Varios tipos de HPV son asociados con el 90-100% del CaCuen el mundo y el HPV16 ha sido detectado en casi el 50% de los casos.

El proyecto de investigación médica "GENOMA HUMANO Y CÁNCER DEL CUELLO UTERINO: identificación de genes de susceptibilidad y protección, y marcadores tumorales" propone identificar los genes que se ven afectados por el HPV16 en sus variantes Asiático-Americana (AA) y Europeo (E), y que originan la aparición del cáncer invasivo del cuello uterino. Este proyecto se realiza en el departamento de Medicina Genómica del Hospital General de México en conjunción con la Facultad de Medicina de la UNAM, y está a cargo del Dr. Jaime Berumen Campos. Una de las líneas de investigación es a través del análisis de la expresión genética en biopsias de CaCupositivos para HPV16 en distintos estadios clínicos, representados en lafigura 1, los cuales van desdela neoplasia intraepitelial cervical (NIC) hasta el cáncer *in situ*.

Por lo anterior, esta investigación tiene relevancia para la salud pública. Los resultados obtenidos pueden formar parte de las bases para mejorar la prevención y tratamiento del padecimiento de CaCu en la población femenina. Así mismo, el sistema multi-agente resultado de este proyecto podrá ser aplicado en estudios de cualquier otro tipo de cáncer basado en la tecnología de microarreglos provenientes de Affymetrix®.

En el caso de la clasificación de tumores se obtiene una propuesta de clasificación de variantes del CaCu de acuerdo a la información proveniente de biología molecular a diferencia de la forma tradicional que se sigue con el estudio histológico de las muestras. Mientras que el método de histopatología, que actualmente se sigue, es en cierto grado subjetivo, la

clasificación por medio de la información numérica que se obtiene de los microarreglos pierde esa subjetividad y se permite repetir y aplicar diversas pruebas para decidir la clase a la que pertenece el tumor.

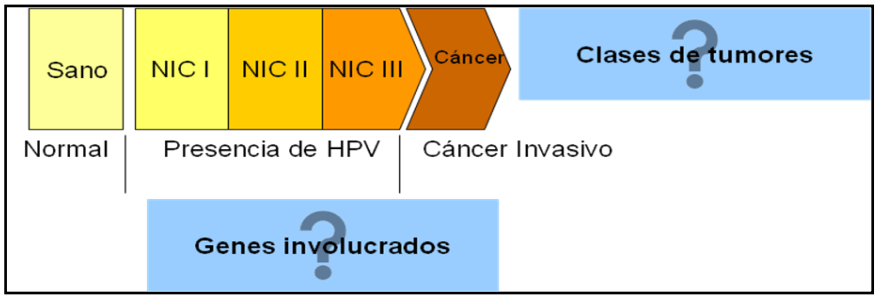


Figura 1. Hay 3 estadios previos al desarrollo de CaCu. Conocer qué genes están involucrados y encontrar una clasificación de tumores basada en los datos de expresión de genes permite a la medicina genómica colaborar en la prevención y tratamiento del cáncer

El análisis de la información biológica obtenida para la expresión de genes que permita la identificación de genes involucrados en el desarrollo de alguna enfermedad requiere del procesamiento de grandes volúmenes de datos generados por los microarreglos de expresión génica, como sucede con nuestro caso de estudio, elCaCu, así como de la utilización de diversas bases de datos biológicos publicadas en internet. Para ello, se requiere del apoyo de herramientas computacionales poderosas como las desarrolladas en el área de inteligencia artificial. El manejo de agentes permite la automatización de este proceso de análisis de expresión genética para obtener conclusiones importantes y con mayor oportunidad al equipo médico que participa en el departamento de Medicina Genómica.

Encontramos la necesidad de facilitar al personal el proceso de identificación de los genes significativos y la posibilidad de la clasificación de muestras, utilizando una herramienta que integre los métodos que aplican actualmente por separado utilizando diferentes aplicaciones especializadas, debe considerarse que el personal encargado de Medicina Genómica cuenta con formación en el área biológica. A partir de ese análisis de la situación pensamos en la propuesta de un sistema con diversos agentes computacionales, cada uno de ellos especializado en una etapa del proceso, logrando la automatización y apoyo en la toma de decisiones. La automatización está centrada en el usuario, los resultados emitidos por el sistema multi-agente son propuestas de solución y toca al usuario (especialista en biología molecular o médico) su interpretación y validación biológica. Sin embargo, debido a que los métodos desarrollados están basados en la experticia y

conocimiento del conjunto de procedimientos que ellos realizan se espera que el resultado esté cerca de los estándares que ellos manejan.

El sistema aquí desarrollado es una aportación en el campo de la bioinformática ya que el objetivo principal no sólo hacer consultas a la información disponible en la Web o hacer únicamente el análisis de los datos numéricos, sino que conjuntamos en un solo sistema ambas metas. Esas tareas que los bioinformáticos realizan utilizando diferentes aplicaciones de software son implementadas en los agentes para que en un sistema de software sea posible discriminar los genes más significativos y crear agrupaciones de tipos de muestras tumorales basadas tanto en listas de genes seleccionados como con el grueso de todos los genes, facilitando así esa labor al personal no especializado y apoyando en la toma de decisiones.

En esta tesis tomamos como caso de estudio el CaCu pero el sistema desarrollado puede ser utilizado para el análisis de expresión de genes de otros padecimientos, pues la metodología es similar.

1.5 Organización de la tesis

El presente trabajo está estructurado en 7 capítulos:

Capítulo 1. En este capítulo hemos dado una introducción al tema de aplicación, el cáncer de cuello uterino, además de incluir los objetivos y la justificación del presente trabajo.

Capítulo 2. Presenta diversos conceptos que integran elconocimiento biológico básico para comprender el dominio del problema. Incluye una introducción a la tecnología de los microarreglos de expresión de genes, la cual es la fuente de los datos numéricos que alimentan el sistema multi-agente, además de algunos aspectos del CaCu.

Capítulo 3. Este capítulo muestra los principales métodos de aprendizaje de máquina utilizados para el análisis de expresión de genes.

Capítulo 4.Proporciona el estado del arte correspondiente a los agentes y sistemas multi-agentes, así como su relación con el tratamiento de datos de bioinformática.

Capítulo 5. Presenta la propuesta de solución basada en el diseño e implementación de un sistema multi-agente creado como herramienta de apoyo en el análisis de expresión de genes.

Capítulo 6. Se incluyen algunas de las pruebas realizadas con datos reales de muestras de casos y controles de CaCu y los resultados obtenidos.

Capítulo 7. Se dan las conclusiones del trabajo desarrollado y trabajos futuros.

Al final de la tesis se incluyen los apéndices A y B. El apéndice A contieneun pequeño glosario con términos biológicos y computacionales utilizados en la tesis. partes de la representación del conocimiento que manejan los agentes y pantallas del sistema multi-agente implementado. El apéndice B incluye el artículo publicado como resultado de esta investigación.

CAPÍTULO 2

Dominio biológico

En la última década del siglo pasado surgió una disciplina nueva, la bioinformática, dedica a la investigación y desarrollo de herramientas tecnológicas útiles para entender el flujo de información proveniente del genoma de cualquier especie. Con la finalidad de conocer a través desde los genes hasta sus estructuras moleculares, su función bioquímica, su conducta biológica y finalmente, su influencia en las enfermedades y en la salud.

La bioinformática utiliza datos que pueden provenir de la expresión de los genes (o expresión genética), los cuales pueden llegar a ser miles en una sola observación. Una de las formas de medir la expresión genética es la tecnología de los microarreglos.

2.1 El dogma central de biología molecular

El ácido desoxirribonucleico (ADN), forma parte de cualquier tipo de vida en nuestro planeta. El ADN contiene la información genética y se puede encontrar en las células de todos los organismos en la Tierra. Los principios básicos de la replicación y la traducción de la información genética, que integran el dogma central de biología molecular son similares en cada organismo.

El ADN está constituido por bloques denominados nucleótidos, los cuales contienen moléculas de azúcar y de fosfato además de una de las 4 bases nitrogenadas: adenina, guanina, citosina y tiamina, conocidas por las letras A, G, C y T, respectivamente.

El dogma central de la biología molecular define tres procesos principales en la utilización celular de la información genética. El primero es la replicación (o duplicación), la copia de un ADN progenitor para formar una molécula de ADN hija que tiene una secuencia de nucleótidos idéntica a la original. El segundo proceso es la transcripción, el proceso por el cual partes de el mensaje codificado en el ADN es copiado en forma deácido ribonucleico(ARN). El

tercer proceso es la traducción, en el cual el mensaje genético codificado en el ácido ribonucleico mensajero (ARNm) es traducido en los ribosomas en una proteína con una secuencia específica de aminoácidos.

El dogma también postula que sólo el ADN puede duplicarse y, por tanto, reproducirse y transmitir la información genética a la descendencia. Los segmentos de ADN que llevan esta información genética son llamados genes y pueden verse como secuencias de bases, (ejem. *ATGCTAGATCGC)*, otras secuencias de ADN tienen propósitos estructurales o toman parte en la regulación del uso de esta información genética. Cada gen contiene una parte que se transcribe a ácido ribonucleico mensajero (ARNm) y otra que se encarga de definir cuándo y dónde deben expresarse. El ARNm resultante, utilizando el código genético, se traduce como la secuencia de aminoácidos, que integran las proteínas, (figura 2).

La información contenida en los genes (genética) se emplea para generar ARN y proteínas, que son los componentes básicos de las células.

Figura 2. El doga central de biología molecular, El ADN puede transcribirse en ARN y éste traducirse a proteína. Además puede darse una transcripción que va de ARN para generar ADN, o copiarse a más ARN.

2.2 Microarreglos que detectan cambios en la expresión génica

Los microarreglos son una técnica de biología molecular que apareció en los años 90s y que muestra la expresión de miles de genes en una matriz.

Desde su aparición la tecnología de microarreglos ha sido una herramienta esencial en el descubrimiento de la información genética. Específicamente, los microarreglos de expresión genética tienen como función presentar el comportamiento de miles de genes bajo condiciones específicas. A partir de ahí, se abre un abanico de posibilidades dentro de la epidemiología, la farmacología o cualquier rama de la medicina en genera.

Un tipo de sondas que se puede utilizar para el diseño de microarreglos son de transcriptoma, otros pueden detectar la pérdida o ganancia de genes y otras mutaciones se pueden detectar en el ADN. La diferencia entre cada uno de ellos es el tipo de ADN que está inmovilizado en las placas de los microarreglos.

Cuando se quiere determinar un cambio en el nivel de expresión de un gen, esto puede ser detectado por análisis de expresión de microarreglos llamada chips o biochips. El ADN para ser estudiado es inmovilizado por hibridación de ácido ribonucléico mensajero (ARNm), un producto génico conocido por ácido desoxhirribonucléico comunicante (ADNc). Esto viene de las células sanas del tejido (control) y de pacientes que padecen una enfermedad (muestras de estudio). Si un gen es expresado en más de un experimento, una mayor cantidad de ADNc se hibrida en un punto (spot) que representa el gen afectado y por lo tanto las intensidades de fluorescencia son diferentes entre el grupo de estudio y el grupo control. Una vez que se caracterizan los genes implicados en ciertas enfermedades, el cADN de las células humanas pueden ser hibridadas para determinar si la persona tiene el patrón de expresión de genes relacionados con la enfermedad y para optimizar el diagnóstico y el tratamiento.

Los chips de expresión de genes también se pueden utilizar para determinar los cambios en la expresión a través del tiempo, tales como durante el ciclo celular. Esto representa una herramienta importante en la investigación del cáncer que podrían identificar nuevos marcadores de cáncer con fines de diagnóstico.

2.3 Análisis de datos de microarreglos

La idea de identificar grupos o particiones de genes es que los genes que tienen patrones de expresión contienen los mismos mecanismos de regulación y su conocimiento es la base para un control preciso de su transcripción.

El análisis de microarreglos es un método de comparación de los microarreglos que proporciona información cuantitativa sobre el perfil de la transcripción completa de las células, lo que podría facilitar el desarrollo de medicamentos y terapéutica, el diagnóstico de la enfermedad, y una mejor comprensión de la biología celular básica. Uno de los retos en el análisis de microarreglos, especialmente en perfiles de expresión génica de células cancerosas, es identificar los genes o grupos de genes que son altamente expresados en las células tumorales, pero no en las células normales y viceversa.

Los experimentos hechos con microarreglos permiten determinar la abundancia de transcripción de los genes de un organismo en condiciones diferentes, como sucede en las muestras de tejido sano y tejido tumoral. Con el análisis de datos de expresión génica se intenta identificar grupos de genes que exhiben un comportamiento similar bajociertas condiciones de los experimentos de microarreglos.El uso principal de un experimento de microarreglos es determinar si ha habido un cambio en el nivel de expresión de un gene cuando hay una condición frecuente, como sucede en un tumor canceroso. El nivel de expresión puede cambiar hacia arriba, en ese caso se dice que el gene está sobre-regulado o el nivel de expresión puede cambiar hacia abajo y en ese caso se dice que el gene está sub-regulado. El objetivo es identificar los genes que han cambiado significativamente su expresión cuando está la condición presente. Estos genes forman una lista conocida como “genelist”.

Este proceso de comparación requiere el pre-procesamiento de los datos que permite ajustar los niveles de expresión entre los diferentes microarreglos. A partir de la lectura de los microarreglos desde los archivos que contienen las intensidades de expresión de los genes en diferentes muestras obtenidas, se crea la matriz numérica con la expresión de genes, el pre-procesamiento continúa con la normalización y estandarización de los datos numéricos.

Existen muchos métodos para el análisis los datos de expresión de genes. Estos métodos se pueden agrupar principalmente en métodos de filtrado y métodos de clasificación supervisados y no supervisados. Los métodos de filtrado son de tipo estadístico y tratan generalmente de comprobar una hipótesis. El filtrado más común consiste en la búsqueda de los genes que presenten los niveles de transcripciónpor encima o por debajo de un cierto

umbral para una condición dada. Estos genes denominados bajo o sobre regulados son importantes porque por su condición especial en la expresión pueden modificar el comportamiento de un organismo debido a que alteran la generación normal de proteínas. A este método se le conoce como análisis diferencial de la expresión ya que comparala expresión delos genes en una condición determinada contra su expresión en una condición de control, la definición de umbrales de expresión está basa en la proporción, media y varianza.

Posterior al filtrado de los genes significativos por ser bajo o sobre regulados, es decir, que son diferenciados significativamente en su expresión se da paso al agrupamiento de las muestras y/o de los genes.Los métodos de agrupamiento intentan caracterizar laestructura generalde la matriz de expresión, presentandoconjuntos separados de genescon comportamientos similares, los métodos pueden ser supervisados o no supervisados.

La agrupación o clasificación supervisada utiliza el conocimiento disponible también llamado conocimiento a priori del problema. La clasificación supervisada está basada en la predicción de clases, análisis discriminante y aprendizaje supervisado. Mientras que laagrupación no supervisada se realiza por predicción automáticay aprendizaje no supervisado. Entre los métodos no supervisados más utilizados para la agrupación de genes con perfiles similares en todas las situaciones, están los métodos de clustering, pertenecientes al aprendizaje de máquina, de los cuales haremos una amplia descripción en el capítulo 3.

Para el presente trabajo se utilizaron microarreglos de expresión de genes denominados Human Gene Focus, de la tecnología Affymetrix. Este tipo de microarreglos contienen ~8638 genes humanos caracterizados en la base de datos Gene Reference.

Los valores de los microarreglos son almacenados en archivos .CEL que son leídos directamente para iniciar el pre-procesamiento de los valores numéricos que se almacenan en la matriz de datos creada (figura 3).

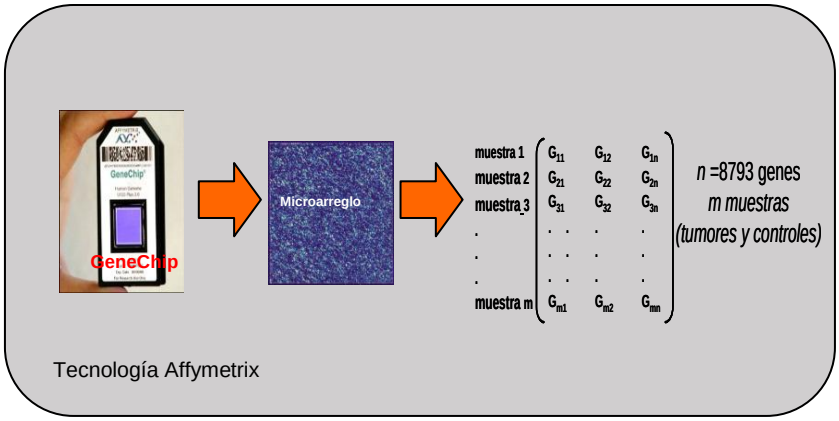


Figura 3. Los microarreglos de Affymetrix representan la imagen de la expresión de miles de genes en un experimento, eso puede traducirse a una matriz numérica para su procesamiento.

2.4 La normalización y estandarización de los datos

La señal obtenida a partir de microarreglos debe ser estandarizada o normalizada debido a la diferencia de expresión en los genes entre las muestras con el objetivo de que los datos de diferentes microarreglos se puedan comparar de forma fiable. Por otro lado, la tecnología de microarreglos y la preparación experimental puedenintroducir ciertos artefactos en la medida de la expresión genética y durante el pre-procesamiento se trata de eliminar estas variaciones sistemáticas en los datos tratando siempre de preservar la variación biológica real. Al ajustar la señal se pueden eliminar errores causados por factores técnicos y biológicos.Con la normalización aplicada a los datos de expresión se ajustan las intensidades de hibridación individuales de genes en los microarreglos.

De los algoritmos más reportados para la normalización de microarreglos de ADN son Microarray Suite (MAS) y Robust Multiarray Average (RMA) [Irizarry 2003], las que eliminan la variación de las intensidades totales de los microarreglos.Con la normalización MAS cada microarreglo se trata de manera independiente, mientras que el algoritmo RMA trabaja con todos los microarreglos en conjunto y no utiliza uno de ellos como base de la normalización, sino que todos los chips son tratados como iguales. El método de normalización asume que las intensidades de los chips siguen una distribución normal y los valores están expresados en log2, esta normalización es la más utilizada.

Finalmente, la fase de preparación de los datos de microarreglos termina con la estandarización de los datos a través de la transformación de la expresión de genes a nivel de un rango igual en todos los chips que permita un espectro continuo de valores, generalmente la transformación log2 se utiliza con los métodos de la agrupación.

2.5 El cáncer de cuello uterino

El cáncer de cuello uterino (CaCu) es la segunda causa de mortalidad en población femenina a nivel mundial incluyendo a México. Es una enfermedad multifactorial donde los factores ambientales, como el tabaquismo, el uso de anticonceptivos orales y algunas deficiencias dietéticas, así como factores genéticos confieren susceptibilidad o resistencia al desarrollo de la enfermedad. El virus del papiloma humano (HPV) esta asociado al desarrollo de este cáncer,ya que se han detectado hasta en el 99.7% de carcinomas cervicales [Berumen2010]. Esto indica que el virus es indispensable en el proceso tumoral.

Existen diferentes virus del papiloma humano, algunos infectan el tracto genital y otros producen verrugas benignas y otras lesiones en la piel y mucosas no genitales. Los HPVs asociados a lesiones genitales pueden ser de alto y bajo riesgo. Se han identificado 13 HPVs de alto riesgo asociados al CaCu y neoplasias intraepiteliales cervicales de alto grado (NIC-AG), los más comunes son los HPVs 16, 18, 31, 33, 35, 39, 45, 52, 56, 58 y 59. Los virus de bajo riesgo, como los HPVs 6, 11, 40, 42, 43, 44, 54, 61, 72, 81, están asociados a neoplasias intraepiteliales cervicales de bajo grado (NIC-BG) e infecciones asintomáticas.

El HPV presenta variantes virales que se comportan de manera distinta en las regiones geográficas. Las variantes del HPV16 tienen una distribución distinta entre los cinco continentes y de allí toman su nombre: las variantes Asiático-Americanas (AA) se encuentran principalmente en México, Centro y Sudamérica y en España; las variantes Africanas (Af) en África, las variantes Asiáticas (As) en el Sudeste de Asia y las variantes Europeas (E) en todas las regiones excepto en África. Las variantes Europeas son las más comunes y las de menor riesgo para desarrollar CaCu.

El CaCu es una enfermedad que se desarrolla por etapas. La clasificación de estas etapas está basada en los cambios morfológicos del epitelio que define el grado de la lesión. La clasificación más utilizada es la neoplasia intraepitelial cervical (NIC) que se divide en grados I, II y III y finalmente cáncer *in situ*. Los criterios para el diagnóstico de neoplasia intraepitelial varían según el criterio del médico-patólogo; aunque los aspectos más importantes a considerar son la desorganización celular, la atipia nuclear y el aumento de la actividad mitótica. La NIC I corresponde a una lesión de bajo grado, ligera displasia, coilositosis y condiloma. La NIC II es una lesión de alto grado y la NIC III es una displasia moderada o severa.

A partir del diagnóstico de cáncer in situ se hace una clasificación de la progresión clínica del CaCu, utilizando los criterios de la International Federation of Ginecology and Obstetrics, conocida por sus siglas como FIGO. Esos criterios tipifican al CaCu en 4 estadios con varios subestadios cada uno, y el pronóstico de la enfermedad varía ampliamente en ellos.

Por la zona en que se desarrolla el CaCu se divide en 3 tipos histológicos : 1) Carcinoma de células escamosas o Epidermoide es el más frecuente (90-95%), se origina en el (exocervix); 2) Adenocarcinoma es menos frecuente, aunque recientes reportes indican el aumento de este tipo cáncer cervical y se origina del endocervix y, 3) Otros tumores epiteliales.

Se han hecho estudios que sugieren que el pronóstico de vida y supervivencia está relacionado con el tipo histológico del tumor. El adenocarcinoma es el de peor pronóstico, y se reporta que tiene un comportamiento más agresivo y de menor supervivencia de la mujer [Platz 1995].

Existen alrededor de 200 tipos diferentes de HPV pero solo 40 son conocidos como HPV genitales y se dividen en dos grupos: los de bajo riesgo (HPV-BR: 6,11,40,42,43,44,54,61,70,72,81) y los de alto riesgo oncogénicos (HPV-AR: 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73, 82). Los tipos virales más predominantes en la región de América latina son el tipo 16 y el tipo 18.

En la actualidad, existe la necesidad de desarrollar nuevos métodos de detección temprana para el CaCu con alta sensibilidad y especificidad al mismo tiempo. Hay un grupo de genes expresados solo en muestras tumorales positivas a HPV y no en tejido sano que pueden ser prometedores como blancos para detección temprana por nuevos métodos sensibles y específicos, así como para la terapia y tratamiento de la enfermedad.

El objetivo de esta tesis es crear un sistema que automatice el proceso de análisis de expresión de genes y contribuir en la identificación de los genes alterados en el CaCu y que sean útiles como blancos de diagnóstico o blancos terapéuticos, así como a una mejor clasificación de muestras como ayuda en el pronóstico terapéutico.

Para la elaboración de esta tesis se trabajo con la Unidad de Medicina Genómica del Hospital General de México. El personal que allí labora nos proporcionó información referente a 55 muestras, el análisis de la expresión de genes se realizó con el ARN aislado de 43 biopsias de tumores y de 12 muestras de epitelio cervical normal. Los perfiles de expresión génica de 43 muestras de CaCu positivos para HPV16 y 12 controles sanos fueron examinados usando el microarreglo de oligonucleótidos de expresión Human Gene Focus (HG Focus, Affymetrix, Santa Clara, CA). Ese arreglo contiene ~8794 sondas que corresponden con 8638 genes humanos caracterizados en la base de datos Gene Reference.

CAPÍTULO 3

Aprendizaje de máquina

El aprendizaje de máquina(Machine Learning) es la rama de la Inteligencia Artificial que tiene como objetivo desarrollar técnicas que permitan a las computadoras aprender. Se ocupa de crear algoritmos capaces de generalizar comportamientos y reconocer patrones a partir de una información suministrada en forma de ejemplos.

Podemos verlo como un proceso de inducción del conocimiento, es decir, un método que permite obtener por generalización nuevo conocimiento. Es descubrir nuevas formas de resolver problemas.

3.1 Aprendizaje de máquina en bioinformática

El aprendizaje de máquina es un área de la computación con una importante aplicación en bioinformática, es una herramienta de automatización para reducir la complejidad de la gran cantidad de datos biológicos y para descubrir patrones y relaciones entre los datos, un ejemplo es la posibilidad de realizar el descubrimiento de conocimiento de biología molecular para la identificación de genes.

Entre las principales razones para aplicar el aprendizaje de máquina en la bioinformática están [Bergeron 2002]:

- Los nuevos enfoques experimentales, basados en biochips (microarreglos), que permiten obtener datos genéticos en gran cantidad, de genomas individuales (mutaciones, polimorfismos) o de enfoques celulares (expresión génica).
- El enorme volumen de datos generados por los distintos proyectos denominados genoma (humano y de otros organismos).
- El acceso universal a las bases de datos de información biológica.

Para dar solución a algunas de estas metas expuestas, existen algunas herramientas desarrolladas y que basan en aprendizaje de máquina:

a) *WEKA (*Waikato Environment for Knowledge Analysis*)* [Hall2009]

Herramienta de entorno para análisis del conocimiento de la Universidad de Waikato, NZ. Es un software libre para aprendizaje automático y minería de datos escrito en Java.

b) GEPAS (gene expression analysis pattern suite) [Tarraga 2008]

Esta herramienta se encuentra en la web para el análisis de microarreglos, en especial para el análisis de expresión de genes. Cuenta con diferentes algoritmos de clustering, de pre-procesamiento y diferenciación de genes.

c) dChip [Li 2003]

Es un software que permite utilizar algunos algoritmos de clustering implementados para el análisis de expresión de genes. Presenta resultados en formato gráfico y en tablas de excel. Maneja muchos parámetros en las ventanas de diálogo para ejecutar acciones.

En las herramientas anteriores se ofrecen diversos algoritmos de aprendizaje de máquinay filtrado ya implementados que pueden ser utilizados para análisis de expresión de genes, sin embargo, no es sencillo su uso por personal neófito de computación. En el caso de Weka, uno de los principales problemas es la escaza documentación que se ofrece del software. Con Gepas una desventaja es que los datos son leídos en los archivos del usuario y enviados para su análisis al servidor web que se encuentra en España, lo que hace muy lento el análisis y poco seguro para la obtención de resultados. Además el usuario depende completamente de la disponibilidad del software y no tiene la flexibilidad de adaptarlo a sus necesidades actuales. El software Dchip, fue desarrollado hace más de 20 años y muchas de sus herramientas ya están obsoletas, no tiene mantenimiento además la interfaz no es fácil de comprender para el usuario común.Considerando los inconvenientes anteriores para el uso de las herramientas de agrupamiento incluimos en el sistema tres agentes autónomos que se encargan de aplicar esos algoritmos de pre-procesamiento y agrupación haciendo transparente para el usuario la aplicación de los métodos y permitiendo incorporar más métodos.

3.2 Métodos de clustering y el análisis de expresión de genes

Clustering corresponde a un conjunto de métodos de aprendizaje de máquina que permite encontrar grupos de datos que tienen características comunes o similares en ciertas condiciones.El objetivo de aplicar clustering es descubrir particiones o grupos de objetos similares, a los que se denomina clusters, actualmente una de las aplicaciones principales del clustering es el análisis de expresión genética.

Como ya se mencionó en el capítulo 2, existe un gran interés en la comunidad bioinformática para encontrar agrupaciones de genes significativos, los grupos formados pueden representar relaciones funcionales, relaciones metabólicas, genes involucrados en ciertos procesos biológicos o que responden a ciertas condiciones del ambiente. Esto puede ser interpretado de diversas formas y alcanzar gran relevancia para la salud. Se han aplicado diversos métodos de clustering no supervisado propuestos para el procesamiento de señales.Una vez aplicados los métodos de clustering es necesario responderse algunas preguntas como: ¿Cuál es la mejor partición de los datos?, ¿Cuál es el número correcto de clusters? O ¿Cuál es el mejor método de clustering?

Los métodos de clustering aplicados en la identificación de genes permiten, por medio de la agrupación de genes, encontrar patrones de expresión para identificar los genes involucrados en enfermedades tales como el cáncer, con lo que es posible llegar a los genes marcadores que están estrechamente relacionados con las enfermedades y que se pueden usar como blanco para el tratamiento y prevención de enfermedades.

Entre los métodos de clustering más utilizados para estas tareas se encuentran [Guojun 2007]:

- k-medias
- c-fuzzy
- clustering jerárquico
- cuantización vectorial

Estos métodos representan una técnica analítica de tipo cuantitativo para la identificación de grupos o conjuntos de puntos en un espacio dimensional dado. Esta agrupación se hace sobre la base de similitudes o diferencias de distancias. Para cada uno de los métodos se debe elegir una medida de similitud o distancia, y esa medida varía de acuerdo con el objetivo del estudio.

La idea principal del cálculo de distancias o similitudes considera que los elementos similares o cercanos deben comportarse de forma similar.La medida de distancia más empleada es la distancia eucludiana, se basa en la obtención de la separación entre dos variables *(X,Y)*, representadas por una serie de puntos en el espacio euclidiano. Otra medida de distancia es el coeficiente de correlación de Pearson, medida que se basa en el grado de relación que existe entre los elementos, verifica si las dos variables varían juntas, entonces se dice que están correlacionadas.

En el caso que nos ocupa, con vectores de expresión de genes, se trata de un espacio n-dimensional, $X=\{x_1,x_2,x_3,\ldots,x_n\}$ e $Y=\{y_1,y_2,y_3,\ldots,y_n\}$, la distancia euclidiana se expresa en la ecuación 1 y el coeficiente de correlación de Pearson en la ecuación 2.

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

Con las medias $\bar{x}$ e $\bar{y}$, y desviaciones típicas σ_x, σ_y:

$$r=\frac{\sum_{i=1}^{n}[(x_i-\bar{x})(y_i-\bar{y})]}{(n-1)\sigma_x\sigma_y} \qquad (2)$$

En el análisis de expresión de genes los datos corresponden a una matriz de genes con su expresión en las diferentes muestras, tumores o controles.

$$\begin{pmatrix} G_{11} & G_{12} & G_{13}\ldots\ldots G_{1n} \\ G_{21} & G_{22} & G_{23}\ldots\ldots G_{2n} \\ G_{31} & G_{32} & G_{33}\ldots\ldots G_{3n} \\ . & . . & . \\ . & . . & . \\ . & . . & . \\ G_{m1} & G_{m2} & G_{m3}\ldots\ldots G_{Nn} \end{pmatrix}$$

N=1…n,muestras

M=1…m, genes

3.2.1 K-medias

El algoritmo de las k-medias es un algoritmo de partición. Básicamente este algoritmo busca formar clusters (grupos) los cuales serán representados por *k* objetos denominados centroides, cada uno de estos *k* objetos es el valor medio de los objetos que pertenecen a dicho grupo.

Es uno de los métodos más conocidos de clustering y por lo tanto muy utilizado para análisis de expresión de genes [Irizarry 2003]. En la mayoría de las herramientas de software desarrollados para esa tarea está implementado este algoritmo. Su utilidad principal radica en cuanto a que puede dar una primera aproximación de las agrupaciones que se pueden formar y el algoritmo converge rápidamente. Sin embargo, la elección del número *k* de grupos que se formarán con los datos requiere conocimiento a priori, es decir, se debe tener mayor conocimiento del comportamiento de los datos y hacer diversas pruebas para encontrar el mejor valor. Los datos seleccionados al inicio para formar los primeros *k*centroides influye en la formación final de los clusters, por lo que varían generalmente los resultados finales.

Algoritmo k-medias

Datos de entrada: Conjunto de *N* vectores de datos, *k* número de clusters

1. Seleccionar *k* vectores de datos que representan las particiones del conjunto inicial, $\{C_1, C_2, C_3, \ldots, C_k\}$
2. Repeat
3. for i=1 to N do
4. for j=1 to k do
5. $d_i(x_i,C_j)$
6. if dMin>$d_i(x_i,C_j)$ then
7. dMin=$d_i(x_i,C_j)$
8. p=j
9. End if
10. End for
11. asignar i al cluster p
12. End for
13. for j=1 to k do
14. C_j=media de los datos que integran el cluster
15. End for
16. until no hay cambios en asignación de datos a los clusters

Tratando de resolver el inconveniente de la variación de resultados finales, se ha creado una versión de este algoritmo que se basa en lógica difusa por lo que recibe el nombre de fuzzy c-medias.

3.2.2 Fuzzy c-medias (FCM)

Este algoritmo es una versión del algoritmo k-medias que implementa lógica difusa. En muchas situaciones cotidianas ocurre el caso de que un dato está lo suficientemente cerca de dos clusters de tal manera que es difícil etiquetarlo en uno o en otro, esto se debe a la relativa frecuencia con la que un dato particular presenta características pertenecientes a clusters distintos y como consecuencia no hay exactitud en su clasificación.

El algoritmo FCM asigna a cada dato un valor de pertenencia a cada cluster y por consiguiente un dato específico puede pertenecer parcialmente a más de un cluster. A diferencia del algoritmo k-means, FCM realiza una partición suave del conjunto de datos, en tal partición los datos pertenecen en algún grado a varios de los clusters.

Algoritmo fuzzy c-medias

Datos de entrada: Conjunto de *N*vectores de datos, *c* número de clusters, *g* grado de fusificación

1. Seleccionar *k* datos que representan las primeras particiones del conjunto inicial, $\{v_1, v_2, v_3, \ldots, v_k\}$
2. Repeat
3. for i=1 to N do
4. for j=1 to k do
5. Obtener grado de membresía de i, μ_i, a v_j
6. if dMin>$d_i(x_i,v_j)$ then
7. dMin=$d_i(x_i,v_j)$
8. p=j
9. End if
10. End for
11. Asignar *i* al cluster *p*
12. End for
13. for j=1 to k do
14. $v_j=\mu_j(x_i)$
15. until no hay cambios en asignación de datos a los clusters

$v_j=\mu_j(x_i)$, corresponde a la ecuación 3.

$$\mu_j(x_i) = \frac{\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{p}\left(\frac{1}{d_{ki}}\right)^{\frac{1}{m-1}}} \qquad (3)$$

Donde, $\mu_j(x_i)$ es el grado de pertenencia del objeto x_i al cluster C_j

d_{ij}es la distancia del objeto x_i al cluster C_j

m es el parámetro de fusificación

p es el número de clusters especificado

d_{ik}es la distancia del objeto x_i en el cluster C_k

Los centroides calculados para los nuevos clusters se obtienen por:

$$c_j = \frac{\sum_i [\mu_j(x_i)]^m x_i}{\sum_i [\mu_j(x_i)]^m} \qquad (4)$$

La restricción del grado de pertenencia de un objeto $x_i$es:

$$\sum_{j=1}^{p} \mu_j\ (x_i) = 1 \qquad (5)$$

En el algoritmo se debe calcular el grado de pertenencia de cada elemento a cada uno de los clusters creados, en la figura4, se aprecia la representación de un elemento que pertenece a 2 grupos a la vez.

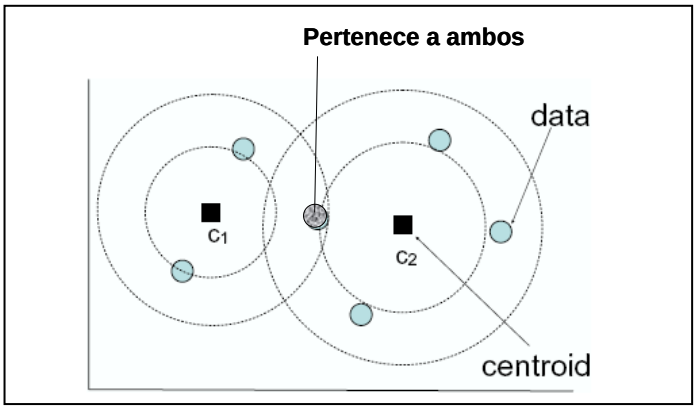


Figura 4. Representación de clusters difusos

Este método resulta muy útil para resolver la clasificación de muestras correspondientes a los diferentes tumores, debido a que se ha encontrado que con los datos de biología molecular, es decir, los provenientes de microarreglos de expresión de genes no siempre corresponde a los grupos formados con la clasificación dada por las pruebas de histopatología, las cuales son hechas principalmente de manera visual por los patólogos quienes aplican lógica bivalente, es decir, pertenece o no pertenece a un grupo de tumores, sin considerar grados de pertenencia. Utilizando métodos difusos como c-means, se permite distinguir el grado de pertenencia a los diferentes grupos y basados en una cuantización objetiva de los datos. Es importante recordar que una clasificación adecuada de las muestras tumorales es muy útil para el tratamiento y pronóstico de vida o evolución del paciente.

Su desventaja es la dificultad al seleccionar el valor de difusividad de los clusters, m, generalmente se le asigna el valor 2. De la asignación de ese parámetro depende la pertenencia a cada grupo y en algunos casos el método falla dando valores de membrecía muy similares para los elementos en diferentes clusters.

3.2.3 Clustering jerárquico

Es uno de los métodos de agrupación de manera aglomerativa, donde los elementos que se encuentran separados forman grupos y esos grupos se modifican durante el proceso, hasta formar un árbol jerárquico con los grupos de genes [Quackenbush 2001]. En general los resultados pueden ser vistos como árboles binarios.

Algoritmo de Clustering jerárquico:

Datos de entrada:Conjunto de *N*vectores de datos

1. Crear N clusters,con 1 objeto cada uno
2. While N>1 do
3. For i=1 to N do
4. Calcular d(C_i, C_{i+1})
5. End_For
6. Encontrar los clusters más cercanos (C_i, C_j),
7. C_{ij}= merge(C_i, C_j
8. N=N-1

9. End While.

Para calcular la distancia entre dos clusters se ocupa alguna de las siguientes medidas de distancia:

• $d(C_i,C_j) = \min d(x, y)$. Single-linkage. Se considera que la distancia o similitud entre dos clusters viene dada por la mínima distancia entre sus componentes.

• $d(C_i,C_j) = \max d(x, y)$. Complete-linkage. La distancia o similitud entre dos clusters se mide a partir de sus elementos más dispares, es decir, la distancia o similitud entre clusters viene dada, respectivamente, por la máxima distancia entre sus componentes.

• $d(C_i,C_j)$ = average $d(x, y)$. la distancia entre los dos clusters se obtiene como la media aritmética entre la distancia, o similitud, de las componentes de dichos clusters.

• $d(C_i,C_j)$ = centroide $d(x, y)$.La semejanza entre dos clusters se obtiene a partir de la semejanza entre sus centroides. Donde sus centroides son los vectores de medias de los individuos del cluster, y son m_i y m_j dimensionales.

En la figura 5 puede verse la representación gráfica que resulta en el clustering jerárquico con la aplicación de las diferentes distancias. Y en la figura 6 puede verse una representación gráfica que resulta en el clustering jerárquico para la expresión de genes.

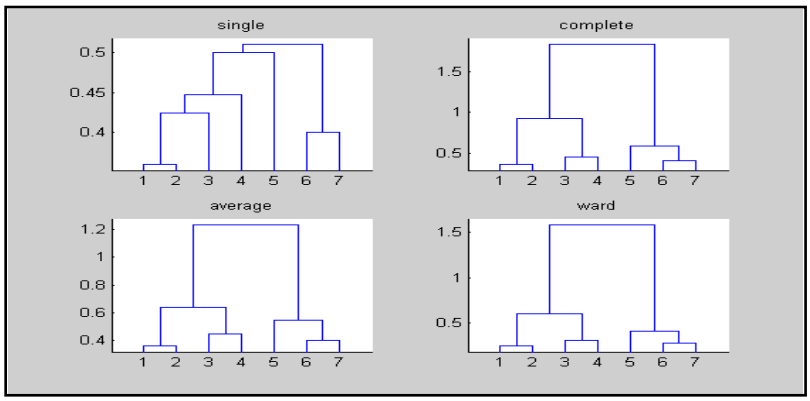


Figura 5. Resultados diferentes en dendogramas creados por clustering jerárquico

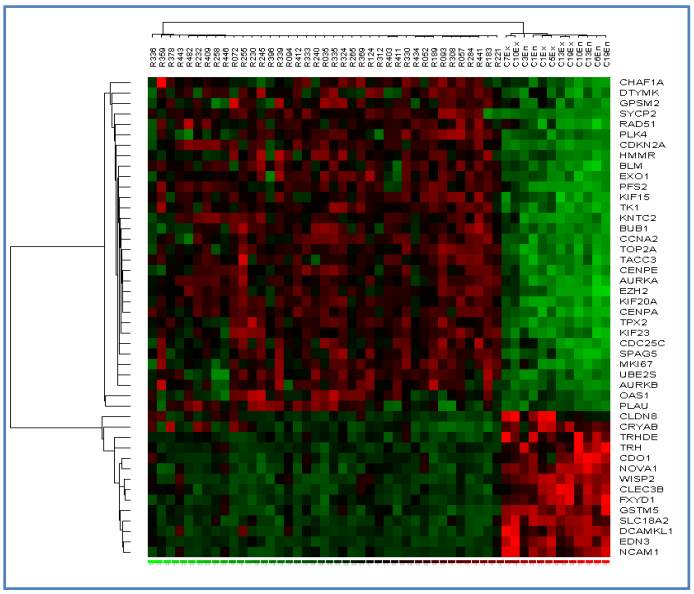


Figura 6. Clustering jerárquico de la expresión de genes de microarreglos

Este tipo de clustering permite la visualización de los resultados por medio de dendogramas y con ellos encontrar diferentes niveles de clusters y de relación entre los elementos que los integran.Aunque los patrones que representan los clusters pueden ser muy amplios y con ello pierden importancia y no es fácil determinar el número de clusters que representan la mejor clasificación.

3.2.4 Cuantización vectorial (VQ)

El método de cuantización vectorial corresponde también al clustering no jerárquico y divisible, consiste en el proceso de dividir el espacio en varias regiones conectadas, llamadas también regiones de Voronoi. Cada región es representada por un centroide, al que se unen los vectores de entrada más cercanos de acuerdo a la medida de distancia aplicada. La cuantización vectorial aquí utilizada está diseñada de manera que puede crear cualquier número de centroides y no sólo en potencia de 2, así permite trabajar con vectores de N dimensiones que se agruparán en M centroides. En la figura 7 puede verse un espacio conformado por 4 regiones con sus centroides representativos.

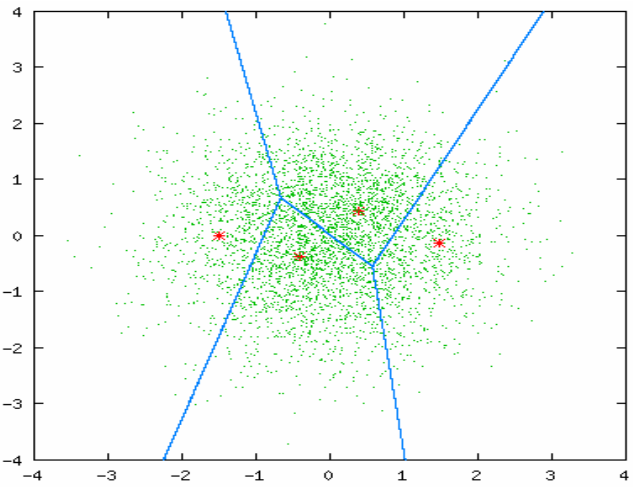


Figura 7. Espacio conformado por 4 regiones de Voronoi

Las técnicas de cuantificación vectorial [Linde 1980] se han utilizado ampliamente para la compresión de datos en el procesamiento de señales digitales y telecomunicaciones. Para la expresión génica, VQ podría ser una buena aproximación para la búsqueda de clases de tumores y también de los genes [Pham2006].

Dado un conjunto de vectores N, $p_j = \{X_{J1}, X_{j2}, \ldots X_{JM}\}$; j = 1 ... N y m = 1 ... 8793 que representa la posición de los puntos en el espacio libre, un conjunto de centroides se encuentra, los centroides representan los vectores en los racimos, y la colección de centroides se llama codebook. A continuación, el codebook está diseñado a partir de una secuencia de entrenamiento representativa de todos los vectores p_j para ser codificada por el sistema. El codebook se crea con el algoritmo [Linde 1980] Linde-Buzo-Gray (LBG), que se basa en el algoritmo de Lloyd generalizada [Lloyd 1982]

Algoritmo de cuantización vectorial:

Datos de entrada: Conjunto de *M* vectores de *N* datos, ε=umbral de distorsión
Encontrar el codebook inicial D_1, con su centroide C_1, L_m=1;

1. Repeat
2. For i=1 to M do
3. For k=1 to P do
4. Calcular $d(x_j,C_k)$, C_k es el centroide;
5. Asignar el vector X_i al centroide con min{ $d(x_j,C_k)$}
6. End For
7. End For
8. $C_{i=} C_{i+}\varphi$
9. Generar un nuevo codebook $D_m=D_m+1$
10. $L_m=L_m+1$

11. Calcular la distorsión, A
12. Until L_m>codebook_size and A<ε

El tamaño del codebook es el número de clusters necesarios que se crean para representarlos vectores de entrada. Para la visualización de las agrupaciones en el espacio es necesario reducir la dimensión de los vectores que representan los centroides, a partir de N-dimensión a 2-dimensión o 3-dimensión, esta tarea es a través del Análisis de Componentes Principales (PCA) [Everitt 1992], donde un componente es equivalente a una dimensión. Con PCA se puede ver gráficamente la distribución en el espacio de los grupos que se encuentran representados por los vectores de centroides.

3.2.5 Análisis de componentes principales (PCA)

El análisis de componentes principales es una técnica muy utilizada en campos como reconocimiento de imágenes y para encontrar patrones en datos de grandes dimensiones. Otro uso dado para PCA es la comprensión de datos, ya que puede reducir la dimensión de vectores de datos sin pérdidas de información, esta técnica es muy utilizada en el análisis de datos. Del total de factores o variables que se representan por las dimensiones se eligenlos que recogen el porcentaje de variabilidad que se considere suficiente, denominados componentes principales.

En tanto que como método de clustering, PCA busca las características principales que permitan separar en grupos a los datos, a través de la reducción de dimensiones. También puede utilizarse como un método previo a la aplicación de otro algoritmo de clustering al simplificar el conjunto de datos.

Algoritmo of Análisis de componentes principales:

Datos de entrada: Conjunto de *M* vectores de *N* datos,

1. Obtener la matriz de covarianza
2. Obtener los eigenvectors y eigenvalues de la matriz de covarianza
3. Ordenar la matriz de eigenvectors de acuerdo a los eigenvalues
4. Seleccionar losprimeros N componentes y crear los vectores de N-dimensión

Para que se pueda realizar el PCA, es necesario que las variables presenten factores comunes. Es decir, que estén muy correlacionadas entre sí. Los coeficientes de la matriz de

correlaciones deben ser grandes en valor absoluto.Como este método permite reducir el número de dimensiones de una matriz es utilizado para graficar vectores de grandes dimensiones utilizando 2 ó 3 dimensiones de los componentes más importantes (figura8).

Este método es considerado por algunos como estadístico debido a que se basa en medidas descriptivas como la media para obtener la covarianza, sin embargo, es más adecuado decir que es uno de los métodos de clustering no supervisado, que permite ver gráficamente los resultados.

La representación de los componentes principales es una matriz con tantas variables como componentes puede diferenciar el método, su dimensión de la matriz es *MxM,* con *M* variables representativas, ver tabla 1.

Tabla 1. Reducción a 5 componentes principales

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5
Variable 1	0.1077	-0.2608	0.1968	-0.268	0.196
Variable 2	0.1241	-0.191	-0.1774	-0.11	-0.174
Variable 3	0.1134	-0.2205	0.1535	-0.205	0.155
Variable 4	0.1259	-0.1885	-0.1488	-0.185	-0.148
Variable 5	0.1052	-0.1885	-0.1364	-0.195	-0.1564

Para poder crear la gráfica de la figura 8, se utilizaron sólo los 2 componentes principales de cada variable, de esta forma se puede visualizar en una gráfica de 2 dimensiones los valores

.

La matriz de expresión de genes es N-dimensional y el método de PCA permite reducir a 2 ó 3 dimensiones para poder graficarlos y visualizar las agrupaciones encontradas por éste u otro método como VQ.

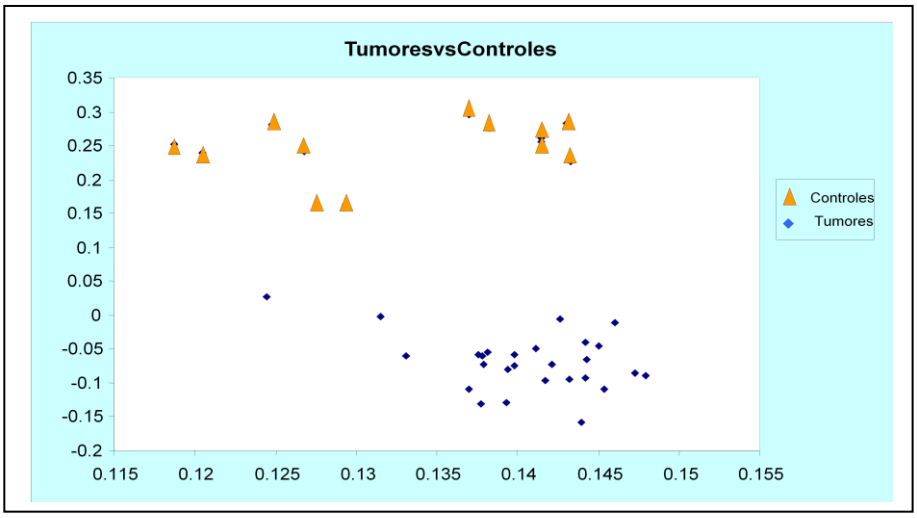


Figura 8. Gráfica en 2D a partir de la aplicación de PCA

3.3. Validación de clusters

Una vez aplicados los métodos de clustering es necesario contestar algunas preguntas como: ¿cuál es la mejor partición de los datos?, ¿cuál es el número correcto de clusters? O ¿cuál es el mejor método de clustering?. Una manera de poder contestar esas cuestiones es por medio de la aplicación de métodos de validación.

Un método de validación de los resultados obtenidos por diferentes métodos de clustering es la validación cruzada (cross-validation). Cross-validation es muy útil en los casos donde se cuenta con pocas muestras para la obtención de un clasificador. Esta técnica permite sacar varios grupos de trabajo de la muestra que se tenga y validar los resultados con muestras no utilizadas para ese entrenamiento.

La validación cruzada se hace en K iteraciones por eso se denomina también K-fold cross-validation, donde los datos de muestra se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado. Este método es muy preciso puesto que evaluamos a partir de K combinaciones de datos de entrenamiento y de prueba. La elección del número de iteraciones

depende de la medida del conjunto de datos. Lo más común es utilizar la validación cruzada de 10 iteraciones (10-fold cross-validation).

Algoritmo de k-fold Cross Validation:

1. Colocar en orden aleatorio los *N* vectores de entrada
2. S=N/k
3. Hacer k grupos con un numero S de vectores
4. For i=1 to k
5. For j=1 to to k
6. If j≠i then
7. Entrenar el clasificador con el grupo j de vectores
8. End if
9. End for
10. Hacer las pruebas del clasificador utilizando el grupo i de vectores
11. E_i = error de clasificación
12. $E_t=E_t+ E_i$
13. End for
14. e= E_t/k ; la media aritmética del error

En el análisis de la expresión de genes con microarreglos utilizar particiones creadas por K-fold puede ser muy útil debido a que el número de muestras de experimentos puede ser reducido, como resultado del costo de la tecnología o que no siempre es posible tener muestras de prueba y entrenamiento distintas. En el sistema los agentes aplican 10-fold cross-validation, para evaluar y comparar los resultados de los métodos de la agrupación, donde las muestras se dividen en dos grupos: para entrenamiento y para las pruebas.

CAPÍTULO 4
Teoría de agentes

En ciencias de la computación el desarrollo de agentes se consolida en los años 80s como una alternativa de solución a los problemas planteados en inteligencia artificial, donde agentes tanto de software como de hardware actúan en un ambiente en el que están inmersos, que perciben y que pueden modificar como resultado de sus acciones.

En este capítulo se presenta los conceptos principales de la teoría de agentes para dar un panorama general del tema.

4.1 Agentes computacionales

Una de las definiciones más aceptadas de agentes se trata de la correspondiente a Wooldrige que nos dice,

Un agente es un sistema de cómputo situado en un medio ambiente, con capacidad de actuar de manera autónoma en su entorno con el fin de cumplir con sus objetivos de diseño.

Ejercicio de la racionalidad del agente:

Percepciones => Razonamiento => Acciones

Los agentes inmersos en un ambiente perciben las características de ese medio y llevan a cabo acciones que van a modificar ese medio. Para que los agentes puedan actuar y cumplir su objetivo deben decidir qué acción es la más conveniente a realizar de acuerdo a las condiciones percibidas en el medio. El agente por lo tanto debe tener la habilidad de percibir, razonar y actuar (figura9).

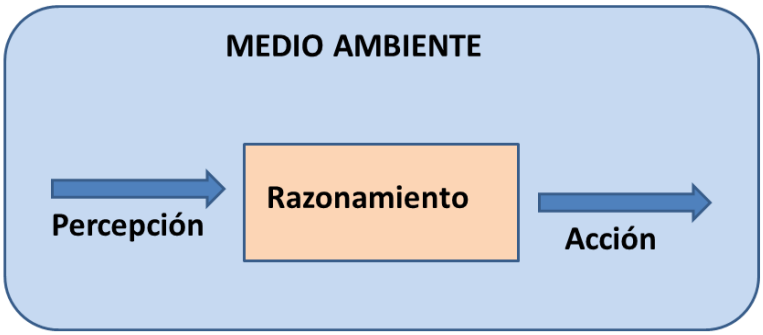


Figura 9. El agente está inmerso en un ambiente que percibe y sobre el cual actúa a partir de su razonamiento.

Cuando el agente cuenta con una autonomía flexible entonces se trata de una agente inteligente. No es necesario que todos los agentes sean inteligentes en los sistemas multi-agente. La flexibilidad del agente es dada por las características de reactividad, pro-actividad y sociabilidad [Wooldridge 1997].

- **Reactividad**. Les permite a los agentes percibir su entorno y responder oportunamente a los cambios que se producen en ella con el fin de satisfacer sus objetivos de diseño.
- **Pro-actividad.** Los agentes son capaces de mostrar un comportamiento dirigido a un objetivo al tomar la iniciativa con el fin de satisfacer sus objetivos de diseño.
- **Sociabilidad**. Los agentes son capaces de interactuar con otros agentes (incluso humanos) con el fin de satisfacer sus objetivos de diseño.

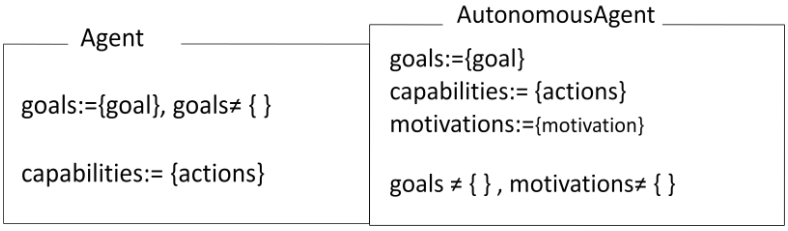
En [Luck 2001]se distinguen 4 tipos de elementos que pueden encontrarse en un ambiente en software creado como un framework de agentes: entidades, objetos, agentes y agentes autónomos.

Una entidad es una descripción abstracta, una colección de atributos. Un entorno puede entonces ser definida como una colección de entidades. Los objetos a diferencia de las entidades además de darnos una descripción a través de atributos dan una descripción de sus capacidades. Las capacidades de un objeto están definidas por un conjunto de primitivas de acción que teóricamente puede ser realizado por el objeto en algún medio ambiente y, en consecuencia, cambiar el estado de ese entorno. Cuando los objetos sirven a un propósito o a un conjunto de metas entonces se convierten en agentes, donde el detalle de descripción

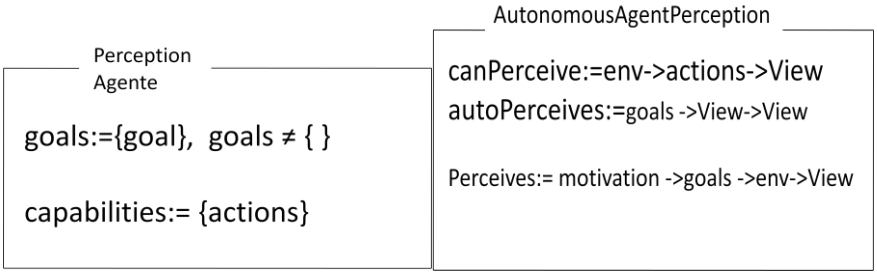
aumenta ya que también define su funcionalidad. Los agentes autónomos aparecen como una subclase de agentes, son agentes auto-motivados en el sentido de que persiguen sus propios intereses en lugar de funcionar bajo el control de otro agente, un agente autónomo como un agente con motivaciones.

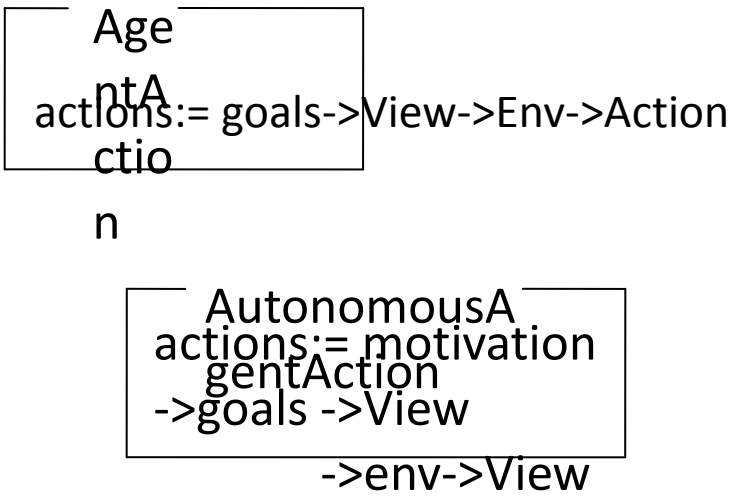
Los agentes son entidades autónomas si son capaces de ejercer elección sobre sus acciones e interacciones. Los agentes no pueden ser invocadas directamente como objetos. Sin embargo, pueden ser construidos usando la tecnología de objetos.

La definición propuesta para los agentes simples y para los agentes autónomos es:



Como puede apreciarse en las definiciones anteriores la principal diferencia entre los agentes autónomo y los simples es la motivación que tienen los primero para actuar por iniciativa propia encaminada a alcanzar sus metas. El agente autónomo cuenta con un conjunto de objetivos que quiere lograr, un conjunto de capacidades que es capaz de realizar, además de un conjunto de creencias que representa su visión acerca de su ambiente. Las motivaciones son la fuerza que afecta el razonamiento de los agentes y los conduce a la satisfacción de sus metas. Las acciones son eventos que al realizarlos cambian el estado del medio donde fueron ejecutados. La percepción de los agentes define el estado que encuentran de su ambiente.





La diferencia entre los agentes simples y los autónomos en todas las definiciones anteriores, desde la percepción hasta la representación de su estado radica en a inclusión de las motivaciones en los agentes autónomos.

Las percepciones del agente acerca de su ambiente corresponden a:

Percepciones={mfs|msgs}

mfs=magnitudes físicas, obtenidas a partir de sensores o mediciones

msgs=mensajes del entorno y/o mensajes de otros agentes.

4.2 Inteligencia artificial distribuida

La inteligencia artificial distribuida (IAD) es una rama de la Inteligencia Artificial que trata de resolver de manera distribuida los problemas, aprovechando así las ventajas propias de la programación distribuida: robustez, paralelismo y escalabilidad. La IAD considera que resolver problemas a través de una conducta colectiva es más eficiente que por una conducta individual [MAS 2005].

4.3 Sistemas multi-agente

Cuando en un mismo ambiente se encuentran dos o más agentes que interactúan para conseguir la meta del sistema, hablamos de un sistema multi-agente. Los sistemas multi-agente (MAS, por sus siglas en inglés), se basan en la inteligencia artificial distribuida, en la cual el problema es descompuesto en un conjunto de subproblemas. Esta teoría se basa en la teoría socio-biológica de la inteligencia de los primates que evolucionó debido a la necesidad de hacer frente a las interacciones sociales [Minsky 1988].

MAS es el subcampo de la IA para la construcción de sistemas complejos que implican múltiples agentes y mecanismos para la coordinación de comportamientos de agentes independientes [Veloso 2000]. MAS permite que los subproblemas de un problema puedan solucionarse a través de diversos agentes con sus propios intereses y objetivos. En estos sistemas las motivaciones son la fuerza que afecta el razonamiento de los agentes y los hace cumplir voluntariamente con normas y permanecer en una sociedad con otros agentes. En los sistemas multi-agente varios agentes modelan cada uno las metas y acciones de ellos mismos y en algunos casos de otros también. Los ambientes en los que se desempeñan los agentes en el MAS son significativamente dinámicos, ya que pueden ser afectados por otros agentes antes de que el agente actúe, esto hace que además el ambiente sea impredecible pues los resultados esperados por el agente al actuar pueden ser diferentes a los esperados por el agente mismo.

Entre las ventajas que puede dar el desarrollo de un MAS se encuentran el paralelismo, la robustes y la escalabilidad.[Veloso 2000].

Contar en un sistema con múltiples agentes podría acelerar el logro de la meta general del sistema al proporcionar un método para computación paralela. En un dominio que se dividir fácilmente en componentes con tareas independientes que pueden realizarse por separado por diferentes agentes puede beneficiarse utilizando MAS. En algunos sistemas, mientras que el paralelismo se consigue mediante la asignación de diferentes tareas o habilidades a diferentes agentes, la robustez es un beneficio de los MAS que tienen agentes redundantes, si el control y las responsabilidades son correctamente compartidas entre los diferentes agentes, el sistema puede tolerar fallos por uno o más de los agentes. Al tratarse de sistemas modulares, debería ser más fácil añadir nuevos agentes a un sistema multi-agente de lo que es añadir nuevas capacidades a un sistema monolítico.

Para una correcta interacción entre los agentes integrantes del MAS es necesario contar con mecanismos adecuados de comunicación. Para lograr la comunicación efectiva que les permita negociar o intercambiar información es necesario seguir un protocolo de comunicación. En el protocolo de comunicación se consideran los siguientes tipos de mensajes:

- Mensajes para proponer una acción
- Mensajes para aceptar una acción
- Mensajes para rechazar una acción
- Mensajes para cancelar una acción
- Mensajes para presentar el desacuerdo en ejecutar una acción

La comunicación entre los agentes les permite lograr sus resultados y mejorara los resultados en común, a través de ella coordinan sus acciones y producen comportamientos coherentes a la sociedad. La coordinación es una propiedad de los MAS donde se cuenta con un ambiente compartido. Para lograr la cooperación entre los agentes cada uno de ellos debe tener el modelo de otro agente para poder anticipar su actuación.

En la comunicación entre agentes existen básicamente dos tipos de mensajes: hechos y preguntas, todos los agentes deben aceptar información. La forma más sencilla de enviar esa información es como un hecho. También los agentes deben contestar a preguntas, teniendo tanto la capacidad de recibir preguntas como de contestarlas por envío de hechos. Los agentes de acuerdo a sus capacidades de envío y recepción pueden considerarse activos o pasivos:

Capacidad de recibir/enviar	Pasivo	Activo
Recibe hechos	*	*
Recibe preguntas	*	*
Envía hechos	*	*
Envía preguntas		*

Además de la comunicación es necesario que en el sistema multi-agente exista cooperación entre los agentes involucrados en el logro de una meta. La cooperación se refiere al hecho de que los agentes pueden necesitar ayuda de otro agente para resolver su tarea. La tarea es dividida entre diversos agentes. Cuando un agente soluciona una sub-tarea dada por otro agente le debe regresar el resultado.

Un modelo propuesto para establecer comunicación y cooperación entre los agentes es el correspondiente a la pizarra. En la pizarra los agentes muestran por medio de mensajes sus necesidades y los agentes al percibir que existen datos y mensajes que les corresponden realiza el trabajo y manda los resultados a la pizarra. En este tipo de comunicación se requiere tener control centralizado para evitar colisiones o conflictos entre los agentes [Erman 1980].

4.4 Arquitecturas para el diseño de agentes

Por sus características individuales los agentes pueden clasificarse en agentes reactivos y agentes cognitivos o deliberativos.

4.4.1 Arquitecturas de agentes reactivos

Los agentes de este tipo de arquitecturas se caracterizan por no tener como elemento central de razonamiento un modelo simbólico, es decir, su representación simbólica del ambiente y de él mismo es mínima, lo necesario para que los agentes puedan responder a los cambios del entorno de una forma inmediata a partir de una serie de reglas básicas.
Su comportamiento inteligente surge a partir de su relación directa con el medio ambiente en el cual está inmerso, fig. 10.

El modelo del agente es:

Estímulo →Respuesta

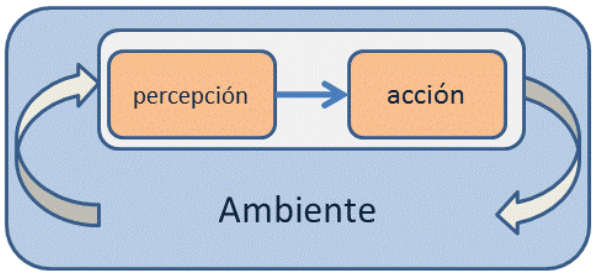


Figura 10. El agente percibe su ambiente y actúa sobre éste de manera inmediata

El comportamiento del agente se rige por la *f(p,a),* donde *p∈P*, p es una percepción que forma parte de las condiciones identificadas por el agente, y *a∈A*, es una acción que responde a la condición cumplida para llevar a cabo esa acción. Este comportamiento del agente se da en un estado s.

Así el comportamiento del agente puede describirse como C={f(p,a)| p∈P y a∈A}, donde *C* está compuesto por un conjunto de reglas *R* que responden a esa interacción.

Roodney Brooks presenta la arquitectura de subsunción [Brooks 1991], basada en el hecho de que el comportamiento inteligente puede ser generado sin utilizar representaciones del modelo simbólico y que la inteligencia es una propiedad emergente de la interacción de comportamientos simples en el entorno común, figura 11. Aquí el comportamiento de los agentes es meramente reactivo, en que reciben eventos procedentes del ambiente o entorno y realizan acciones según el evento recibido y el estado interno del agente. Una de las formas de realizar el control interno de sus acciones es a través del uso de redes neuronales o en general mecanismos de aprendizaje sub-simbólico [MAS 2005].

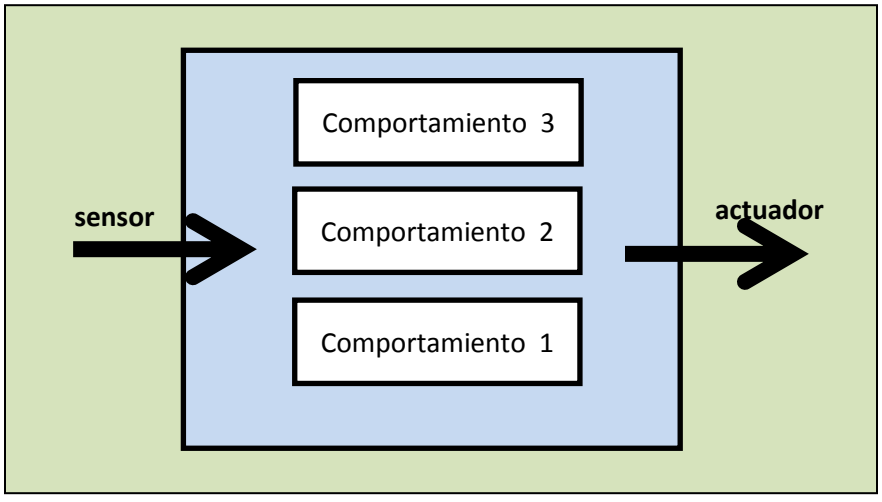


Figura 11. Para la reactividad el agente utiliza solo la información que recibe del exterior para decidir sobre su siguiente acción. Base de la arquitectura de Subsunción definida por Brooks.

Entre las ventajas que se encuentran utilizando esta arquitectura se mencionan:

- Son simples conceptualmente.
- Consumen pocos recursos y son poco costosas desde el punto de vista computacional.
- Son relativamente robustas frente a fallos.
- Imitan comportamientos biológicos simples.

Y como desventajas pueden mencionarse:

- La información local debe ser suficiente para cumplir los objetivos del sistema.
- 'Ocultan' la forma en que se crea la inteligencia (el conocimiento).
- Dificultad para construir sistemas que cumplan objetivos concretos.
- Dificultad para construir agentes con comportamientos complejos (múltiples percepciones y acciones posibles)

4.4.2 Arquitecturas de agentes deliberativos

Los agentes basados en estas arquitecturas utilizan modelos completos de representación simbólica del conocimiento, parten de un estado inicial y son capaces de generar planes para alcanzar sus objetivos. Estos agentes deliberativos se ocupan de representar el modelo del mundo de una forma simbólica completa y las decisiones se llevan a cabo por medio de razonamientos de tipo lógico.

Representan simbólicamente la información acerca de entidades complejas del mundo real y sus procesos pero además se espera que razonen con esta información de forma eficiente para obtener resultados útiles. La arquitectura representativa de estos modelos es la conocida como BDI, por sus siglas en inglés de Believes-Desires-Intentions.

Los agentes llevan a cabo la deliberación de qué metas deben realizar y cómo lograrlas. Se trata de agentes capaces de razonar sobre sus creencias e intenciones y que planifican al incluir creencias e intenciones en sus planes (fig.12).

- Creencias - *B (Believes)*. Conocimiento del agente sobre el entorno y sobre sí mismo.
- Deseos –*D (Desires)*. Metas del agente
- Intenciones –*I (Intentions)* . Manejan y conducen a acciones dirigidas hacia las metas, persisten y en ellas influyen las creencias.

El agente estructura en forma de creencias su conocimiento y se revisan continuamente. Son de 2 tipos:

1. suposiciones acerca del entorno (del ambiente y otros agentes) a partir de las percepciones.
2. suposiciones acerca de las propias capacidades del agente.

Los deseos representan fines que persigue el agente, se suelen definir con un alto nivel de abstracción. Puede haber más de uno e incluso ser incompatibles.

Las intenciones se traducen como acciones inmediatas que piensa realizar el agente para cumplir los deseos.

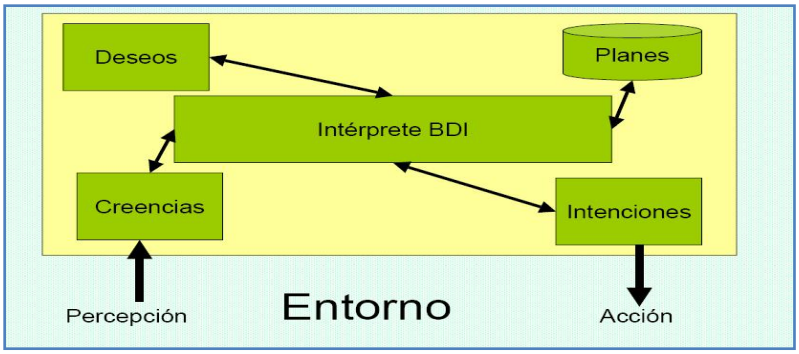


Fig. 12. Partes integrantes de un agente deliberativo con arquitectura DBI.

Los Planes son secuencia de intenciones ejecutadas para conseguir un objetivo.

Entre las desventajas de los agentes meramente deliberativos tenemos:

- Una excesiva revisión trae como consecuencia que el agente no actúa.
- Una pobre revisión no considera que el mundo evoluciona y los objetivos se vuelven inalcanzables.
- Requiere un equilibrio entre pro-actividad y reactividad.

4.4.3 Arquitecturas híbridas

Tratando de resolver los inconvenientes de las arquitecturas anteriores, se conjugan elementos de las arquitecturas reactivas y las deliberativas en una nueva denominada arquitectura híbrida. Se basa en un razonamiento práctico y en la necesidad de buscar tanto reacciones rápidas como en razonar acerca de los siguientes objetivos a alcanzar.

4.5 Sistemas multi-agente y bioinformática

Entre las aplicaciones de los sistemas multi-agente se encuentran las correspondientes a la solución de problemas bioinformáticos, ya que por su grado de complejidad en la toma de decisiones y la cantidad de tareas requeridas es un campo idóneo para la colaboración de diversos agentes.

El interés por parte de los diferentes grupos de investigación de aprovechar las características de los sistemas basados en agentes en las aplicaciones de bioinformática puede verse en los congresos y revistas relacionados con los temas de agentes inteligentes, sin embargo, hemos encontrado que en menor medida los relacionados con el análisis cuantitativo de la expresión de genes.

En seguida se presentan algunos trabajos relacionados a sistemas multi-agente y bioinformática.Entre los años 2005 y 2008 se realizó el Workshop Multi-Agent and Grid Systems for Medicine, Computational Biology, and Bioinformatics. Y desde entonces dentro de las conferencias de Sistemas multi-agentes o sistemas inteligentes se incluye el rubro correspondiente a las aplicaciones para bioinformática, pues como ya se mencionó es un campo fértil para el desarrollo de aplicaciones de la inteligencia artificial, tanto en el

desarrollo de sistemas multi-agentes como de técnicas de descubrimiento de conocimiento o aprendizaje de máquina.

A continuación se presentan características de algunos sistemas multi-agente publicados para bioinformática:

a) An intelligent agents architecture for ADN-microarray data integration, [Angeletti 2001]. Propuesta de una arquitectura de agentes inteligentes para análisis de secuencias de ADN. Los agentes se encuentran distribuidos en diferentes sitios. Utilizan las bases de datos externas. La información recopilada por todos los agentes es concentrada para trabajar con un sólo mapa de Kohonen. Los agentes no pueden desplazarse entre diferentes sitios, cada sitio tiene su propio agente, todos los agentes tienen la misma misión, se activan o desactiva en los sitios. Este sistema fue probado con sitios web simulados por ellos, no con sitios reales.

b) Agents in Bioinformatics, computational and systems biology, [Mirelli 2005]. Reporte del trabajo de grupo acerca de Agente en bioinformática (BioAgents), fundado en julio de 2004. El trabajo sobre BioAgents fue creado para impulsar el campo de la bioinformática con el diseño e implementación de herramientas de información y comunicación para apoyar en el análisis de datos biológicos y distribuir la computación de grandes cantidades de datos. Querían promover el uso de agentes inteligentes entre la comunidad bioinformática. Consideran que los agentes en bioinformática son para integrar información, manejar conocimiento y resolver problemas biológicos, además de realizar actividades que consumen tiempo y son repetitivas.

c) Applying agents to bioinformatics in GeneWeaver, [Luck 2001]. Geneweaver es un sistema multi-agente hecho para manejo y análisis de datos bioinformáticos. Está compuesto por una comunidad de agentes que interactúan entre sí. Los agentes de este sistema pueden tratar con otras bases de datos como las de secuencia y utilizan herramientas que ya existen o pueden almacenar y presentar resultados.

d) Gene expression analysis in multi-agent environment, [Lam 2006]. Es un sistema para análisis de expresión de genes por medio de tecnología de multi-agentes. Mencionan la posibilidad de utilizar aprendizaje de máquina para el análisis de datos. En su arquitectura

propuesta cuentan con 6 agentes, entre los que destaca uno que es tiene la función de una unidad de control del sistema, uno para interactuar con el usuario, otro para aplicar métodos de normalización, otro para métodos estadísticos y uno más para el manejo de bases de datos.

f) A multiagent Framework to Integrate and visualize gene expression information, [Li 2005]. Los autores presentan un sistema multi-agente para bioinformática, BioMas, obtiene información genómica de varias bases de datos y recursos web y analiza la expresión de genes de un organismo. El sistema multi-agente trabaja con bases de datos que son heterogéneas y constantemente actualizadas. En su arquitectura BioMas cuneta con agentes para revisar la secuencia y la función de los genes, hay un agente especializado para la interacción con el usuario.

g) Using multi-agent system for gene expression classification, [Stiglic2004]. Este sistema utiliza los datos de expresión de genes provenientes de microarreglos para encontrar genes significativos que permitan la clasificación de las muestras presentadas de cáncer de colón y leucemia. Los agentes se encargan de buscar pares de genes que permitan diferenciar los tipos de cáncer. Los agentes aplican métodos de clustering para probar la capacidad de los pares de genes de clasificación las muestras presentadas. No se presenta la arquitectura del sistema ni especificaciones de los agentes.

h) Sistema multi-agente para el análisis de expresión de genes (MAS- GEN) [Márquez 2015]. En este sistema se presenta un grupo de agentes encargados de las tareas necesarias para el análisis de expresión de genes con los objetivos a cumplir de selección de genes significativos y de separación de muestras.Los agentes especializado procesanla información cuantitativa proveniente de los microarreglos, aplican métodos de filtrado y aprendizaje de máquina, y consultas de bases de datos genómicas en la Web para finalmente producir listas de genes y hacer la separación de muestras.

La mayoría de las propuestas anteriores utilizan los sistemas multi-agente para resolver problemas de consulta a bases de datos externas, obtener secuencia de genes o hacer minería de texto, principalmente.En el último inciso[Márquez 2015], puede apreciarseunsistemaen el cuál el objetivo principal no es hacer consultas a la información disponible en la Web o hacer sólo el análisis de los datos numéricos, en este trabajo conjuntamos en un solo sistema ambas

metas. Esas tareas que los bioinformáticos realizan utilizando diferentes aplicaciones de software son implementadas en los agentes para que en un solo sistema de software puedan discriminarse los genes más significativos y crear agrupaciones de tipos de muestras tumorales basadas tanto en listas de genes seleccionados como con el grueso de todos los genes.

CAPITULO 5

El sistema multi-agente para el análisis de expresión de genes

El sistema multi-agente desarrollado y que se presenta en esta tesis para la identificación de genes involucrados en el CaCu y clasificación de muestras, denominado MAS-GEN, es descrito tomando conceptos vistos en el capítulo anterior, uniendo principios de diferentes autores y metodologías para lograr la definición completa del sistema. Para ilustrar el modelado de agentes utilizamos el lenguaje unificado de modelado (UML) con el que se realizaron algunos diagramas de agentes.

5.1 Arquitectura del sistema multi-agente para análisis de expresión de genes (MAS-GEN)

El sistema multi-agente para el análisis de expresión de genes implica la realización de diversas tareas que pueden desarrollarse de manera independiente y distribuida por diversos agentes autónomos. El sistema se descompone en diversos procesos: pre-procesamiento de datos, identificación de genes, clasificación de tumores y manejo de bases de datos externas. En cada uno de los módulos citados se asigna un agente operacional especializado en la realización de las tareas correspondientes para automatizar el procedimiento actual y tomar las decisiones adecuadas que den como resultado conocimiento confiable y oportuno.

Además de los agentes operacionales (agente de pre-procesamiento de datos, agente de identificación de genes, agente de clasificación de tumores y agente para el manejo de bases de datos externas), se requiere un agente administrador encargado de centralizar el control y de mediar la comunicación entre los agentes por medio del uso de una pizarra, es un agente coordinador y organizador, en la figura 13 puede verse la estructura general del sistema propuesto.

Los agentes requeridos crean una representación del conocimiento de tipo sub-simbólico y razonan para tomar decisiones acerca de qué hacer pensando en su objetivo y más aún en el objetivo del sistema completo considerando que no es necesaria su respuesta en tiempo real. Para su toma de decisiones es conveniente que se basen en un sistema de reglas de producción, en este sistema de reglas de producción cada agente cuenta con una base de hechos, un conjunto de reglas y un mecanismo de inferencia, esto le permitirá al agente representar sus creencias, elaboradas a partir de sus percepciones del ambiente, por medio de la base de hechos y con las reglas deliberar acerca de qué acción debe ejecutar.

Para la implementación de la plataforma de agentes se trabaja en el lenguaje JAVA, donde se crean las clases correspondientes a los agentes integrantes del MAS-GEN y para la toma de decisiones basadas en el manejo de reglas de producción, base de hechos y motor de inferencia para cada agente se utilizó el lenguaje CLIPS. La implementación en el lenguaje java nos permite trabajar con threads para dar paralelismo en la ejecución de diversas tareas dentro del sistema.

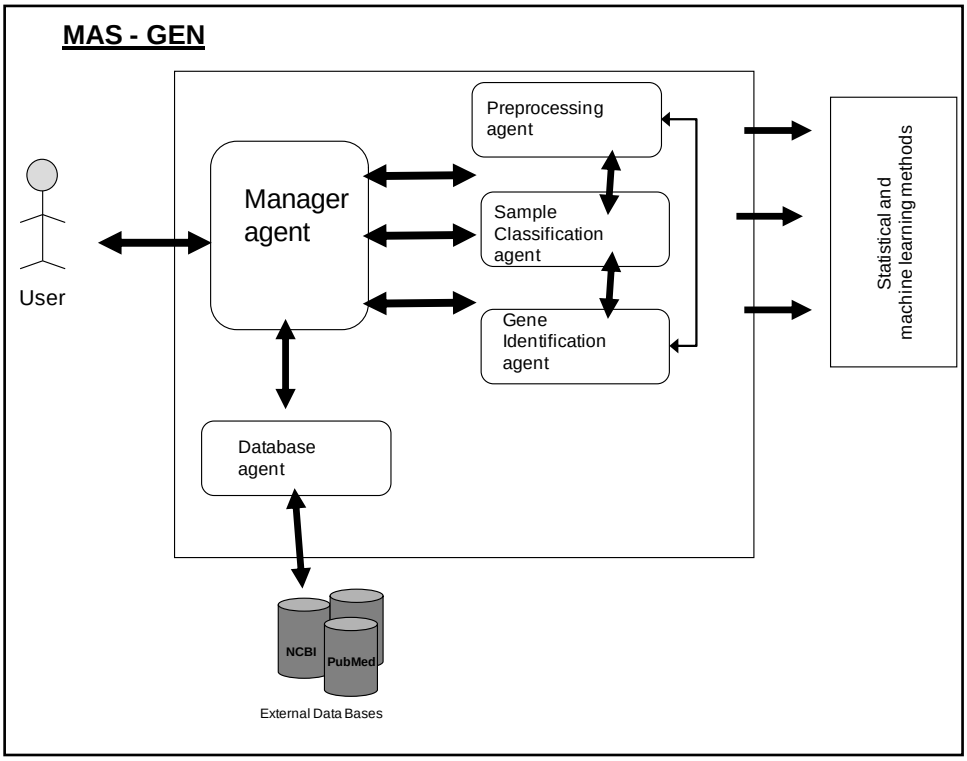


Figura13. Sistema multi-agente para análisis de expresión de genes (MAS-GEN). La plataforma de agentes operacionales coordinada por un administrador. Los agentes operacionales utilizan funciones generales que se encuentran fuera de la plataforma de agentes. El agente de base de datos externas consulta información publicada en la internet.

Para dar mayor flexibilidad de crecimiento y adaptación al sistema lo cual requiere que los procesos realizados por los agentes puedan modificarse, eliminarse o incluir nuevos, se implementa ese conjunto de procedimientos fuera de la plataforma de agentes, es decir, como un conjunto de funciones y procedimientos que los agentes pueden llamar para el logro de sus metas pero que no les pertenecen directamente sino que son independientes de los agentes, como se muestra en la figura 13. Esta independencia también ha permitido que los procedimientos y funciones que realizan los agentes hayan sido implementados en diferentes lenguajes de programación (Java, C y R), lo cual otorga también la capacidad de crecimiento del sistema al incorporar, sin causar conflictos mayores para su funcionamiento, nuevas funciones para el procesamiento, análisis y presentación de los datos.

La comunicación entre los agentes necesaria para su cooperación y coordinación es mediante el agente administrador, no hay paso de mensajes directo entre los agentes sino que cada agente solo puede enviar mensajes a una pizarra para solicitar la colaboración de otro agente. Los agentes están en continúa percepción o revisión de esa pizarra para responder a las tareas que ellos tienen implementadas.

5.1.1 Agente de pre-procesamiento de datos

El agente es utilizado para la lectura de archivos que contienen las intensidades representantes de la expresión de los microarreglos de oligonucleotidos de Affymetrix, archivos cuya extensión es .CEL. o para la recuperación de archivos de intensidades de genes previamente leídos de microarreglos y almacenados en formato de texto. Para la recuperación de los datos de intensidades de archivos de Affymetrix se utilizan las funciones desarrolladas en el lenguaje R implementadas especialmente para este tipo de microarreglos por Bioconductor [NationalCenter for Biotechnology information 2014]. Las intensidades pueden provenir de un solo microarreglo o de varios a la vez, de acuerdo a la solicitud del usuario. Posteriormente, se realiza la normalización de los datos por el método que el usuario seleccione. La normalización es con los algoritmos RMA (Robust Multichip Averaging) y MAS (Affymetrix Microarray Suite), los cuales son los más utilizados actualmente[Affymetrix 2012, Irizarry 2013].

De acuerdo a la tarea que se realizará con los datos, identificación de genes o clasificación de tumores, este agente da a los datos el formato adecuado para ser utilizados por los métodos de aprendizaje de máquina y estadística correspondientes a los siguientes módulos.

Para la clasificación de tumores se utiliza un vector de expresión de genes para un individuo (muestra de tumor o control):

$$\text{Individuo1} = (G_{11}, G_{12}, ..., G_{1n})$$

Teniendo una matriz:

$$\begin{array}{l} \text{muestra 1} \\ \text{muestra 2} \\ \text{muestra 3} \\ . \\ . \\ . \\ \text{muestra m} \end{array} \begin{pmatrix} G_{11} & G_{12} & G_{1n} \\ G_{21} & G_{22} & G_{2n} \\ G_{31} & G_{32} & G_{3n} \\ . \ . & . & . \\ . \ . & . & . \\ . \ . & . & . \\ G_{m1} & G_{m2} & G_{mn} \end{pmatrix}$$

n =8793 genes
m muestras
(tumores y controles)

vector de expresión de genes para varios individuos (muestra de tumor o control):

$\text{Gen1} = (I_{11}, I_{12}, ..., I_{1m})$

Con una matriz de la forma:

$$\begin{array}{l} \text{gen 1} \\ \text{gen 2} \\ \text{gen 3} \\ . \\ . \\ . \\ \text{gen n} \end{array} \begin{pmatrix} I_{11} & I_{12} & I_{1m} \\ I_{21} & I_{22} & I_{2m} \\ I_{31} & I_{32} & I_{3m} \\ . \ . & . & . \\ . \ . & . & . \\ . \ . & . & . \\ I_{n1} & I_{n2} & I_{nm} \end{pmatrix}$$

n =8793 genes
m muestras
(tumores y controles)

El agente de pre-procesamiento debe responder a las solicitudes del administrador recibidas por mensajes, en la figura 14, se muestra las tareas del agente de pre-procesamiento para el análisis de expresión de genes a partir de la solicitud enviada por el agente de interfaz de usuario.

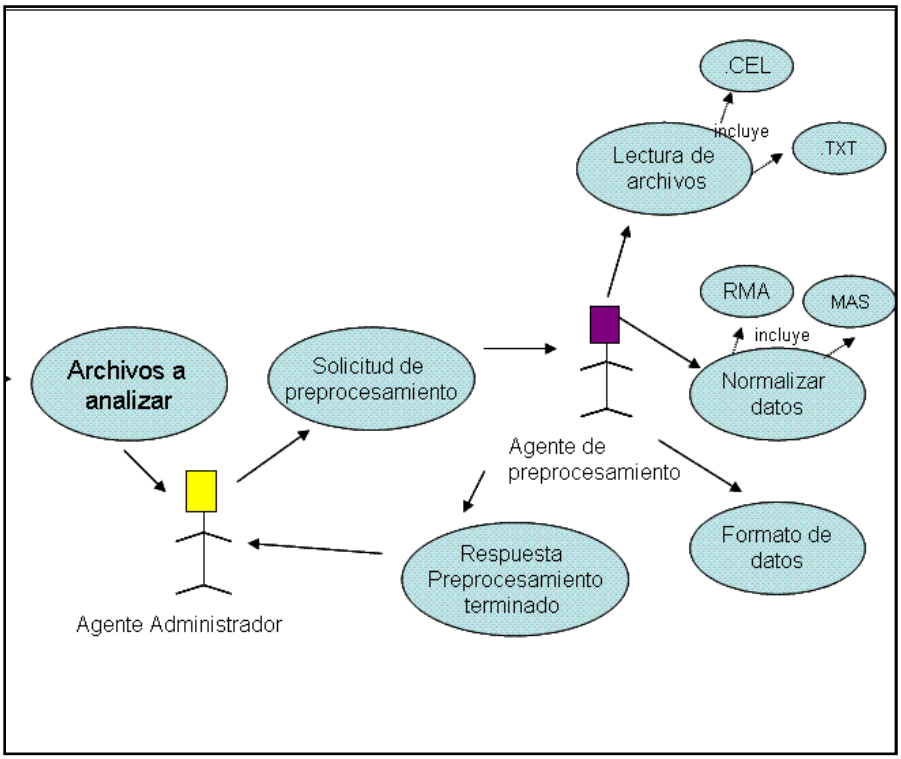


Figura14. Diagrama de caso de uso agente de pre-procesamiento de datos

5.1.2Agente de identificación de genes

Para este agente han sido implementados métodos de pruebas estadísticas ya previamente probados y empleados para esta tarea, como son t-test y el análisis significativo de microarreglos (SAM) [Tusher2001], además de permitir que trabaje con métodos de clustering -cuantización vectorial y mapas auto-organizados- estos últimos pueden aplicarse tanto a los genes filtrados previamente con los métodos estadísticos como al conjunto inicial de genes, para obtener grupos de genes. Este agente tiene como objetivo la obtención de listas de genes que representen genes candidatos a marcadores, genes bajo y sobre-regulados o grupos de genes formados. La tarea de identificación de genes requiere la colaboración del agente de interfaz de usuario que recibe la solicitud y la lista de archivos que contienen los datos de expresión; del agente de pre-procesamiento, el cual lee los archivos y proporciona los datos de expresión de genes a analizar ya normalizados y con el formato necesario; del agente de base de datos externas para la caracterización de genes con la información publicada acerca de los genes en bases de datos. También el agente de clasificación de genes

puede colaborar en esta tarea al realizar clasificación de muestras utilizando las listas de genes encontradas. Para establecer esa colaboración con los demás agentes mantiene un envío y recepción de mensajes con el agente administrador que sirve como mediador entre los agentes del sistema. En la figura 15 se muestrana través de un diagrama de caso de uso tipo UML para presentar las acciones correspondientes al agente de identificación de genes.

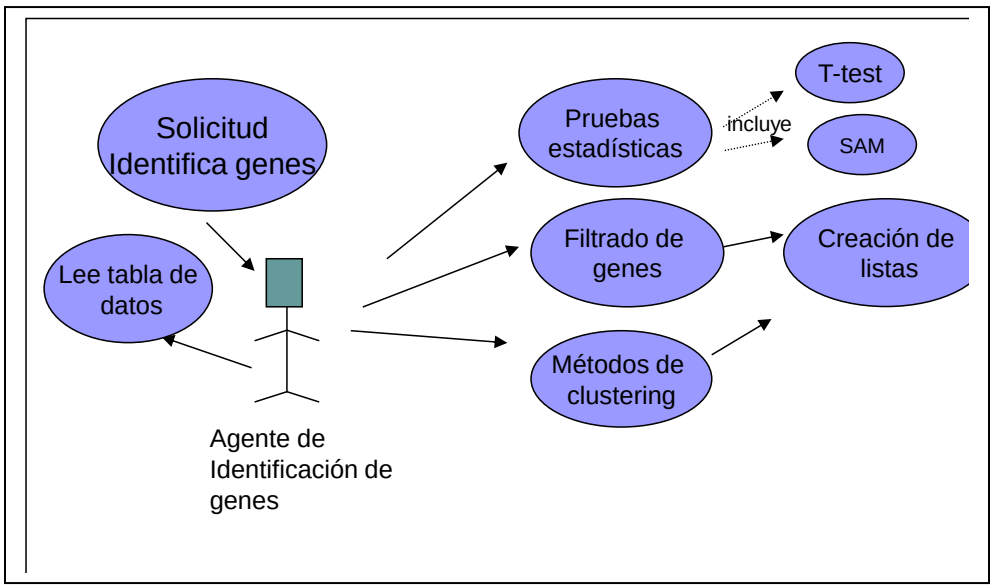


Figura15. Diagrama de caso de uso para el agente de identificación de genes

5.1.3 Agente de clasificación de tumores

Este agente está orientado a clasificar las muestras de tumores y controles utilizando la expresión de los genes. Tiene como finalidad crear grupos de muestras que representan la diferenciación de grupos de las muestras que provienen de individuos sanos de aquellos que están enfermos. También busca encontrar la separación por tipos y variantes de cáncer, esta última clasificación puede variar con respecto a la clasificación histológica que se hace debido a que este agente toma en cuenta la información numérica de la expresión de genes. Para resolver las separaciones de muestras el agente hace uso de los métodos de aprendizaje de máquina que estén implementados, como mapas auto-organizados (SOM), cuantización vectorial (VQ) y análisis de componentes principales (PCA). Este agente puede aplicar todos los métodos de clustering implementados y comparar los resultados obtenidos para decidir

una clasificación final. Otra tarea que puede realizar este agente es la predicción de la clase a que clase pertenece la muestra de un nuevo paciente una vez que se cuenta con los grupos de tumores, lo cual contribuye para determinar el tratamiento a seguir. La figura 16muestra el comportamiento que puede realizar el agente de clasificación de muestras.

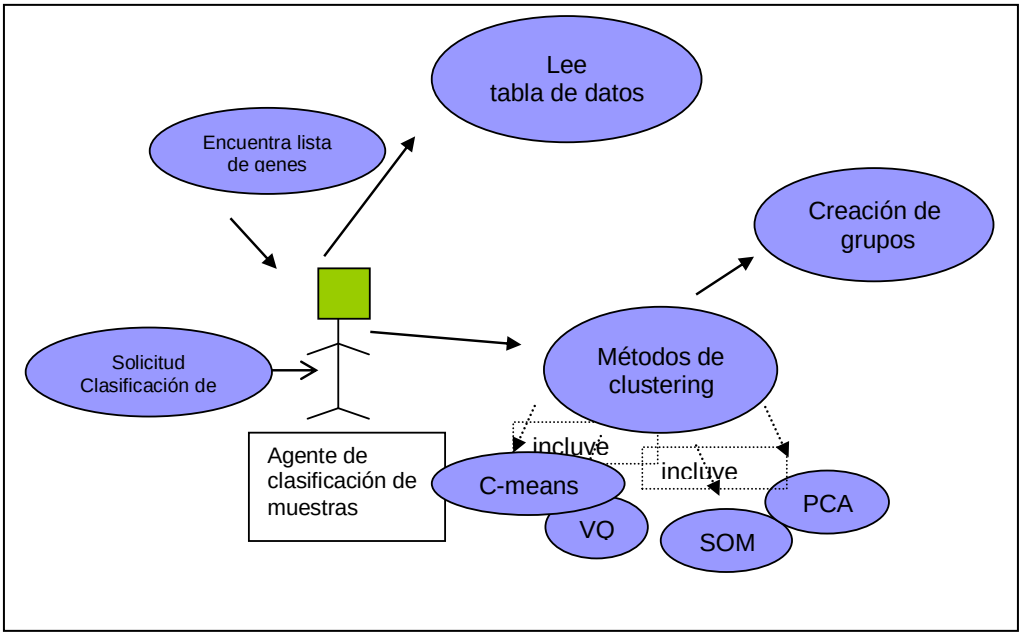


Figura16. Diagrama de caso de uso para el agente de clasificación de muestras, donde se distingue las posibilidades de su comportamiento

5.1.4 Agente de bases de datos

Para lograr la identificación de los genes se requiere de su caracterización por medio de información que existe en bases de datos externas y este agente se encarga de la búsqueda de dicha información. Existen en la actualidad diversas bases de datos genómicas que dan acceso a su información por medio de la Internet, información que ayudará en nuestro sistema al conocimiento de los genes, es importante por lo tanto poder establecer la conexión y obtención de datos de dichas bases. Como se observa en la figura 17, este agente por medio del uso de webservices se encargará de interactuar con las bases de datos externas y homogeneizar la información que pueda ser utilizada por otro agente para la caracterización de los genes.

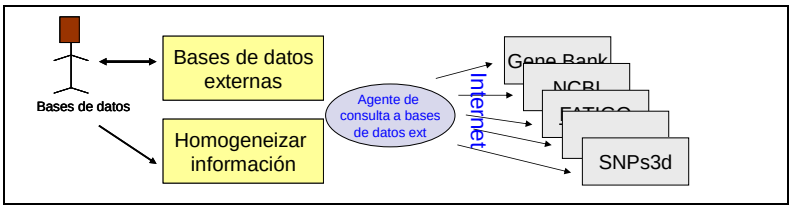


Figura17. Diagrama de agente de base de datos

5.1.5 Agente administrador

El agente administrador coordina todas las acciones del sistema multi-agente. Es el encargado de recibir las peticiones y respuestas de los agentes operacionales y enviar los mensajes necesarios entre ellos para lograr el funcionamiento del sistema. Hace la planeación global del objetivo a resolver y libera la información de lo que los agentes han hecho para que otro agente continúe las tareas de acuerdo a su especialización. Toda la comunicación que se requiere entre los agentes es mediada por este agente, ya que es el encargado del envío y recepción de mensajes por parte de los agentes.

De acuerdo con [Luck 2001] que indica que podemos clasificar las entidades en un entorno en objetos, agentes y agente autónomo. En nuestro sistema tenemos agentes simples y agentes autónomos, ya que creemos que todas las entidades tienen una meta a alcanzar. Los agentes autónomos son los agentes motivados por ellos mismos para actuar en el medio ambiente, como los agentes de clasificación de muestras, el de pre-procesamiento y el de identificación de genes. El agente de base de datos no es autónomo porque otros agentes lo invocan para hacer su tarea.

Pre-processing agent

Autonomous Agent
Goal-type: active.

Goal: provide the data genes.

Perception:
Perceiving actions: gene expression data read from files.
CanPerceive: state of expression data, aims for data.
WillPerceive: new necessities of other agents with data.

Actions:
Reading files of gene expression data, Creation a matrix of data.
data normalization and standardization, format the data genes.

Sample Classification Agent

Autonomous Agent
Goal-type: Active.

Goal: Create sets of samples.

Perception:
Perceiving actions: need to create sets of samples, creation of lists of genes to identification of genes.
CanPerceive: Sets of genes, sets of samples.
WillPerceive: Sample differentiation.

Actions:
*C*reate sets of samples using gene expression profiles through machine learning methods.

Gene IdentificationAgent

AutonomousAgent
Goal-type: Active.

Goal: Create sets of relevant genes.

Perception:
Perceiving actions: Need to create sets of genes to get relevant genes.
CanPerceive: profiles expression of genes.
WillPerceive: grade of correlated of genes in the lists, sample differentiation, and characterization of joined genes in the lists.

Actions:
Create sets of genes using gene expression profiles through machine learning methods, evaluated the relation of genes into a list, evaluation of information gotten by Database Agent about genomic medicine.

DataBases Agent

Non autonomous Agent
Goal-type: Pasive.

Goal: Characterization of genes.

Perception:
Perceiving actions: Request of characterization of lists of genes to identification of genes.
CanPerceive: Messages from manager agent.

Actions:
Consult of public databases to integrate the knowledge about genes related with the disease of study.

Manager Agent

Autonomous Agent
Goal-type: Active.

Goal: Facilitate the communication between agents, interaction with the user

Perception:
Perceiving actions: Request of autonomous agents to non-autonomous, requests of the user, conflicts between agents.
CanPerceive: Communication of the user, results from other agents.

Actions:
Interaction with the user system, communication with agents, registration of agents, and resolution of conflicts.

5.1.6Responsabilidades de los agentes

En la tabla 2 se presenta una lista de responsabilidades de los agentes de operación como resultado del análisis funcional.

Tipo de Agente	Responsabilidades
Agente:Pre-procesamiento de datos	Responder a solicitudes de lectura de datos desde archivos Realizar la normalización de datos Responder a solicitudes de análisis estadístico de datos Dar formato a datos para métodos de análisis Enviar mensajes a la pizarra.
Tipo de Agente	Responsabilidades
Agente:Identificación de genes	Responder a solicitudes de creación de listas de genes Aplicar pruebas estadísticas a genes Realizar filtrado de genes Generar listas de genes Realizar clustering para listas de genes Comparar entre listas de genes Solicitar información de listas de genes Caracterización de genes Obtener estadísticas de genes Recibir y enviar mensajes a la pizarra
Agente:Clasificación de tumores	Obtener estadísticas de muestras (tumores y controles) Responder a solicitudes de creación de grupos de muestras Aplicar clustering a muestras Generar grupos de muestras Evaluar grupos de muestras Recibir y enviar mensajes a la pizarra
Agente: Base de datos externas	Comunicarse con bases de datos externas que lo permitan Integrar información de genes Revisar la información de los genes proveniente de diferentes

	bases de datos Responder a la solicitud de información de genes Recibir y enviar mensajesal Agente Administrador
Agente Administrador	Coordinar a los demás agentes Comunicar a los agentes, Establecer comunicación con el usuario

Tabla 2. Agentes del sistema MAS-GEN y sus responsabilidades

5.2 Implementación de los agentes del MAS-GEN

Para la implementación del sistema multi-agente MAS-GEN, se crearon las clases correspondientes de los agentes operacionales y del administrador, que heredan a partir de una superclase Agente que cuenta con los atributos y métodos que todos los agentes requieren como base para su comportamiento. En la figura 18 se presenta la definición de clases para los agentes y para los mensajes que permitan establecer la comunicación entre los agentes.

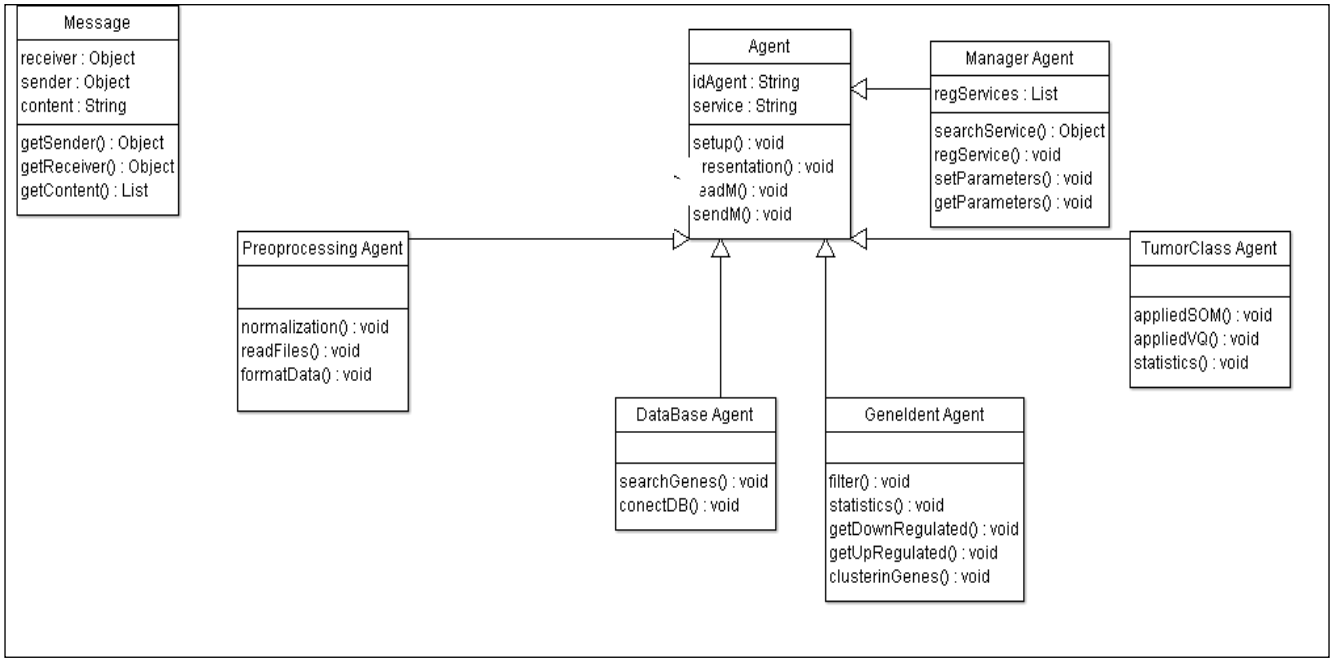


Figura 18. Diagrama de clases principales del sistema multiagente, MAS-GEN.

Los mensajes que se enviarán de los agentes operacionales hacia el administrador y viceversa tiene el siguiente formato:

Sender: Agente

Receiver: Agente

Content: Lista

Debido a que solo se pueden enviar mensajes entre el administrador y los gantes operacionales, siempre en cualquiera de los dos atributos, sender o receiver, uno de ellos será el administrador. El contenido está integrado por una lista de argumentos que indican el tipo de tarea que se solicita al agente que realice, junto con los parámetros que debe considerar para la toma de decisiones.

Para la toma de decisiones de los agentes se definieron algunas reglas y hechos en el lenguaje Clips para interactuar con el lenguaje JAVA en el que están implementados los agentes. A continuación se presentan algunas de las reglas definidas.

Algunas de las reglas definidas para la toma de decisiones de los agentes se presentan a continuación, escritas con la sintaxis que marca el lenguaje Clips.

Regla para integrar en el ambiente de los agentes los parámetros básicos para el filtrado de genes

```
(defrule ponParam
(solicitud (tipo setParam))
(normal ?n) (pvalue ?pv)   (fch ?fc) (folder ?d)
=>
(assert (parametros (norma ?n) (pvalue ?pv) (fch ?fc) (folder ?d)))
)
```

Regla para actuación del agente de identificación de genes para el filtrado de genes

```
(defrule filtraGenes
?f<-(solicitud (tipo identGenes) (estado inactivo) (subTipo filtrado))
```

```
  (readFiles yes)
  (parametros (norma ?n) (pvalue ?pv) (fch ?fc) (folder ?d))
=>
 (modify ?f (estado ejecuta))
 (comando filtrado ?n ?pv ?fc ?d)
)
```

Reglas para actuación del agente de clasificación de muestras

```
(defrule clasificaMuestras
 f<-(solicitud (tipo classSamples) (estado inactivo) (subTipo listaFiltrado))
=>
 (modify ?f (estado ejecuta))
 (comando clustering VQ SOM PCA)
)
------
(defrule clasificaMuestras2
 f<-(solicitud (tipo classSamples) (estado inactivo) (subTipo listaGenes))
 (readFiles yes)
=>
 (modify ?f (estado ejecuta))
 (comando clustering)
                                                                    )
```

CAPÍTULO 6

Pruebas y resultados

Parala realización de esta tesis presentamos en este capítulo el análisis de expresión de genes de casos y controles respecto al caso de estudio que nos ocupa, el CaCu. A partir de los datos de 52 microarreglos almacenados en archivos con formato numérico y la solicitud al sistema de la identificación de genes o la clasificación de muestras, los agentes inician su actuación para alcanzar la meta planteada. Los resultados alcanzados se presentan aquí en forma de tablas y con gráficas para su visualización.

6.1 Caso de estudio

Las muestras de obtenidas de los microarreglos generados en el Departamento de Medicina Genómica del Hospital General de México, de 39 mujeres afectadas con CaCu y 12 de mujeres sin control, sin padecimiento de CaCu, nos fueron proporcionados los 51 archivos que contienen la representación numérica, con extensión .CEL, generados por el software de Affymetrix lector de los microarreglos. Losmicroarreglos o microchips biológicos se hacen con materiales de vidrio o silicio fabricados e implementados con sondas. Estas sondas se utilizan para recopilar y analizar los datos a nivel de expresión que a continuación se clasifican como valores más categorizados en dos grupos, a saber, MM (desajuste) y PM (perfecto).

Para la obtención del conocimiento que debía implementarse en los agentes se realizaron diversas visitas al Departamento de Medicina Genómica del Hospital General de México para la observación directa y comprensión del procedimiento, que se realiza por el personal médico y estudiantes de esta Unidad. Posteriormente se definió la arquitectura del sistema multi-agente que se encarga de la automatización de todo el proceso, con la finalidad de obtener una herramienta inteligente que les facilite esta tarea, obteniendo resultados confiables.

Además de la observación directa del proceso se revisa el software que utilizan para el análisis de expresión de genes y las técnicas estadísticas empleadas, así como, la revisión en la literatura (libros y revistas) de los métodos de aprendizaje de máquina y técnicas estadísticas recomendados para el análisis de expresión de genes.

Por otro lado, se identifican las bases de datos externas que son de libre acceso en la Web, considerando ¿cuáles son las más importantes?, ¿qué información ofrecen acerca de los genes? y ¿qué puede ser relevante para la toma de decisiones al momento de la selección de genes?,, es importante definir aquellas que serán consultadas directamente por el agente que dentro del sistema será el encargado de esta tarea.

Algunos de los métodos de aprendizaje de máquina, clustering, que han probado ya su utilidad para la agrupación de genes y clasificación de tumores, fueron implementados en el sistema para ser utilizados por los agentes. De la misma forma se utilizan técnicas estadísticas para el filtrado o preselección de genes recomendadas por los expertos de Medicina Genómica y en las publicaciones relacionadas. Todos estos métodos son integrados en el set de procedimientos y funciones que los agentes operacionales pueden utilizar, junto con el conocimiento que para ello requieren los agentes.

6.2 Resultados alcanzados

Los agentes que integran el sistema multi-agente fueron implementados con la definición de las clases, de los mensajes y las reglas para la toma de decisiones que les corresponden.

Los métodos de aprendizaje de máquina implementados son los clustering de tipo mapas auto-organizados (SOM), cuantización vectorial (VQ), fuzzy c-means, clustering jerárquico, y análisis de componentes principales (PCA). Los dos primeros se aplican tanto para hacer clusters de genes como de muestras en diferentes etapas por los agentes de identificación de genes y clasificación de muestras.

6.2.1 Clasificación de muestras

Con 39 muestras de tumores (13 adenocarcinomas , encontrados en células glandulares, y 26 epidermoides, en células del epitelio) y 12 muestras control, utilizando todos los 8793 genes provenientes de los microarreglos se obtuvieron los siguientes resultados:

a) Por cuantización vectorial

Con 2 clusters, tratando de separar entre muestras de cáncer y controles, se obtiene una clasificación exacta considerando los resultados esperados, como muestra la tabla 3. Posteriormente se intento separar por tipos de cáncer, adenocarcinoma y epidermoide, los resultados están en tabla 4. Allí puede apreciarse que sin problemas puede separar controles de tumores, pero en la separación de tipos de tumores, 3 de tipo epidermoide los puso en el cluster de adenos y una muestra de adenos en el cluster de epidermoides.

Tabla 3. Agrupación en dos clusters, controles y tumores, con VQ.

	Cluster1	Cluster2	Total
Tumores	39	0	39
Controls	0	12	12
Total	39	12	51

Muestras clasificadas con 0% de error

Tabla4. Con 3 clusters, clasificación por tipo de cáncer (epidermoide y adenocarcinoma) y controles, con VQ

	Cluster1	Cluster2	Cluster3	Total
Epidermoide	23	3	0	26
Adenocarcinoma	1	12	0	13
Controles	0	0	12	12
Total	24	15	12	51

El error de clasificación es de 7.8%

b) Con el método de mapas auto-organizados

En esta parte se presentan los resultados alcanzados al utilizar el método de clustering denominado mapas auto-organizados. En tabla 5 se presenta la separación en dos grupos: tumores y controles, la cual pudo realizarse con un error mínimo al haber puesto un control dentro del cluster de tumores. En la tabla 6, se muestra la separación de grupos que hace al método SOM con los genes, el agente para hacer la separación entre los tipos de muestras hacer una látice de 2x2 pero solo se agrupan con 3 para los subtipos de tumores.

Tabla 5. Agrupación de tumores y controles, en 2 clusters, con SOM:

	Cluster1	Cluster2	Total
Tumores	38	1	39
Controls	0	12	12
Total	38	13	51

Muestras clasificadas con 2% de error

Tabla 6. Con una látice de 2x2, se crearon sólo 3 clusters y el cuarto quedó vacío, utilizando SOM.

	Cluster1	Cluster2	Cluster3	Total
Epidermoide	3	6	17	26
Adenocarcinoma	12	0	1	13
Total	15	6	18	39

El error de clasificación es 7.8%

c) Por Fuzzy C-means.

La clasificación de muestras por este método permitió diferenciar completamente los controles de los tumores, además de ofrecer los grados de membresía que en la división de 3 clusters resulta muy interesante para diferenciar los tipos de tumores. En las tablas 7 y 8,

puede verse el resultado de clasificar las muestras por este método, con 2 y 3 clusters respectivamente.

Tabla 7. Dos grupos de muestras encontrados por c-means, uno de controles y el otro de tumores

	Cluster1	Cluster2	Total
Tumor	39	0	39
Controles	0	12	12
Total	39	12	51

El error de clasificación es 0%

Tabla 8. Tres grupos de muestras encontrados por c-means, uno de controles y dos para los tumores

	Cluster1	Cluster2	Cluster3	Total
Tumor	20	0	19	39
Controles	0	12	0	12
Total	20	12	19	51

El error de clasificación es 0%, si consideramos que separo casos y controles

d) Por PCA

Por este método podemos visualizar de manera gráfica la agrupación de las muestras estudiadas, los resultados se presentan en la figura. 19.

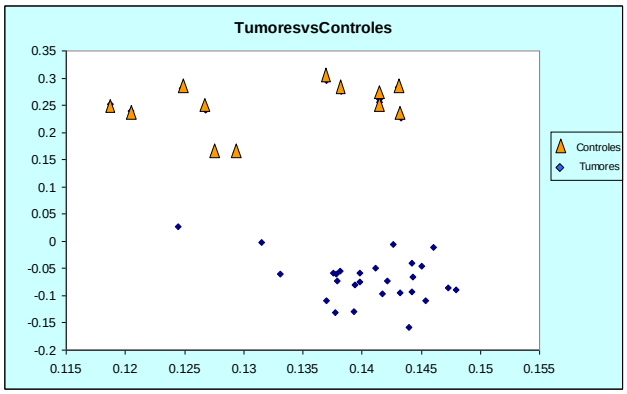


Figura 19. Gráfica de PCA de la clasificación de tumores y controles

Para los tipos de cáncer se compara con respecto a la clasificación histopatológica dada, sin embargo se encontró que en las muestras que tanto en VQ como SOM son las mismas muestras las que se presentan mal clasificadas, es decir que la expresión de sus genes no corresponde a su clasificación histopatológica.

6.2.2 Identificación de genes

Para la identificación de genes se han realizado diversas pruebas, donde intervienen los agentes de pre-procesamiento, de identificación de genes, de clasificación de tumores y de bases de datos externas coordinados por el agente administrador.Utilizando los datos de microarreglos proporcionados por el depto. deMedicina Genómica del Hospital General, de 51 muestras de mujeres mexicanas, 39 con cáncer cervical y 12 controles (mujeres sin CaCu).

Los datos son presentados al sistema en un formato emitido por el software que lee los microarreglos y los almacena como .CEL, estos son manipulados por el agente de pre-procesamiento para leer los archivos y normalizarlos para emitir al final una tabla de datos numéricos que representan las intensidades de expresión de los casi 9000 genes de cada microarreglo. Los genes son filtrados para reducir su número y quedarse solo con aquellos que se expresan diferencialmente para tumores y controles. La parte de filtrado de genes la hace el agente de identificación de genes con la aplicación de la pruebas estadísticas como T-student y SAM (Significance analysis of microarrays), que en combinación con el método fold change, permite obtener listas de los genes que se expresan diferencialmente en el análisis de expresión de genes. Separa los sobre-expresados y los sub-expresados y de acuerdo al corte que se tome en el parámetro fold change se obtiene las listas de genes. El agente de identificación de genes prueba varios cortes para obtener las listas y encontrar la lista de genes más significativos para los experimentos. Toma tanto sobre como sub expresados.

Después de filtrar la lista original de genes, aquellos que han sido seleccionados para candidatos como genes diferenciales, este mismo agente de identificación de genes aplica métodos de clustering para formar nuevos subgrupos de genes, los clusters obtenidos permiten considerar la relación estrecha que pueden tener algunos genes entre sí.

Estos nuevos grupos de genes los toma el agente de clasificación de muestras y con ellos aplica métodos de clustering para probar la capacidad de estos genes en grupo de clasificar las muestras, tanto de controles con tumores como del tipo de tumores.

El agente de clasificación de muestras presenta al grupo que mejores resultados da para la clasificación de muestras. Posteriormente se elige uno de esos subgrupos de genes para consultar información de cada gen mediante el agente de bases de datos externas.

El agente de bases de datos externas es activado para consultar las bases de datos externas que se encuentran en internet y tratar de distinguir aquellos genes que estén relacionados con el padecimiento de CaCu.

Las listas de genes que se prueban pueden ser muchas y hacerlo de manera manual, es decir, el usuario hacer esa selección resultaría tardado y tedioso, por esto es importante que los agentes puedan encargarse de la automatización para agilizar la obtención de resultados y poder hacer diversas pruebas con diferentes listas de genes diferenciados. En la tabla 8 se presentan dos de las listas seleccionadas por el agente como propuesta de genes diferenciados.

Con los datos de nuestro caso de estudio se trabajó también por separado con la lista de los genes que están filtrados como sobre-expresados, de esa lista surgen más de 100 genes que presenta dentro de sus principales genes varios que ya han sido reportados por su relación con CaCu, como MCM5, MCM2, CDKN2A y CCNA2, además de otros que coinciden con los seleccionados por las investigaciones que realizan en el departamento de Medicina Genómica.

Cuando el agente de identificación de genes genera una lista de genes seleccionados el agente de clasificación de muestras hace su trabajo de verificar la capacidad de cada lista para la agrupación de muestras aplicando los algoritmos de clustering disponibles. Los resultados que da al aplicarle los algoritmos de clustering con el agente de clasificación de tumores para verificar algunas agrupaciones de muestras se presentan en la tabla 9para la separación en 2 grupos: tumores y controles.

Tabla 9 . Listas de los mejores genes seleccionados

Lista	Genes	#Genes reportados
List 1 (7 genes)	• PRC1, CDC20, TOP2A, ZWINT, CDC2, MCM5, CCNB2.	5
List 2 (5 genes)	• NUSAP1, RFC4, CKS2, MCM2, UBE2C.	4

En la figura 20 puede apreciarse la gráfica del patrón de expresión de los genes de cada lista seleccionada, el patrón de esos genes es muy similar en su expresión de genes en las muestras estudiadas.

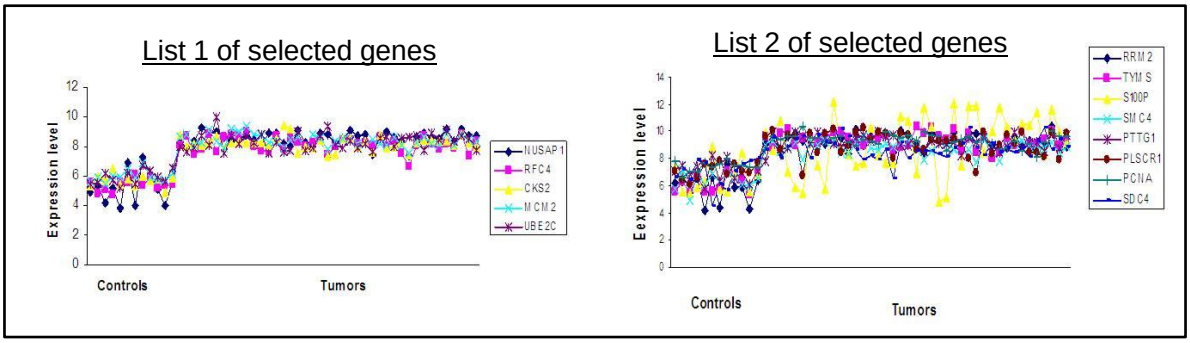


Figura20. La lista de genes seleccionados permite una correcta separación entre tumores y controles.

En las siguientes figuras se aprecia la capacidad de esas dos listas seleccionadas para separar las muestras, el agente de clasificación de muestras genera esas agrupaciones de las muestras al aplicar los diferentes métodos que tiene a las listas que el agente de identificación de genes libera. Las dos listas seleccionadas pueden separar perfectamente las muestras de tumores y controles con clustering jerárquico, ver figura 21.

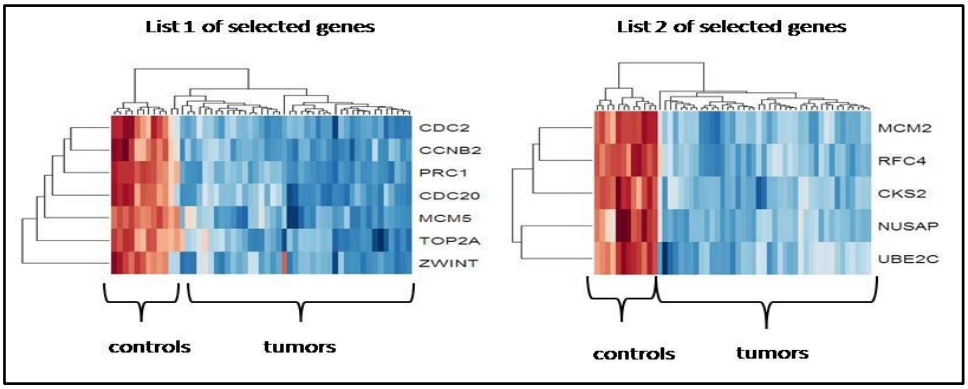


Figura21. Las 2 listas de genes seleccionados permite una correcta separación entre tumores y controles con el método de clustering jerárquico.

Si se aplica el método de análisis de componentes principales, como pueden verse en las gráficas de 2 dimensiones de la figura 22, a las dos listas de genes se encuentra que la separación de tumores y controles es correcta.

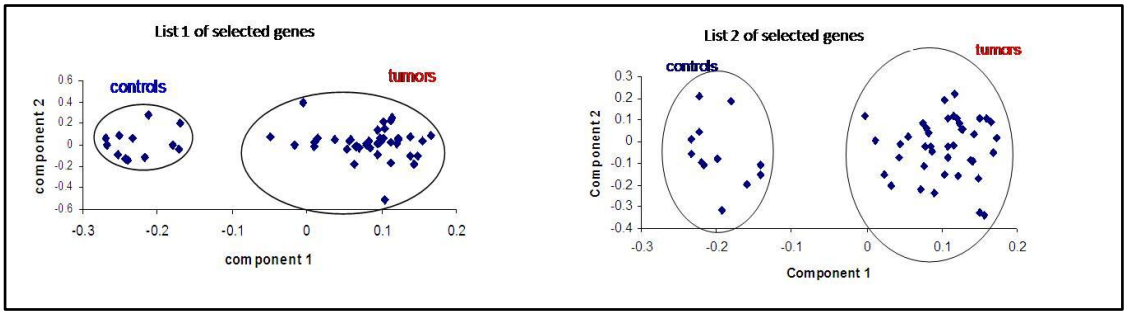


Figura22. Las 2 listas de genes seleccionados con el método de PCA permite una correcta separación entre tumores y controles, utilizando 2 componentes principales.

Entre las consultas que hace el agente de bases de datos para la caracterización de los genes seleccionados está la correspondiente a las funciones biológicas de los genes, en la tabla 10 se presentan las funciones biológicas de los genes integrantes de las listas, el porcentaje de genes de cada lista que se relacionan con esa función biológica.

Tabla10. Funciones biológicas de los genes por listas seleccionadas

Lista	**Funciones Biológicas**	**% genes**
1	Cyclin-dependent protein kinase activity	14.3
	Protein kinase binding	14.3
	Ubiquitin binding	14.3
	ADN binding	28.6
	Atpase activity	57.1
	Protein binding	57.1
2	ADN replication origin binding	20.0
	Protein kinase regulator activity	20.0
	ADN binding	40.0
	Nucleotide binding	40.0
	Atpase activity	80.0

Finalmente se hizo la validación de clusters por 10-fold para los diferentes métodos de clustering aplicados a la clasificación de muestras, la tabla 11 indica el porcentaje de certeza para las listas de genes.

Tabla 11. Resultados de la validación 10-fold para la clasificación de muestras de las 2 listas seleccionadas de genes

Genes	Method	Tumors Vs. Controls	AdenosVs epidermoides
Lista 1 de genes (PRC1, CDC20, TOP2A, ZWINT, CDC2, MCM5, CCNB2)	C-means	99%	67%
	SOM	95%	70%
	VQ	100%	73%
Lista 2de genes (NUSAP1, RFC4, CKS2, MCM2, UBE2C)	C-means	100%	74%
	SOM	97%	59%
	VQ	100%	55%
	C-means	100%	61%

Los resultados aquí obtenidos en la clasificación de las muestras entre los casos y controles tienen una clasificación errónea mínima, por lo que los grupos de genes seleccionados se podrían utilizar para determinar si una muestra corresponde o no a tejido sano. En la clasificación de las muestras por los subtipos de CaCucomo el adenocarcinoma y el carcinoma de células escamosas la tasa de errores de clasificación es relativamente alta debido al desacuerdo entre las clasificaciones histológicas por la morfología citológica del tumor y los datos cuantitativos que se tienen en los microarrays. MAS-GEN es un sistema

inteligente para ayudar a los biólogos y equipos médicos en el análisis de la expresión génica para entender los detalles genéticos de la enfermedad, representa una mejora significativa en el estado de análisis de expresión génica. Como un sistema multi-agente MAS-GEN, proporciona una herramienta flexible, utilizando agentes dirigidos a un objetivo que permiten naturalmente la descomposición del complejo problema en submetas de los agentes en un sistema de software único, esta es una ventaja clave sobre el uso de varias aplicaciones de diferentes software,por el ahorro de tiempo y reducción de errores en la gestión del proceso a través de diferentes formatos de datos y la reducidaintervención del usuario.

CAPÍTULO 7

Conclusiones y trabajo futuro

7.1 Conclusiones

El análisis de expresión de genes por la diversidad de tareas y su posibilidad de distribución en subsistemas autónomos e independientes es viable a través de un sistema multi-agente, en el que colaboren los agentes operacionales especializados en tareas como pre-procesamiento de datos, identificación de genes, clasificación de tumores, y manejo de bases de datos.

Este sistema logra hacer más fácil la realización del análisis de expresión de genes debido a que automatiza en un solo software los pasos que el investigador realiza para dicha tarea, en cuanto a la identificación de genes y clasificación de muestras. Hace transparente para el usuario las tareas a realizar y las decisiones que toma para encontrar genes que puedan diferenciarse en ciertos procesos o crear grupos de muestras de datos. La labor del usuario se reduce a proporcionar los archivos de donde se obtendrán los datos de microarreglos e indicar cuáles son las muestras de casos y controles. El sistema está utilizando el conocimiento del experto en análisis de expresión para que otros puedan realizar estas tareas.

Es importante resaltar que el objetivo principal del sistema de automatizar el proceso deanálisis de expresión de genes por medio de un solo sistema que integralas actividades que los encargados de esa labor, la cual les puede demorar varios días debido a que una utilizan diversos programas de software, nuestra aplicación permite hacerla completamente con resultados muy similares a los de encontrados por los especialistas a través del método tradicional. Entre las ventajas alcanzadas por el sistema a través de la utilización de agentes destaca la facilidad para usuarios nuevos o inexpertos en el análisis de expresión de genes para obtener resultados en un tiempo muy reducido y sin tener que capacitarse ampliamente en el dominio.

La implementación de los agentes da flexibilidad en su desarrollo y permite su sencillez en comparación con las herramientas ya existentes que requerían de la implementación de múltiples agentes para la administración del sistema, en MAS-GEN esto se concretó sólo en el agente administrador.

La separación de los métodos de filtrado y aprendizaje de máquina de la plataforma de agentes permite modificar, agregar o eliminar procesos en el sistema, sin que los agentes operativos se vean afectados en cuanto a su diseño e implementación. Esto resulta de gran utilidad debido a que por razones de crecimiento del área de bioinformática, continuamente surgen nuevos procedimientos o mejoras a los ya existentes por lo que el sistema puede crecer o actualizarse evitando quedar obsoleto.

7.2 Trabajo futuro

Las herramientas de consulta de datos acerca de los genes en la Web cambian constantemente, aparecen y desaparecen con relativa frecuencia. La tarea del agente de base de datos encargado de obtener la información de los genes proveniente de la internet podría mejorar haciéndolo autónomo para que éste pudiera buscar con libertad y de acuerdo a las necesidades planteadas, ya que el agente no tiene posibilidad de toma de decisiones que le permita evaluar y mejorar sus respuestas.

Debido a la posibilidad que brinda MAS-GEN de agregar métodos para el análisis de genes pensamos que puede dar un mejor soporte al resultado obtenido por el sistema para la clasificación de tumores y para la listas de genes obtenidas si se agregan más métodos de filtrado y aprendizaje de máquina que han surgido en los últimos años, como se aprecia en [Baena 2013, Vanitha 2015,Chen 2009, Garroa 2016].

Otro aspecto en el que puede trabajarse más adelante sería en la parte de visualización de los resultados, actualmente sólo es posible visualizar de manera gráfica los resultados obtenidos por el clustering jerárquico a través del dendograma y de la gráfica resultante del análisis de componentes principales. Enlos demás casos se obtienen listas de genes o de muestras en pantalla y en archivo. La presentación gráfica es de gran ayuda para el usuario

final, por lo que podrían utilizarsemás gráficas para presentar los resultados obtenidos que por el momento es principalmente mediante tablas.

Creemos que el sistema puede ser aplicado en otros dominios no solo para el caso que aquí presentamos. Existen un gran número de dominios en los cuales se requiere procesamiento de grandes volúmenes de datos con la finalidad de descubrir patrones de datos que no resultan sencillos por su cantidad y distribución de los datos. Dentro de los trabajos en los que podría probarse el sistema es el correspondiente a Big data o datos a gran escala. En el Big data se hace manipulación de grandes conjuntos de datos con métodos de aprendizaje de máquina.Debido a que los agentes aquí desarrollados utilizan diversos métodos de aprendizaje de máquina, en especial de clustering, los agentes podrían adaptarse para aplicar sus métodos en esos volúmenes de datos. Además gracias a la flexibilidad que se tiene de poder agregar más métodos para el análisis de los datos es posible hacer crecer la funcionalidad de los agentes de análisis de datos de manera relativamente sencilla.

Pensando en que el trabajo que realiza el sistema obtenido puede ser una herramienta de apoyo para el análisis de datos numéricos en un futuro próximo podría ponerse disponible en la web.

BIBLIOGRAFÍA

http://www.affymetrix.com/support/technical/manual/expression_manual.affx(2002).

Angeletti M., et al. (2001), An intelligent agents architecture for DNA-microarray data integration, NETTAB Workshop on Agents and Bioinformtics, Italia

Alshamlan H., Badr G., Alohali Y., A study of cancer microarray gene expres-sion profile: objectives and approaches, Proceedings of the World Congress on Engineering 2, (2013)

Baena R., Urda D., Subirats J., Franco L., Jerez J., Analysis of cancer microarray data using constructive neural networks and genetic algorithms, International Work-Conference on Bioinformatics and Biomedical Engineering, España, (2013), 55–63.

Bergeron B., Bioinformatics Computing, Prentice Hall, USA, (2002).

Berman J. (2004) Tumor classification: molecular analysis meets Aristotle. BMC Cancer.

Berumen J. (2003), Nuevos Virus del Papiloma Humano descubiertos en México: su asociación a la alta incidencia del cáncer del cervix, Gaceta Médica de México, Vol.139, No. 4, s3-s10.

Berumen J. et al, Salud en las mujeres, *V II*,.UACM, México, (2010).

Brooks R., Intelligence without reason, Proceedings of 12th Int. Joint Conf. on Artificial Intelligence, Sydney, Australia, August (1991), 569--595.

Bryson K, Luck M, Joy M and Jones D (2000) Applying agents to bioinformatics in GeneWeaver. Lect. Notes Artif. Intell, 1860, 60-71.

Chen W., Lu H., Wang M., Fang C., Gene expression data classification using artificial neural network ensembles based on samples filtering, Inter-national Conference on Artificial Intelligence and Computational Intelligence, volume 1, (2009), 626–628.

DAVID Bioinformatics Resources (2014).

Espinosa A., et al. (2013) Mitosis is a Source of Potential Markers for Screening and Survival and Therapeutic Targets in Cervical Cancer, PLOS ONE, Vol. 8, Num 2.

Erman L., et al. (1980) The hearsay II speech understanding system: integrating knowledge to resolve uncertainty, Computer surveys.

Everitt B., Dunn G.: Applied Multivariate Data Analysis, Oxford, University Press, New York (1992)
Garroa B., Rodríguez K., Vázquez R., (2016) Classification of DNA microarrays using artificial neural networks and ABC algorithm, Applied Soft Computing No. 38, 548–560

Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S (2005) Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer. http://www.bioconductor.org/.

Guojun G.,. Chaoqun M., Jianhong W., Data clustering Theory, algorithms, and applications, SIAM, USA, (2007).

http://david.abcc.ncifcrf.gov/

http://www.hsph.harvard.edu/cli/complab/dchip/

https://sites.google.com/site/dchipsoft/

http://transcriptome.ens.fr/gepas/tools.html

http://www.r-project.org/index.html, The R Project for Statistical Computing

Hall M. et al. (2009) The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Irizarry R. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostat., 4, (2003) 249-264

Jennings N. (2001), An Agent-Based Approach for Building Complex Software Systems. Communications of the ACM, Vol 44, No4.

Karasavvas K., Burger A., Baldock R. (2004) A multi-agent bioinformatics integration systems with adjustable autonomy. JouARNl of Biomedical Informatics, No. 37, 205-219. http://clipsrules.sourceforge.net/

Kohonen T. (1995) Self Organizing Maps. Springer, Berlin.

Koutkias V., Malousi A., Maglaveras N. Engineering Agent-Mediated Integration of Bioinformatics Analysis Tools, Multiagent and Grid Systems, (2007), IOS Press.

Lam H., Vazquez M., Junega B., Gene expression analysis in multi-agent environment, InteARNtional Transactions on Systems Science and Applications, Vol 1, (2006).

Laureano-Cruces, A.L., Verduga-Palencia, D.O. (2010). Simulación de un juego de futbol utilizando una arquitectura Multi-agente-Reactiva. In Libro: Desarrollo Tecnológico. (Alfa-Omega) ISBN: 978-607-707-097-9, pp. 485-493. XXIII Congreso Nacional y XI Congreso InteARNcional de Informática y Computación de la ANIEI. México, Puerto Vallarta 11-15 de octubre.

Li, C, Wong H. (2003). DNA-chip analyzer (dChip), p. 120-141, Acad. Sci. USA 98:31- 36.

Li, J. et al., A Multiagent Framework to Integrate and Visualize Gene Expression Information, IEEE ICDM Workshop on MADW & MADM, (2005)

Linde Y., Buzo A., Gray R.: An Algorithm forVector Quantizer Design. IEEE Transactions on Communications, (1980) 84-95

Lloyd. S. Least Squares Quantization in PCM. IEEE Transactions on information Theory, IT-28: (1982) 127-135

Luck M., D'Inverno M. (2001) A Conceptual Framework for Agent Definition and development, The Computer Journal, vol. 44, Num. 1.

Luck M., et al., Applying Agents to Bioinformatics in GeneWeaver**,** Lecture Notes in AI, Springer-Verlag, (2000), USA

Márquez E, Savage J, Espinosa A, Berumen J, Lemaitre C (2008) Gene expression analysis for tumor classification using vector quantization. Third IAPR InteARNtional Conference on Pattern Recognition in Bioinformatics.

Márquez E., Savage J., Espinosa A., Berumen J., LemaitreC., (2015), A decision support system based in multi-agent technology for gene expression analysis.

Mas Ana (2005), Agentes, Software y Sistemas Multiagente, conceptos, arquitecturas y aplicaciones, PearsonEducación, Madrid-España.

Merelli E., et el. , An intelligent agents architecture for ADN-microarray data integration, NETTAB (2001).

Merelli E., et el. *Agents in bioinformatic,* Knowledge Engineering Review, Vol. 20, Num. 2 117-125, Cambridge University Press, (2005).

Minsky M. (1986), The Society of Mind. New York, Simon & Schuster.

National Center for Biotechnology information (2014). http://www.ncbi.nlm.nih.gov/

Pham T., Dominik B., Hong Y. Spectral Pattern Comparison Methods for Cancer Classification Based on Microarray Gene Expression Data, IEEE transactions on circuits and systems, vol. 53, no. 11, november, (2006) 2425-2430

Platz E. (1995), Female genital tract cancer, Cancer vol. 75 No. 1, (270-294)

Quackenbush J.Computational analysis of microarray data, Nature Reviews, (2001).

Sánchez-Guerrero, L., Laureano-Cruces A.L., Mora-Torres, M, Ramírez-Rodríguez, J., Silva-López R.B. (2013). A Multi-Agent Intelligent Learning System: An Application with a Pedagogical Agent and Learning Objects. In Creative Education 2013. Vol.4, No.7A2, 181-190. Published Online July 2013 in SciRes (http://www.scirp.org/jouARNl/ce). Scientific Research.

Štiglic G., Kokol P., Using Multi-Agent System for Gene Expression Classification, Proceedings of the 26th Annual InteARNtional Conference of the IEEE EMBS, 2952-2955, (2004).

Tarraga J, Medina I, Carbonell J, et al. "GEPAS, a web-based tool for microarray data analysis and interpretation." Nucleic Acids Res. (2008)

Tinker A., Boussioutas A., Bowtell D. (2006), The challenges of gene expression microarrays for the study of human cancer, Volume 9, Issue 5, Pages 333-339

Tusher V., et al., Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA, (2001), 5116-5121

Veloso M., Stone P., Multiagent Systems: A Survey from a Machine Learning Perspective, Autonomous Robotics, USA, (2000), 2-59.

Vanitha C., Devaraj D., Venkatesulu M., (2015) Gene expression data classificationusing support vector machine and mutual information-based gene selection,Procedia Comp. Sci. 47 13–21.

Wooldridge M. (1997) Agent-based software engineering, IEEE proc. On software engineering, 144, USA.

Wooldridge M. (2002) An introduction to multiagent systems, John Wiley, USA.

Apéndice A

- Terminología básica manejada dentro del presente documento de tesis.
- Partes de la implementación del sistema de agentes para el análisis de expresión de genes(MAS-GEA).

Glosario

Adenocarcinoma. Tipo de cáncer cuyo origen son células que constituyen el revestimiento interno de las glándulas de secreción externa.

ADN. Acido desoxirribonucleico. El ADN es la molécula que contiene la información genética utilizada por una célula para la creación de proteínas. Almacena las instrucciones genéticas usadas en el desarrollo y funcionamiento de todos los organismos vivos.

Agente. Es el software o hardware que está en un medio ambiente que percibe y sobre el que actúa para cumplir con sus objetivos.

ARN. Acido ribonucleico. Acido nucléico de cadena sencilla compuesto por los nucleótidos Adenina (A), Uracilo (U), Guanina(G) y Citosina (C). En las células copia la información genética del ADN y en el citoplasma dirige la síntesis de proteínas según su secuencia de nuecleótidos.

ARNm. Acido ribonucleico mensajero. Se encarga de transferir la información genética que proviene del ADN.

CaCu. Cáncer del cuello uterino, también denominado cáncer cervical o cáncer cérvico-uterino.

Centroide. Elemento representativo de un cluster.

Clustering. Creación o formación de agrupaciones de elementos por su similitud. Conjunto de métodos aplicados a un conjunto de objetos o elementos para separarlos de acuerdo a sus semejanzas. Los elementos similares se juntan en un grupo.

Clusters. Grupos de elementos, agrupaciones encontradas para los elementos con características similares.

Dendograma. Representación gráfica tipo árbol para mostrar diferentes niveles o agrupaciones de objetos.

Downregulated gene. Gen bajo regulado o bajo expresado

Epidermoide. Tipo de cáncer cuyo origen son células del epitelio escamoso, que se encuentra en la capa superficial de la piel, en el revestimiento de la cavidad del cuerpo o en los vasos sanguíneos.

GeneChip. Es un microarreglo comercial creado por la empresa Affymetrix.

HPV16. Virus del papiloma humano tipo 16. Virus de mayor prevalencia en las muestras de mujeres diagnósticadas con Cáncer cervical.

MAS. Sistema multi-agente. Sistema donde participan 2 o más agentes para alnazar el objetivo del sistema.

MAS-GEA. Sistema multi-agente para el análisis de expresión de genes.

Medio ambiente. Medio que perciben los agentes, sobre el cual actúan y modifican los agentes.

Microarreglo. Es una matriz de datos basado en la síntesis o fijación de sondas en un sobre un sustrato sólido (vidrio o silicio), y que representan a los genes, proteinas, o metabolitos, de una muestra. Se mide un nivel de fluorescencia y por análisis de imágenes se mide la expresión con respecto a una sonde objetivo.

Oligonucleótidos. Son moléculas formadas por varios nucleótidos, se trata de secuencias cortas de ADN o ARN.

Pre-procesamiento. Métodos aplicados a los datos para adecuarlos para aplicar métodos de análisis. Incluye la normalización y estandarización de datos.

Upregulated gene. Gen sobre regulado o sobre expresado

/*Ejemplo de código para el agente de Pre-procesamiento:
Utilizando la tecnología de programación orientada a objetos,con Java, se definen los agentes, con sus características y sus comportamientos que les permiten realizar sus tareas. Dentro del código se intercalan definiciones para interactuar con las reglas de razonamiento que están hechas en lenguaje declarativo, basado en Clips, y con lo que los agentes deciden que acción realizar, a partir de lo que perciben de su medio.
*/

```
// DEFINICIÓN DE LA CLASE A PARTIR DE LA SUPER CLASE AGENT

public class PreprocessingAgent extends Agent {
  ManagerAgent magent;
  Rete reJess=new Rete();
PreprocessingAgent(ManagerAgent ma,Rete r) {
        magent=ma;
        reJess=r;
        setup();
presentation();
   }

//DEFINICION DE LA CONFIGURACION DEL AGENTE, DONDE INDICA SUS CAPACIDADES
protected void setup(){
idAgent="PreprocessingAgent";
servicio= new ArrayList();
servicio.add("ReadCel");
servicio.add("ReadTxt");
servicio.add("NormalizationRMA");
servicio.add("NormalizacionMAS");
servicio.add("FormatGenes");
```

```
servicio.add("FormatSamples");
    }

//ESTABLECIMIENTO DE COMPORTAMIENTOS DEL AGENTE

void addBehaivore(String process,List lista,ManagerAgent a){
                System.out.println("comportamiento de Preprocessing");
                magent=a;
                if(((String)lista.get(0)).compareTo("createListGenes")==0){
                  String t;
                  t=(String)lista.get(1);
                if (t.compareTo(".CEL")==0){
                        FileNav f=(FileNav)lista.get(3);
                          ListaFilesAgent test = new ListaFilesAgent
(((List)lista.get(2)),f.getDir(),f.getSelectedD(),magent.getTipoNorm());
                                //test.setDefaultCloseOperation(JFrame.EXIT_ON_CLOSE);
                          }
                        else
                        if (t.compareTo(".TXT")==0)
                                genera();}}}


---------------------------------------
/*  Clase Abstracta Agent es la base para crear los agentes*/
//SUPER CLASE AGENT

abstract class Agent{
   String idAgent;
abstract void setup();
abstract void presentation();
  // abstract void addBehaivore();
abstract void readM(Message msg, String nameA);
abstract void sendM(List msg, String nameA);
```

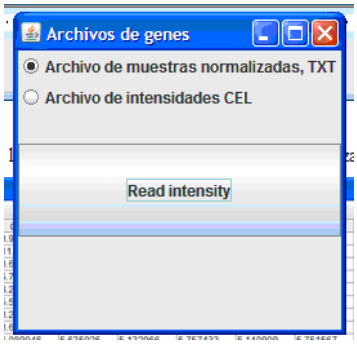
```
   List servicio;
   String getIdAgent(){return idAgent;          }
List getServicio(){  return servicio;}}
```

/*EJEMPLO DE CÓDIGO EN LENGUAJE DELARATIVO PARA LA TOMA DE DECISIONES DE LOS AGENTES UTILIZANDO REGLAS DE PRODUCCIÓN. HECHO CON CLIPS */

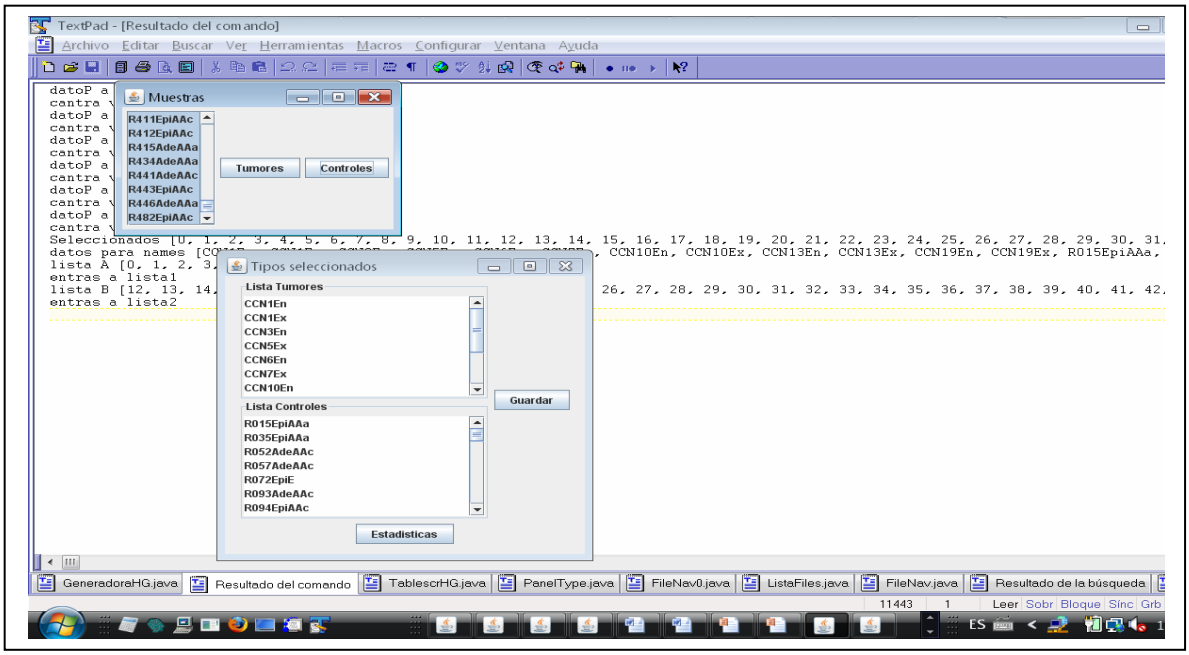
```
(defrule filtraGenes
?f<-(solicitud (tipo identGenes) (estado inactivo) (subTipo filtrado))
(readFiles yes)
   (parameters  (pval ?pv) (fc1 ?fc) )
=>
  (modify ?f (estado ejecuta))
  (comando filtrado ?n ?pv ?fc ?d)
)

(defrule clasificaMuestras
?f<-(solicitud (tipo classSamples) (estado inactivo) (subTipo listaFiltrado))
=>
  (modify ?f (estado ejecuta))
  (comando clustering VQ SOM PCA)
)

(defrule clasificaMuestras2
?f<-(solicitud (tipo classSamples) (estado inactivo) (subTipo listaGenes))
  (readFiles yes)
=>
  (modify ?f (estado ejecuta))
(comando clustering ? ? ?)
)


(defrule parametros
  ?f<-(read parametros Identification)
   (parameters  (fc1 ?f1) (fc2 ?f2) (tfc ?tfc) (tpv ?tpv) (upG ?up) (dnG ?dn) (pval ?pv1) (delta
?d1) (normal ?n1))
=>
  (bind  ?*fc1* ?f1) (bind  ?*fc2* ?f2) (bind  ?*tfc* ?tfc) (bind  ?*tpv* ?tpv) (bind  ?*upG* ?up)
(bind  ?*dnG* ?dn) (bind  ?*pval* ?pv1) (bind  ?*delta* ?d1) (bind  ?*normal* ?n1)
(retract ?f)
    (printout t "para normalizar " ?n1 crlf)
```

En seguida pueden verse algunas pantallas del sistema en funcionamiento, la interfaz está hecha con Java y a través de ella el usuario solicita al sistema el análisis de expresión de genes para sacar los genes diferenciados o para clasificación de muestras.

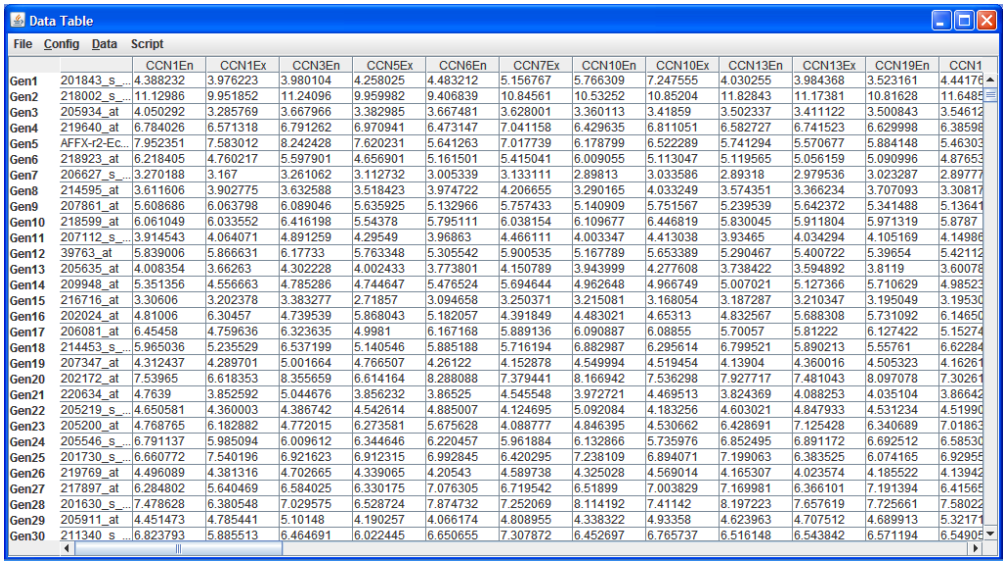
A Pantalla para selección de tipo de archivo origen de datos:



B Pantallas presentadas al Usuario:

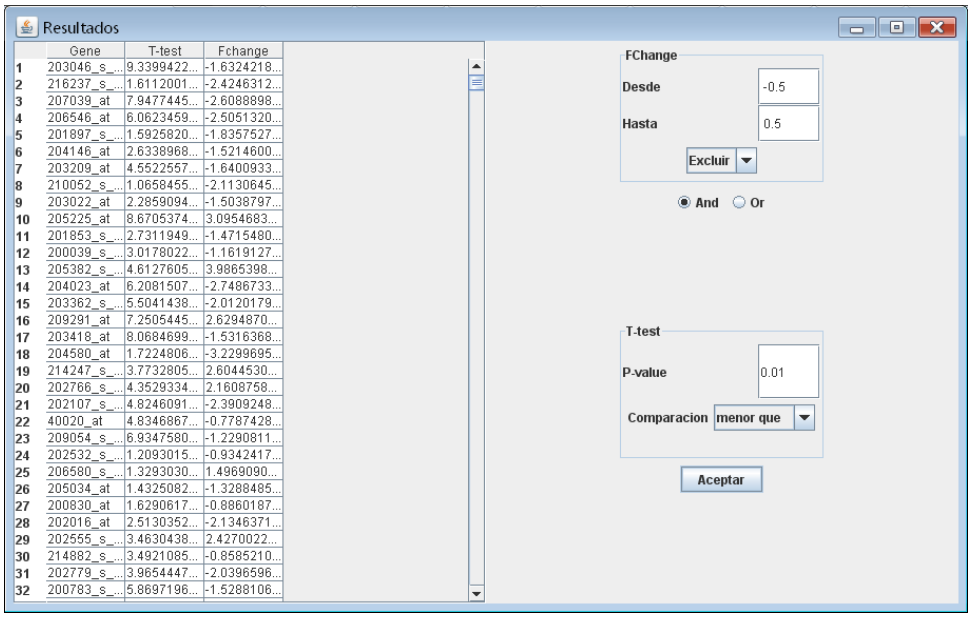


C Pantalla de datos leídos desde los microarreglos de expresión y ya normalizados:



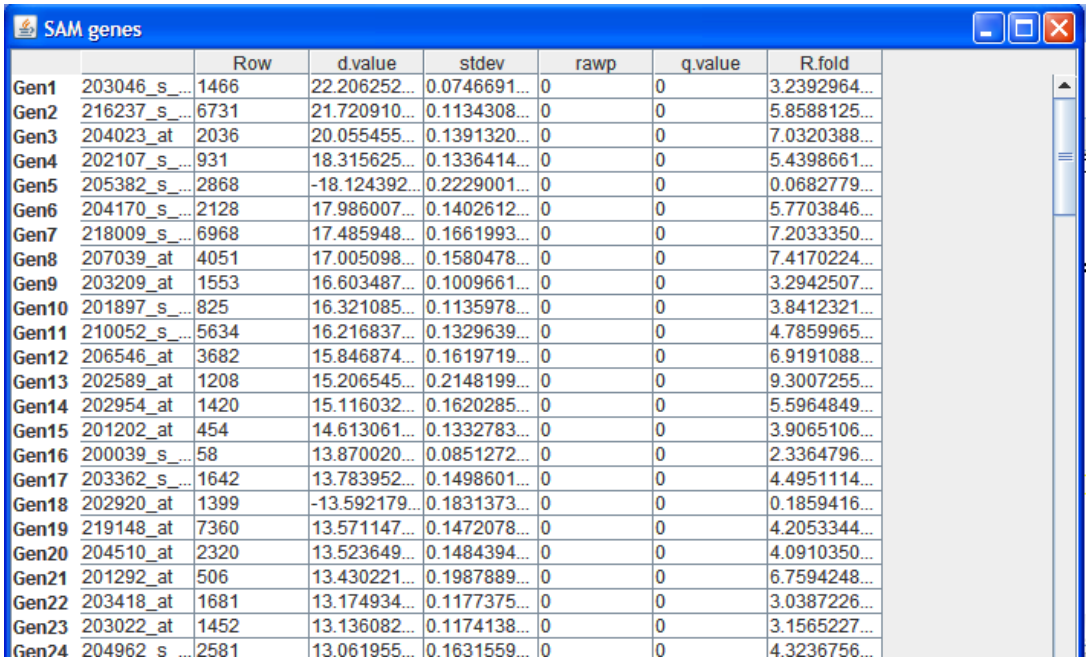
		CCN1En	CCN1Ex	CCN3En	CCN5Ex	CCN6En	CCN7Ex	CCN10En	CCN10Ex	CCN13En	CCN13Ex	CCN19En	CCN1
Gen1	201843_s_...	4.388232	3.976223	3.980104	4.258025	4.483212	5.156767	5.766309	7.247555	4.030255	3.984368	3.523161	4.44176
Gen2	218002_s_...	11.12986	9.951852	11.24096	9.959982	9.406839	10.84561	10.53252	10.85204	11.82843	11.17381	10.81628	11.6485
Gen3	205934_at	4.050292	3.285769	3.667966	3.382985	3.667481	3.628001	3.360113	3.41859	3.502337	3.411122	3.500843	3.54612
Gen4	219640_at	6.784026	6.571318	6.791262	6.970941	6.473147	7.041158	6.429635	6.811051	6.582727	6.741523	6.629998	6.38598
Gen5	AFFX-r2-Ec...	7.952351	7.583012	8.242428	7.620231	5.641263	7.017739	6.178799	6.522289	5.741294	5.570677	5.884148	5.46303
Gen6	218923_at	6.218405	4.760217	5.597901	4.656901	5.161501	5.415041	6.009055	5.113047	5.119565	5.056159	5.090996	4.87653
Gen7	206627_s_...	3.270188	3.167	3.261062	3.112732	3.005339	3.133111	2.89813	3.033586	2.89318	2.979536	3.023287	2.89777
Gen8	214595_at	3.611606	3.902775	3.632588	3.518423	3.974722	4.206655	3.290165	4.033249	3.574351	3.366234	3.707093	3.30817
Gen9	207861_at	5.608686	6.063798	6.089046	5.635925	5.132966	5.757433	5.140909	5.751567	5.239539	5.642372	5.341488	5.13641
Gen10	218599_at	6.061049	6.033552	6.416198	5.54378	5.795111	6.038154	6.109677	6.446819	5.830045	5.911804	5.971319	5.8787
Gen11	207112_s_...	3.914543	4.064071	4.891259	4.29549	3.96863	4.466111	4.003347	4.413038	3.93465	4.034294	4.105169	4.14986
Gen12	39763_at	5.839006	5.866631	6.17733	5.763348	5.305542	5.900535	5.167789	5.653389	5.290467	5.400722	5.39654	5.42112
Gen13	205635_at	4.008354	3.66263	4.302228	4.002433	3.773801	4.150789	3.943999	4.277608	3.738422	3.594892	3.8119	3.60078
Gen14	209948_at	5.351356	4.556663	4.785286	4.744647	5.476524	5.694644	4.962648	4.966749	5.007021	5.127366	5.710629	4.98523
Gen15	216716_at	3.30606	3.202378	3.383277	2.71857	3.094658	3.250371	3.215081	3.168054	3.187287	3.210347	3.195049	3.19530
Gen16	202024_at	4.81006	6.30457	4.739539	5.868043	5.182057	4.391849	4.483021	4.65313	4.832567	5.688308	5.731092	6.14650
Gen17	206081_at	6.45458	4.759636	6.323635	4.9981	6.167168	5.889136	6.090887	6.08855	5.70057	5.81222	6.127422	5.15274
Gen18	214453_s_...	5.965036	5.235529	6.537199	5.140546	5.885188	5.716194	6.882987	6.295614	6.799521	5.890213	5.55761	6.62284
Gen19	207347_at	4.312437	4.289701	5.001664	4.766507	4.26122	4.152878	4.549994	4.519454	4.13904	4.360016	4.505323	4.16261
Gen20	202172_at	7.53965	6.618353	8.355659	6.614164	8.288088	7.379441	8.166942	7.536298	7.927717	7.481043	8.097078	7.30261
Gen21	220634_at	4.7639	3.852592	5.044676	3.856232	3.86525	4.545548	3.972721	4.469513	3.824369	4.088253	4.035104	3.86642
Gen22	205219_s_...	4.650581	4.360003	4.386742	4.542614	4.885007	4.124695	5.092084	4.183256	4.603021	4.847933	4.531234	4.51990
Gen23	205200_at	4.768765	6.182882	4.772015	6.273581	5.675628	4.088777	4.846395	4.530662	6.428691	7.125428	6.340689	7.01863
Gen24	205546_s_...	6.791137	5.985094	6.009612	6.344646	6.220457	5.961884	6.132866	5.735976	6.852495	6.891172	6.692512	6.58530
Gen25	201730_s_...	6.660772	7.540196	6.921623	6.912315	6.992845	6.420295	7.238109	6.894071	7.199063	6.383525	6.074165	6.92955
Gen26	219769_at	4.496089	4.381316	4.702665	4.339065	4.20543	4.589738	4.325028	4.569014	4.165307	4.023574	4.185522	4.13942
Gen27	217897_at	6.284802	5.640469	6.584025	6.330175	7.076305	6.719542	6.51899	7.003829	7.169981	6.366101	7.191394	6.41565
Gen28	201630_s_...	7.478628	6.380548	7.029575	6.528724	7.874732	7.252069	8.114192	7.41142	8.197223	7.657619	7.725661	7.58022
Gen29	205911_at	4.451473	4.785441	5.10148	4.190257	4.066174	4.808955	4.338322	4.93358	4.623963	4.707512	4.689913	5.32171
Gen30	211340_s_...	6.823793	5.885513	6.464691	6.022445	6.650655	7.307872	6.452697	6.765737	6.516148	6.543842	6.571194	6.54905

D. Filtrado de resultados de pruebas estadísticas aplicadas a los datos



	Gene	T-test	Fchange
1	203046_s_...	9.3399422...	-1.6324218...
2	216237_s_...	1.6112001...	-2.4246312...
3	207039_at	7.9477445...	-2.6088898...
4	206546_at	6.0623459...	-2.5051320...
5	201897_s_...	1.5925820...	-1.8357527...
6	204146_at	2.6338968...	-1.5214600...
7	203209_at	4.5522557...	-1.6400933...
8	210052_s_...	1.0658455...	-2.1130645...
9	203022_at	2.2859094...	-1.5038797...
10	205225_at	8.6705374...	3.0954683...
11	201853_s_...	2.7311949...	-1.4715480...
12	200039_s_...	3.0178022...	-1.1619127...
13	205382_s_...	4.6127605...	3.9865398...
14	204023_at	6.2081507...	-2.7486733...
15	203362_s_...	5.5041438...	-2.0120179...
16	209291_at	7.2505445...	2.6294870...
17	203418_at	8.0684699...	-1.5316368...
18	204580_at	1.7224806...	-3.2299695...
19	214247_s_...	3.7732805...	2.6044530...
20	202766_s_...	4.3529334...	2.1608758...
21	202107_s_...	4.8246091...	-2.3909248...
22	40020_at	4.8346867...	-0.7787428...
23	209054_s_...	6.9347580...	-1.2290811...
24	202532_s_...	1.2093015...	-0.9342417...
25	206580_s_...	1.3293030...	1.4969090...
26	205034_at	1.4325082...	-1.3288485...
27	200830_at	1.6290617...	-0.8860187...
28	202016_at	2.5130352...	-2.1346371...
29	202555_s_...	3.4630438...	2.4270022...
30	214882_s_...	3.4921085...	-0.8585210...
31	202779_s_...	3.9654447...	-2.0396596...
32	200783_s_...	5.8697196...	-1.5288106...

E Pantalla que presenta genes seleccionados



		Row	d.value	stdev	rawp	q.value	R.fold
Gen1	203046_s_...	1466	22.206252...	0.0746691...	0	0	3.2392964...
Gen2	216237_s_...	6731	21.720910...	0.1134308...	0	0	5.8588125...
Gen3	204023_at	2036	20.055455...	0.1391320...	0	0	7.0320388...
Gen4	202107_s_...	931	18.315625...	0.1336414...	0	0	5.4398661...
Gen5	205382_s_...	2868	-18.124392...	0.2229001...	0	0	0.0682779...
Gen6	204170_s_...	2128	17.986007...	0.1402612...	0	0	5.7703846...
Gen7	218009_s_...	6968	17.485948...	0.1661993...	0	0	7.2033350...
Gen8	207039_at	4051	17.005098...	0.1580478...	0	0	7.4170224...
Gen9	203209_at	1553	16.603487...	0.1009661...	0	0	3.2942507...
Gen10	201897_s_...	825	16.321085...	0.1135978...	0	0	3.8412321...
Gen11	210052_s_...	5634	16.216837...	0.1329639...	0	0	4.7859965...
Gen12	206546_at	3682	15.846874...	0.1619719...	0	0	6.9191088...
Gen13	202589_at	1208	15.206545...	0.2148199...	0	0	9.3007255...
Gen14	202954_at	1420	15.116032...	0.1620285...	0	0	5.5964849...
Gen15	201202_at	454	14.613061...	0.1332783...	0	0	3.9065106...
Gen16	200039_s_...	58	13.870020...	0.0851272...	0	0	2.3364796...
Gen17	203362_s_...	1642	13.783952...	0.1498601...	0	0	4.4951114...
Gen18	202920_at	1399	-13.592179...	0.1831373...	0	0	0.1859416...
Gen19	219148_at	7360	13.571147...	0.1472078...	0	0	4.2053344...
Gen20	204510_at	2320	13.523649...	0.1484394...	0	0	4.0910350...
Gen21	201292_at	506	13.430221...	0.1987889...	0	0	6.7594248...
Gen22	203418_at	1681	13.174934...	0.1177375...	0	0	3.0387226...
Gen23	203022_at	1452	13.136082...	0.1174138...	0	0	3.1565227...
Gen24	204962_s_...	2581	13.061955...	0.1631559...	0	0	4.3236756...

Apéndice B

Artículo publicado como resultado del trabajo doctoral en Technology for Gene Expression Analysis. InteARNtional JouARNl of Intelligence Science,.