GOBIERNO DEL
ESTADO DE MÉXICO

TESTT
TIANGUISTENCO

ISC
INGENIERÍA EN
SISTEMAS COMPUTACIONALES

TECNOLÓGICO DE ESTUDIOS SUPERIORES DE TIANGUISTENCO

DIVISIÓN DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

TESIS PARA OBTENER EL GRADO DE:
INGENIERO EN SISTEMAS COMPUTACIONALES

"Desarrollo de una herramienta de visualización Integral de datos omicos"

PRESENTA:

JUAN JOSÉ MARTÍNEZ ULLOA

ASESOR:

DR. JESÚS ESPINAL ENRÍQUEZ
MTRA. KAROL BACA LOPEZ

TIANGUISTENCO ESTADO DE MÉXICO, MARZO 2018

Resumen.

Abstract.

Agradecimientos.

Índice

1. Introducción **2**
1.1. Planteamiento del Problema 2
1.2. Justificación 3
1.3. Delimitación 3
1.4. Hipótesis 3
1.5. Objetivos 3
1.5.1. Objetivo General 3
1.5.2. Objetivos Específicos 4
1.6. Aportaciones de la tesis 4

2. Estado del Arte **4**

3. Tecnologías genómica. **4**

4. Microarreglos **4**
4.1. Formato 6

1. Introducción

1.1. Planteamiento del Problema

El cáncer de mama es una enfermedad compleja y heterogénea con más de 1,300,000 casos y 450,000 muertes cada año en todo el mundo. Esta enfermedad se caracteriza por diferentes ápectos biológicos como desregulazación de la expresión génica, alteraciones genómicas del ADN, etc. Todo esto da lugar al inicio y desarrollo del carcinoma de mama. En éstos últimos años el uso de datos ómicos, como los basados en microarreglos (microarrays) y secuenciación, esta en su pleno () en el campo de la biomedicina. Todos estos datos permiten estudiar enfermedades desde un punto de vista biomolecular. Con esto, se ofrecen grandes oportunidades para mejorar tanto la comprención de la enfermedad, como el desarrollo de nuevos métodos para el diagnóstico y tratamiento del paciente, sin embargo, el análisis de estos datos producidos por estas tecnologías, es bastante complejo por lo que es necesario la aplicación de avanzadas técnicas de análisis y cálculos computacionales que permiten obtener la información biológica disponibles. Hasta el día de hoy todos estos datos óptenidos de diversos experimentos, datos de muy alta calidad, datos clinicamente bien anotados, y una gran cantidad de datos de canceres analizados con el fin de encontrar anomalias recurrentes que sean importantes de la enfermedad y estos datos se guardan en diversas plataformas que permiten tener uns gran cantidad de información , pero estos datos no son tan faciles de analizar, por ello, la bioinformática ayuda a manejar estructurar y organizarla para que sea mas fácil de comprender. Circos plot es una de las herramientas que existen para la visualización datos, ideal para explorar las relaciones entre objetos y posiciones. Esta herramienta es flexible, aunque originalmente fue

diseñado para visualizar datos genómicos , se puede crear figúras a partir de datos en cualquier campo, desde la genómica hasta la visualización de la migración al arte matemático. Esta herramienta puede ser automatizada. Está controlado por archivos de configuración de texto plano, lo que lo hace díficil su incorporación en *pipeline* de adquisición de datos, análisis e informes. Todo esto hace dificil su utilización en instalacion y desarrollo de la visualización.

1.2. Justificación

La cantidad de información generada a través de las tecnologías de secuenciación masiva son enormes. La integración de toda esa información puede ayudar a tomar decisiones de índole biomédica e incluso clínica. Aunque existen plataformas ya conocidas capaces de presentar esa información de un modo amigable, la programación de dichas herramientas continúa siendo complicada para un usuario final. Es por esto, que generar una herramienta con gran capacidad de visualización, integración de información y facilidad de implementación para un usuario final es de la mayor importancia.

Con el fin de mejorar el rendimiento y la cantidad de tiempo de los investigadores del Instituto Nacional de Medicina Genómica, es fundamental sistematizar este software para reducir el tiempo de programación y la investigación. La sistematización de circos plot, brindará la posibilidad de que el investigador ahorre en tiempo de programación o en dado caso que no se conozca nada del uso nativo de circos plot, leer todo el manual de uso de dicho software, para que ocupe su mayor de tiempo en la investigación y solo tome varios minutos para diseñar su grafica circular llamada circos plot.

1.3. Delimitación

El desarrollo de tecnologías de secuenciación masiva es abrumador. Para poder integrar toda la información proveniente de dichas tecnologías es necesario contar con herramientas adecuadas para la obtención, análisis y visualización de las tecnologías anteriormente mencionadas.

Ante esta problemática, resulta altamente relevante generar una herramienta capaz de condensar en un solo golpe de vista varias capas de información, para que, de este modo, se pueda analizar con un mayor detalle los datos obtenidos.

Circos plot es una herramienta de apoyo visual que permite lograr la visualización a nivel de genoma completo varios tipos diferentes de datos: Expresión, mutaciones, metilación, citobandas, etc., de una manera amigable para el usuario final.

Aunque circos plot tiene grandes ventajas en cuanto a la visualización, su programación es altamente complicada, por lo que generar un back-end que integre de modo sencillo varios tipos de datos, será de la mayor utilidad.

1.4. Hipótesis

El desarrollo de una herramienta de Visualización integral permitirá el análisis de la información de los datos derivados de múltiples plataformas.

1.5. Objetivos

1.5.1. Objetivo General

Implemetar una herramienta de visualización que integre datos genómicos derivados de múltiples plataformas.

1.5.2. Objetivos Específicos

- Realizar el pretatamiento de los datos de trascriptoma (Microarreglos).
- Realizar el pretatamiento de los datos de Metilación.
- Desarrollar algoritmo computacional para hacer la búsqueda en la referencia del genoma humano con los datos de trascriptoma y metilación.
- Desarrollar software para preprocesar datos de distintos tipos de tecnologías genómicas.
- Desarrollar software para integrar al backbone de circos plot varias capas de información proveniente del paso anterior.
- Generar opciones de visualización de fácil manejo para el usuario final.

1.6. Aportaciones de la tesis

La presente Tesis aportara

2. Estado del Arte

3. Tecnologías genómica.

Las tecnologías genómicas es el conjunto de herramientas orientadas al estudio integral del funcionamiento, contenido, evolución del genoma. Es una de las áreas más vanguardistas de la biología. La genómica usa conocimientos derivados de distintas ciencias como la biología molecular, la bioquímica, la informática, la estadística, las matemáticas y la física. Para entender un poco más de estas tecnologias y de los datos que se obtiene de las antes mencionadas, hablaremos de las tecnologias genómicas que son: Microarreglos y Metilación.

4. Microarreglos

Un microarreglo de ADN (ácido desoxirribonucleico) tambien llamado DNA chip, tambien llamado **oligonucleotido de DNA chip o gene chip** que consiste en pequeños fragmentos de ADN de los cuales representa un gen diferente [1], es una superficie sólida a la cual se une una colección de fragmentos de ADN. Las superficies empleadas para fijar el ADN son muy variables y pueden ser de vidrio, plástico e incluso de silicona. Los chips de ADN se usan para analizar la expresión diferencial de genes. Su funcionamiento consiste, básicamente, en medir el nivel de hibridación entre la sonda específica (probe, en inglés), y la molécula diana (target), y se indican generalmente mediante fluorescencia y a través de un análisis de imagen, lo cual indica el nivel de expresión del gen.
Por lo tanto, los microarreglos son una potente fuente de obtencion de perfiles de expresion de genes sometidos a diferente condiciones. Identificar los patrones de los niveles de expresion sera muy util para compararlos y poder estudiar las respuestas de los genes.
Aplicando una serie de procesos experimentales y computacionales sobre los microarreglos se obtiene una una matriz numerica bidimensional que consta de los genes de poblaciones distintas como individuos y de las condiciones experimentales a las que expusieron as celulas como variables en el caso que

se quiera estudiar a los genes, o a la inversa, si es que se quiere realizar un estudio comparativo de las condiciones a que se somete. Cada uno de los valores de la matriz representa el nivel de expresion de un determinado gen bajo una cierta condicion experimental.
Estas matrices son de grandes dimensiones puesto que existen una gran cantidad de condiciones experimentales y genes. En la figura 1.1 se puede observar el modelo de la presentacion de los microarreglos . Cada fila representa un gen, el cual debe de ser indicado, cada columna represnta una condición experimental, cuyo nombre tambien debe de ser identicado. Los valores de la matriz son los niveles de expresión de los genes para la condicion experimental.

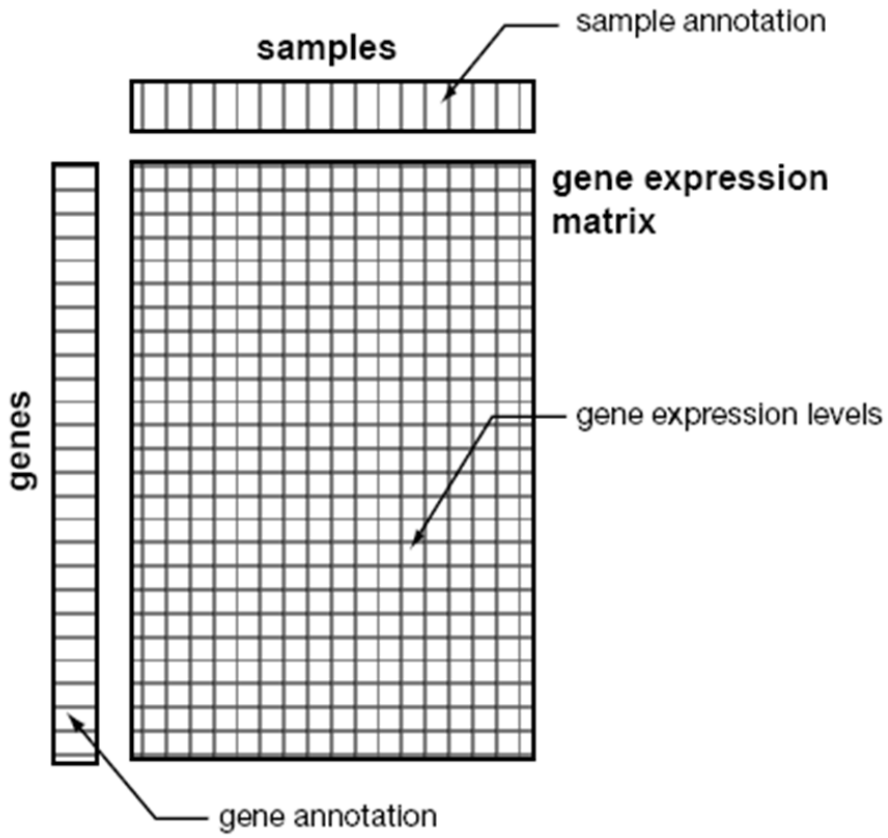

Figura 1: Chip de un Microarreglo

Dado que realizar un analisis de estas matrices de grandes dimensiones es una tarea practicamente imposible, se hace necesarias tecnicas computacionales que permitan analizar todos estos datos y entonces realizar el analisis biologico.
Actualmente existen diferentes bases de datos a nuestro alcance a traves de Internet que unifican y facilitan toda esta informacion genetica ademas de ofrecer diversas herramientas para el analisis de esta gran cantidad de informacion. Algunas de estas bases de datos por ejemplo son las que hay en el EMBL (European Molecular Biology Laboratory), el SIB (Swiss Institute of Bioinformatics), el EBI (European Bioinformatics Institute) o el NCBI (National Center for Biotechnology Information). El EBI y el NCBI son los que mas informacion contienen y por lo tanto los mas utilizados. El tamño de los microarreglos es de 1.28 cm x 1.28 cm, hay 500,000 ubicaciones en cada matriz y por lo general tiene millones de cadenas de ADN construidas en cada ubicación, cada cadena contiene 25 pares bases (Figura 1).

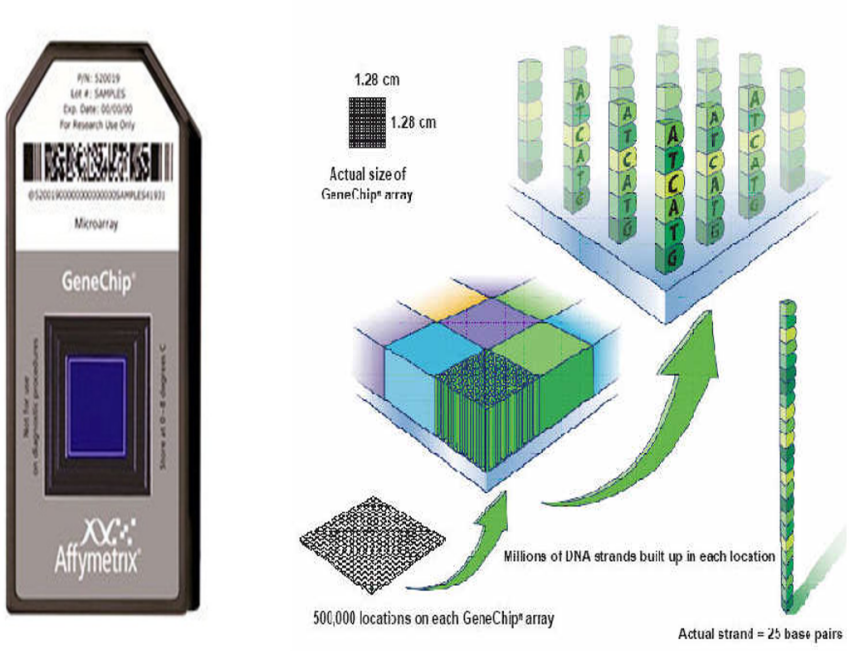

Figura 2: Chip de un Microarreglo.

En estos chips se imprimen las secuencias biologicas en un chip, de manera que se puede cuantificar el transcripción en una matriz numérica.

4.1. Formato

Los archivos están disponibles en un formato de valores separados por comas (CSV). Estos son archivos de texto sin formato con cada fila terminada por un carácter de nueva línea. Los datos en campos separados están entre comillas y separados por comas. Ninguno de los campos de datos contiene ninguno de estos caracteres: comillas, nueva línea, retorno de carro o tabulación.

Estos archivos se usen principalmente en aplicaciones de hojas de cálculo y programas de bases de datos (como bases de datos SQL). Los datos estan formateados de tal manera que estos dos usos sean relativamente fáciles. Se tiene en cuenta que algunos de los archivos y los campos de datos en ellos son grandes.

La primera fila de cada archivo contiene los títulos de los campos que figuran en las filas siguientes.

Cada fila después de la primera fila contiene anotaciones para un solo conjunto de sondas. Todas las anotaciones para ese conjunto de sonda están contenidas en esa única fila. En algunos campos, como las anotaciones de dominio de proteínas, puede haber más de una anotación para un único conjunto de sondas. En este caso, los valores múltiples están separados por la cadena ’///’.

En muchos tipos de anotaciones, los subcampos están separados por ’//’. Por ejemplo, una anotación para un "GO Biological Process"puede aparecer como "7155 // cell adhesion // predicted / computed". En este caso, las secciones corresponden a ÏD // Descripción // Evidencia", pero el significado de los subcampos varía entre los diferentes tipos de anotación, como se describe a continuación.

Los campos vacíos se indican con ’- - -’ . El hecho de utilizar una cadena de este tipo en lugar de dejar el campo vacío es que hace que la naturaleza columnar de los datos sea más visible en ciertos programas de hoja de cálculo. Algunas columnas en algunos archivos no contienen datos. Para ayudar a los usuarios a combinar datos de varios archivos, dichas columnas vacías no se eliminan. Por lo tanto, cada archivo tiene las mismas columnas en el mismo orden.

Algunos campos, como Çhip", contienen el mismo valor para cada conjunto de sonda en un archivo. Aunque estos datos son redundantes en cualquier archivo individual, son útiles para los usuarios que combinan datos de varios archivos.

Referencias

[1] Microarreglos de ADN y sus aplicaciones en investigaciones biomédicas. Revista CENIC. Ciencias Biológicas, 38 (2), 132-135.