EVALUACION Y APLICACIÓN DE HERRAMIENTAS COMPUTACIONALES DE MINERIA DE TEXTO PARA LA IDENTIFICACION Y CARACTERIZACION DE INTERACCIONES MOLECULARES RELACIONADAS CON LA PROTEINA EVI1 EN LA LITERATURA CIENTIFICA PARA EL MODELAMIENTO DE UNA RED BAYESIANA

Daniel González Olano

Raquel Silva Hernández

UNIVERSIDAD INDUSTRIAL DE SANTANDER

ESCUELA DE INGENÍERA DE SISTEMAS E INFORMÁTICA

FACULTAD DE INGENIERÍAS FÍSICO MECÁNICAS

BUCARAMANGA

2011

EVALUACION Y APLICACIÓN DE HERRAMIENTAS COMPUTACIONALES DE MINERIA DE TEXTO PARA LA IDENTIFICACION Y CARACTERIZACION DE INTERACCIONES MOLECULARES RELACIONADAS CON LA PROTEINA EVI1 EN LA LITERATURA CIENTIFICA PARA EL MODELAMIENTO DE UNA RED BAYESIANA

Daniel González Olano

Raquel Silva Hernández

Trabajo de grado para optar el titulo

Ingeniero de sistemas

Director

MPe. Henry Arguello Fuentes

Codirector

Dr. Herman José Arteaga Narváez

UNIVERSIDAD INDUSTRIAL DE SANTANDER

ESCUELA DE INGENÍERA DE SISTEMAS E INFORMÁTICA

FACULTAD DE INGENIERÍAS FÍSICO MECÁNICAS

BUCARAMANGA

2011

A Dios, quien es el principal artífice de este logro, quien ha guiado mis pasos y me ha dado la fe, la fortaleza para seguir luchando y ha sido mi conforto en momentos difíciles.

A mi madre Myriam Elissa a quien debo mucho por sus enseñanzas, por haber velado por mi salud, mi educación, por su amor, sacrificio, paciencia y cariño, por inculcarme valores y formarme con buenos sentimientos.

A mis hermanas Melissa , Adriana y Andrea quienes me han brindado su cariño y apoyo y con quienes he compartido tristezas y alegrías, y me han acompañado en el transcurrir de mi vida.

A mi novia Jenny Sofía, quien me ha brindado su amor, cariño, confianza, apoyo constante y comprensión, gracias a ella por su paciente espera para que pudiera terminar mi carrera, a ella debo el hecho de sonreír con sólo evocarla y de llevar mi alma viajando entre rieles de magia.

A los que nunca dudaron que alcanzaría esta meta: mi tía Mireya, Laura Bent Djafar, María Catalina, Damiano, Diana y Abdú Balah.

A mi sobrinito Cristian David quien no tiene ni idea de lo que se trata mi tesis y tampoco lo que significa para nosotros desde que llego a nuestro hogar llenándolo de ternura y de felicidad.

A la Universidad por abrirme sus puertas y todos los amigos que no siendo anonimos me han acompañado en este largo camino.

Daniel Gerónimo

Dedico este libro a Dios que es el motor de cada una de las cosas que realizo en mi vida.

A mis padres Benito Silva y Angela Hernandez por su fe en mi, su continuo apoyo, sus plegarias y su amor incondicional.

A mis hermanas Helia, Edilia y Mónica por su gran apoyo y confianza.

A mis sobrinos Cristian, Sebastian y Pedro que son la inspiración para salir adelante.

A mi novio Alexander que me apoyó y me acompañó a lo largo de este camino gracias amor.

A toda mi familia…..

A mis amigos que con su ayuda, comprensión han hecho de mi carrera una realidad.

Raquel Silva Hernandez

AGRADECIMIENTOS

Agradecemos la realización de este proyecto a:

M.S.C. Henry Arguello, – por su dirección en la tesis.

Dr. H.Jose Arteaga director del grupo de Inmunología y Epidemiologia de la universidad industrial de santander, – por su valiosa colaboración durante todo el desarrollo de la tesis como codirector del proyecto.

Esp. Jenny Sofía Gómez, por su apoyo y colaboración en cada etapa del desarrollo del proyecto.

Fabian Cardozo, por aportar ideas y sugerir soluciones a problemas que se presentaron.

Al grupo de Inmunologia y Epidemiologia Molecular integrado por: Maria Fernanda Silva, Eddy Betancourt, Vladimir Pabon, Lenis Alvarez, David Colon, Juan Guillermo, Andrea Rocio Baron, Yesid Estupiñan, quienes con sus aportes hicieron posible este logro.

Andrea Reyes, Dario Delgado, estudiantes de la UIS por su amistad y colaboración.

Y a todos los que de alguna u otra manera contribuyeron al desarrollo de esta tesis de grado.

CONTENIDO

TABLA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

RESUMEN

Título: EVALUACION Y APLICACIÓN DE HERRAMIENTAS COMPUTACIONALES DE MINERIA DE TEXTO PARA LA IDENTIFICACION Y CARACTERIZACION DE INTERACCIONES MOLECULARES RELACIONADAS CON LA PROTEINA EVI1 EN LA LITERATURA CIENTIFICA PARA EL MODELAMIENTO DE UNA RED BAYESIANA *

Autores: Daniel Gonzalez Olano, Raquel Silva Hernandez **

Palabras claves: Minería de Texto, EVI-1, Redes Bayesianas, Extracción de Interacciones.

Descripción:

El desarrollo de esta tesis de grado buscó apoyar al investigador de la salud facilitándole el reconocimiento de entidades en los textos, la predicción de procesos biológicos, formulación de nuevas hipótesis, la extracción y representación del conocimiento con el fin de conocer y comprender las interacciones entre proteínas, para intentar crear modelos comprensibles a partir de información que se encuentra de manera no estructurada en millones de artículos científicos.

En primer lugar se destacó la importancia de la minería de texto como herramienta que proporciona la comprensión y extracción de conocimiento de una gran red de interacciones de alta complejidad para el manejo humano. Posteriormente se implementaron herramientas de minería de texto para recopilar información en archivos de texto plano de las proteínas relacionadas con la proteína EVI-1, luego se evaluaron las herramientas comparándolas con un patrón proporcionado por el experto. Mediante un algoritmo aplicado a los archivos, se filtraron datos de interacciones para obtener una matriz de adyacencia que permitió crear la estructura de la red Bayesiana especificando las conexiones entre las proteínas.

Este estudio permite hacer enfasis en la existencia de genes y proteínas que contribuyen a originar un cáncer, los cuales, en sus "versiones normales" regulan el crecimiento y la vida celular. Llegando a creer que muchos tumores son el resultado de una alteración mutagénica no reparada en el ADN situación que deriva en el cáncer.

Este acercamiento con el área de la biología, que bien puede llamarse biología de sistemas, permitió unir campos que han trabajado separados, para reducir el coste en los procesos de investigación del Instituto de medicina molecular de la UIS, no sólo en lo relacionado con las interacciones entre proteínas sino también en otros trabajos de investigación.

---

* Trabajo de Investigación

** Facultad de Ingenierías Físico Mecánicas, Escuela de Ingeniería de Sistemas e Informática. Director: Henry Arguello Fuentes, Codirector: H. José Arteaga.

ABSTRACT

Title: EVALUATION AND APPLICATION OF COMPUTER TOOLS OF TEXT MINING FOR THE IDENTIFICATION AND CHARACTERIZATION OF MOLECULAR INTERACTIONS RELATED TO THE PROTEIN EVI1 IN THE SCIENTIFIC LITERATURE FOR BAYESIAN NETWORK MODELING*

Authors: Daniel Gerónimo González Olano, Raquel Silva Hernández **

Key Words: Text Mining, EVI-1, Bayesian Networks, Extraction Interactions.

Description:

Development of this thesis sought to support the health researcher, facilitating the recognition of entities in the texts, the prediction of biological processes, formulation of new hypotheses, extraction and representation of knowledge with the purpose of know and understand the interactions between proteins, to try to create comprehensible models from information found in an unstructured way in millions of scientific articles.

First, highlighted the importance of text mining as a tool to provide understanding and knowledge extraction from a large network of highly complex interactions to human management. Subsequently we are implemented text mining tools to gather information in plain text files the protein related with protein EVI-1, and then the tools are evaluated comparing them with standard provided by the expert. Using an algorithm applied to the files are filtered interaction data to obtain an adjacency matrix which created the structure of Bayesian network specifying the connections between the proteins.

This study allows to make emphasis in the existence of genes and proteins that contribute to originate a cancer, which, in his " versions normales" they regulate the growth and the cellular life. Getting to think that many tumors are the result of a mutagenic alteration nonsudden shying in the DNA situation that derives in the cancer.

This approach with the area of the Biology, that can well be called Biology of systems, allowed to unite fields that have worked separated, to reduce the cost in the processes of investigation of the molecular medicine Institute of the UIS, not only in the related thing to the interactions between proteins but also in other works of investigation.

---

* Researching Work

** Physical Mechanics Engineering Faculty. Informatics and Systems Engineering School. Director: Henry Arguello Fuentes, Co-director: H. José Arteaga.

INTRODUCCIÓN

Existen grandes repositorios que contienen millones de artículos que poseen datos de interés para los científicos pero la mayoría de las veces no son visibles y tampoco se encuentran de manera estructurada. Los recientes avances en biología molecular han originado un aumento en el volumen en la información relacionada con proteínas, lo cual sugiere un mayor esfuerzo computacional en su manejo. Muchos tipos de datos se encuentran almacenados en forma organizada pero son insuficientes para determinadas investigaciones como las relacionadas con las interacciones entre proteínas conduciendo a los investigadores

Sin embargo resulta imposible para el experto mantenerse al día con toda la literatura relevante de forma manual, incluso en temas especializados de su dominio. Por esta razón ha sido necesaria la creación de nuevas técnicas para la solución de estos problemas que han favorecido la integración de muchos expertos procedentes de áreas como la Biología, la Informática, la Física o las Matemáticas, dando lugar a una nueva ciencia interdisciplinaria conocida como Biología de sistemas.

El solo protagonismo de la biología en estos temas no abarca una comprensión global de todas las entidades biológicas, sean células, genes o proteínas, se necesita una comprensión de las interacciones de los genes y proteínas en todos los seres vivos para ayudar en el estudio de enfermedades como el cáncer, porque en estas interacciones están implicadas un importante número de genes y perturbaciones externas.

Existen genes y proteínas que contribuyen a originar un cáncer los cuales, en sus "versiones normales", regulan el crecimiento y la vida celular. Se cree que muchos tumores son el resultado de una alteración mutagénica[1] no reparada en el ADN situación que deriva en el cáncer. Las alteraciones resultantes hacen que las células inicien un proceso de proliferación descontrolada e invadan tejidos normales. El desarrollo de un tumor maligno requiere de muchas transformaciones genéticas, por esto es importante conocer la interacción de estos genes al interior de la célula.

[1] Una sustancia o agente físico que causa mutaciones, es decir, que altera de forma permanente el ADN de las células

Las técnicas que se necesitan para conocer las interacciones de determinadas proteínas se basan principalmente en extracción de información relevante que permitirá una mejor comprensión de los sistemas biológicos y ayudará a la predicción de procesos biológicos y formulación de nuevas hipótesis, transformando toda esta información en conocimiento.

Por lo anterior es necesario crear modelos para estudiar las interacciones moleculares, inferir sobre el comportamiento de ciertas moléculas bajo determinadas condiciones. En este contexto surge la minería de textos como herramienta que facilita la comprensión y conocimiento de una gran red de interacciones de alta complejidad para el manejo humano.

La complejidad en la extracción de interacciones implica una tarea ardua en cuanto a la obtención de los datos los cuales no se encuentran de manera estructurada. Ante este reto es necesario innovar con otras técnicas y desarrollar herramientas de software especializadas para extraer conocimiento de la literatura científica con el fin de crear modelos que contribuyan de manera cualitativa y cuantitativa a la comprensión de todos estos problemas. Un instrumento importante que está en boga actualmente para la solución de estos problemas es la minería de textos, la cual permite extraer información relevante para facilitar la investigación, descubrir conocimiento nuevo y favorecer la creación de nuevas hipótesis. Para cumplir esta tarea la minería de texto se apoya en otras técnicas como el procesamiento del lenguaje natural PLN, el reconocimiento de entidades (RE)y la recuperación de información (RI).

Es importante saber si la información que se extrae es pertinente, hoy en día ha sido difícil establecer una métrica de evaluación común y normalizada, sin embargo, a pesar de estas limitaciones, se han propuesto muchas ideas y realizado esfuerzos que converjan a una adecuada evaluación de herramientas, optado por adecuar las métricas que existen para evaluar los sistema RI para ponderar la eficacia de los métodos.

De los resultados de la evaluación el experto puede generar nuevas hipótesis, o pasar los resultados por otro filtro que mejore las predicciones, representando el conocimiento derivado de los experimentos, en este caso la minería de textos. En la actualidad existen muchos métodos para hacerlo, el presente trabajo pretende examinar la viabilidad de las redes bayesianas como herramienta de causalidades para representar conocimiento que permita predecir nuevas interacciones de proteínas.

1 PRESENTACIÓN DEL PROYECTO

1.1 OBJETIVOS

1.1.1 Objetivo General

Evaluar y aplicar diversas herramientas computacionales de minería de texto para la identificación y caracterización de las interacciones moleculares relacionadas con la proteína Evi1 encontradas en la literatura científica, para la creación de la estructura de una red bayesiana.

1.1.2 Objetivos Específicos

- Examinar las posibles aplicaciones de la minería de texto para el análisis de textos científicos en las áreas de la biología molecular y celular.

- Estudiar y clasificar diferentes herramientas computacionales de minería de texto para obtener información de las proteínas que interactúan con EVI-1.

- Utilizar metodologías de construcción de redes bayesianas para la inferencia de interacciones entre proteínas con base en la información obtenida por minería de texto.

- Construir la estructura de la red Bayesiana que define las diferentes interacciones moleculares existentes a partir de los resultados obtenidos mediante el proceso de minería de texto sobre la información relacionada a proteínas que interactúan con EVI-1.

1.2 PLANTEAMIENTO Y JUSTIFICACIÓN DEL PROBLEMA

La biología molecular se ha preocupado por entender y completar la identificación de la secuencia del genoma humano. Los avances técnicos han permitido ir descifrando poco a poco intrincados problemas biológicos hasta llegar a facilitar en nuestros días una visión precisa y de gran complejidad de los organismos vivos y de la celula en particular. Sin embargo, a medida que se avanza en los procesos investigativos el volumen de datos ha ido aumentando extraordinariamente originado una gran complejidad en su manejo. Como consecuencia de la complejidad y cantidad de información que se necesita manejar emerge la Biología de Sistemas como una disciplina académica que pretende integrar diferentes niveles de información biológica, en procesos tales como la extracción y representación de conocimientos, con el fin de entender el funcionamiento de los sistemas biológicos para intentar crear modelos comprensibles de sistemas a partir del estudio de las relaciones y las interacciones entre las diferentes partes de un sistema biológico (por ejemplo, las redes génicas y las redes de interacción de proteínas implicadas en la señalización celular o las rutas metabólicas de las células).

Actualmente se han identificado un gran número de genes y proteínas que contribuyen a originar un proceso de transformación tumoral, los cuales, en sus "versiones normales", regulan el crecimiento y la proliferación celular. Se cree que muchos tumores son el resultado de mutaciones no reparada en el ADN que codifica para un gen en particular, situación que puede derivar en un tipo de cáncer determinado. Estas alteraciones hacen que las células inicien un proceso de proliferación no controlada e invadan tejidos normales. El desarrollo de un tumor maligno requiere de muchas transformaciones genéticas, que afectan redes muy complejas de transmisión de señales moleculares. Por lo anterior es importante conocer las interacciones específicas de estos genes o sus productos dentro de redes de interacciones moleculares de gran complejidad.

Para estudiar las interacciones moleculares determinantes de un proceso celular normal o patológico es necesario conocer el comportamiento de ciertas moléculas bajo determinadas condiciones, predecir posibles resultados sobre la función de un estudio y proponer modelos que pudieran ser útiles incluso para el diagnostico o el tratamiento de una condición patológica específica.

Afortunadamente en la actualidad se han desarrollado una gama de herramientas para ayudar a construir modelos de cómputo, de muy alta complejidad. Para crear estos modelos, dada la alta complejidad, se necesita un proceso de abordaje estratégico. La obtención de los datos puede ser una tarea ardua y dispendiosa. Hasta ahora, en la gran mayoría de los casos de interés, los datos se encuentran

en sistemas o bases de información no estructuradas. En estos casos, las técnicas de minería de datos podrían aportar soluciones interesantes para extraer información útil y significativa para problema de estudio en particular, como por ejemplo la identificación y el mapeo de la red de interacciones moleculares de la proteína Evi1 que se asocian a la transformación tumoral de células del sistema hematológico.

Finalmente es muy útil y conveniente estructurar y representar el conocimiento derivado del uso de herramientas como la minería de datos. En la actualidad existen muchas opciones metodológicas para hacerlo; Nuestro estudio pretende además, examinar la viabilidad de las redes bayesianas como herramienta para establecer relaciones de causalidad, las cuales permitirían representar la información en forma de una red de interacciones moleculares, que permitan inferir posibles nuevas interacciones no detectadas, y aprender sobre los mecanismos moleculares implicados en las patologías tumorales asociados con expresión anormal de la proteína Evi1.

1.3 ESTADO DEL ARTE

Muchas aproximaciones a la extracción de interacciones entre proteínas han usado fuentes de datos de alto procesamiento o métodos computacionales de naturaleza estadística para analizar la información almacenada en bases de datos que contienen información detallada de las proteínas. Un primer acercamiento a este tipo de trabajos lo realizaron (Friedman, C. et al. , 2001), (Pe'er, 2005) y (Hartemink, 2001) quienes optaron por aplicar aprendizaje de vías biológicas directamente de datos altamente procesados, este aprendizaje lo hicieron sobre una estructura de red bayesiana para datos de expresión de los genes muy distinto al presente trabajo que trabaja sobre datos obtenidos de la literatura.

Los estudios de (Friedman, C. et al. , 2001) y (Pe'er, 2005) presentaron un alto costo, se medían miles de moléculas cientos de veces y la cantidad de datos para intervención era insuficiente siendo difícil inferir. (Hartemink, 2001) empleó datos de estas relaciones, indicando que genes activaban y que cascada se formaba de

los genes que seguían rio abajo en la red de proteínas. Otro investigador, Segal seleccionó potenciales vías reduciendo la necesidad de intervenir en los datos.

(Sachs, Perez, Pe'er, & LauffenburgeR, 2005) unificaron los avances de los anteriores trabajos utilizando algoritmos de aprendizaje bayesiano sobre un grafo que se obtuvo de correlaciones analizadas de varios parámetros obtenidos con la

citometría de flujo sometido a diferentes condiciones para hacer una red bayesiana y analizar relación por relación teniendo en cuenta las moléculas que activaban o inhibían.

Por otro lado (Saric, J. et al., 2004) implementaron técnicas de minería de texto con normas sintácticas y semánticas para la extracción de información sobre relaciones en los resúmenes de la literatura biomédica sobre la levadura, obteniendo una precisión alta al considerar un número más grande que genes que los utilizados en otros trabajos hechos sobre las interacciones de la levadura, para que posteriores trabajos sean aplicables a otros organismos como el hombre.

Sin embargo, a pesar del aumento de la actividad en esta área, aún hay relativamente pocas herramientas de minería de texto para inferir interacciones entre proteínas. Un listado de las herramientas más usadas se encuentran en la tabla (4). Muchos grupos que han hecho acercamientos en aplicaciones de minería de textos, han abordando diferentes problemas, a menudo utilizando conjunto de datos privados (Hirschman, et al., 2002).

En cuanto a evaluación de herramientas en la minería de textos en la biología muchos equipos de investigadores han organizado reuniones, talleres y conferencias anuales para compartir sus avances logrando como resultado un número cada vez mayor de evaluaciones comunes para la minería de textos aplicada a la biología, así como los recursos y conjuntos de punto de referencia. Estas evaluaciones incluyen los siguientes grupos:

CAPRI (Critical Assessment Of Prediction Of Interatrions) comunidad que evalúa las técnicas de predicción de las interacciones de proteínas.

GASP Genome Annotation Assessment Project, concurso para la evaluación de las técnicas de bioinformática para el genoma.

TREC: Text Retrieval Conference. Conferencia anual para la Evaluación de la recuperación de la información y los enfoques de clasificación de documentos bajo la organización del US Nacional Institute for Standards and Technology (NIST). Anualmente, desde 2003 se celebra TREC Genomics para discutir una gran variedad de resultados sobre diferentes corpus predominando los trabajos sobre reconocimiento de entidades.

BioCreAtIvE: Evaluación crítica de Extracción de la información en Biología, se centra en dos tareas: la identificación de genes y proteínas mencionados en la

literatura y la normalización, y la anotación funcional de entidades biológicas usando términos GO (Gene Ontology).
BioNLP: Etiquetado de los nombres de biológicos en los resúmenes de MEDLINE. Esta evaluación se realizó como parte de la Internacional Joint Workshop on Natural Language Processing in Biomedicine and its Applications JNLPBA en 2004.

MUC: Message Understanding Conferences, Evalúa desde 1990 los métodos de extracción de información proporcionando datos elaborados y definición de tareas, además de proporcionar un software totalmente automatizado de puntuación para medir la máquina y el rendimiento humano.

Otras colecciones de pruebas desarrolladas por investigadores han servido para testear diferentes sistemas de minería biomédica: Knowledge Discovery from Database (KDD) Challenge Cup (YEH *et al*., 2003) El propósito de KDD fue analizar cómo las técnicas de la minería de textos pueden ayudar a los curadores, encargados del mantenimiento de las bases de datos Biológicas.

La mayoría de avances en de evaluación de los sistemas de minería textos se han realizado por separado por grupo de investigadores de forma individual. Por tanto, es imprescindible la normalización de las evaluaciones para unificar criterios de medida que permitan mejorar los sistemas.

2 MARCO TEÓRICO

2.1 PROTEÍNAS

Las proteínas son macromoléculas biológicas formadas por unidades denominadas aminoácidos[2], juegan un papel muy importante en las células de todos los seres vivos porque intervienen en el control del ciclo celular, diferenciación celular, plegamiento de proteínas, señalización, transcripción, traducción, y transporte. Por un lado, forman parte de la estructura básica de los tejidos (músculos, tendones, piel, uñas, etc.) y, por otro, desempeñan funciones metabólicas y reguladoras (asimilación de nutrientes, transporte de oxígeno y de grasas en la sangre, inactivación de materiales tóxicos o peligrosos, etc). También son los elementos que definen la identidad de cada ser vivo, ya que son la base de la estructura del código genético (ADN) y de los sistemas de reconocimiento de organismos extraños en el sistema inmunitario.

Adoptan una estructura espacial tridimensional organizada y ordenada, estrechamente relacionada con su función específica. La función desempeñada por cada proteína puede estar regulada mediante la interacción con otras proteínas y diferentes tipos de moléculas de bajo peso molecular (moléculas orgánicas pequeñas o metales) denominadas ligandos, cuya función puede ser de regulación (actuando como activadores e inhibidores) o constitutiva (cofactores), de modo que dicha interacción permite o imposibilita la función de la proteína. Normalmente la interacción con otras moléculas provoca un cambio en la conformación tridimensional de una proteína y, por tanto, modula su función.

El término proteína, proviene del griego “proteíos”, que significa ser el primero en influencia, indica que todas las funciones básicas en biología dependen de proteínas específicas (Robles V. 2003). Se puede decir que no existe vida sin proteínas. Están presentes en cada célula y en cada componente celular. La gran diversidad de estructuras de las proteínas indica la enorme cantidad de funciones que realizan:

Catalización biológica (enzimas). La mayor parte de las reacciones químicas que suceden en los sistemas biológicos están catalizadas por enzimas, que son proteínas aunque realmente no todos los enzimas son proteínas, como por ejemplo la rihozima que es una molécula de ARN.

Almacenamiento y transporte. Las proteínas están involucradas en el almacenamiento y transporte de partículas, que pueden ir desde electrones hasta macromoléculas.

[2] Los aminoácidos son pequeñas moléculas cuya unión forma a las proteínas**.**

Regulación biológica (hormonas). Las proteínas están presentes en la transmisión de impulsos nerviosos, actuando como receptores de pequeñas moléculas que cruzan las uniones que separan las células nerviosas. Dentro de un organismo los procesos biológicos deben estar coordinados entre células del mismo tejido e, incluso, entre organismos diferentes. Esto se logra a través de las moléculas de señalización llamadas hormonas. Algunas hormonas son proteínas, como por ejemplo la insulina.

Función inmunológica (anticuerpos). El sistema inmunológico depende de la Producción de anticuerpos, que son proteínas capaces de acoplarse a partículas específicas exteriores tales como bacterias y virus.

Función regularizadora. La información requerida para sintetizar proteínas es almacenada en genes (secuencias de ADN). La orquestación necesaria para la actividad celular necesita que varias proteínas estén presentes en las cantidades apropiadas en el momento correcto. Las enzimas sintetizan proteínas traduciendo secuencias de ADN. Esta producción puede ser estimulada o inhibida por otras proteínas en complejos mecanismos de retroalimentación.

Función estructural. Algunas proteínas tienen un papel estructural, proporcionando mecanismos de soporte. El esqueleto de una célula consiste en una compleja red de filamentos proteicos. A gran escala, la contracción muscular depende de la acción de grandes cadenas de proteínas. Otro material orgánico, como el pelo y los huesos, están también basados en proteínas.

En este presente trabajo nuestro interés en las proteínas estriba en las interacciones entre ellas independiente de su función, se intentará aplicar la minería de texto para representar y extraer redes de proteínas.

2.1.1 Proteína EVI1

EVI-1*(*ecotropic viral integration site 1*)*, es un gen expresado en tejido hematopoyético[3] durante la embriogénesis[4] que actúa como factor de transcripción, activador o inhibidor de vías de señalización. Pero su expresión en

[3] El tejido hematopoyético es el responsable de la producción de células sanguíneas.
[4] Proceso de división y diferenciación celular que se inicia tras la fertilización de los gametos para dar lugar al embrión, en las primeras fases de desarrollo de los seres vivos pluricelulares.

tejido adulto, aumenta la proliferación, supervivencia y resistencia a apoptosis [5] e inhibe diferenciación, especialmente mieloide. *(*Laricchia-Robbio1, 2008).

Varios estudios han confirmado que la sobreexpresión de EVI1 es un factor de mal pronóstico en pacientes con LMA (Leucemia mieloide aguda); hay muchas cuestiones que quedan por resolver con referencia al mecanismo de sobreexpresión del gen EVI-1 y de cómo coopera en la transformación leucémica. (Odero, 2009).

Las leucemias mieloides agudas (LMA) son enfermedades que surgen como consecuencia de alteraciones genéticas adquiridas en células progenitoras hematopoyéticas. Se caracterizan por una proliferación incontrolada de células inmaduras, blastos[6], que infiltran la médula ósea e invaden la sangre periférica y otros órganos.

2.2 Interacciones Proteína-Proteína

Gran parte de las funciones biológicas están mediadas por interacciones entre proteínas. La interacción puede ser física (p. ej. las proteínas se unen) está a su vez pueden ser estables o trascendentes, o lógica (p. ej. cuando una proteína afecta el comportamiento de otra).

Como se anota anteriormente las interacciones entre proteínas son importantes en muchos procesos biológicos. Por ejemplo, las interacciones proteína-proteína de las moléculas de señalización median el paso al interior de la célula señales procedentes del exterior. Este proceso, llamado transducción de señales, juega un papel fundamental en muchos procesos biológicos y enfermedades (ej. cáncer). Las proteínas pueden interactuar durante mucho tiempo para formar parte de un complejo proteínico; pueden transportar material a otra proteína, pueden interactuar brevemente con otra proteína para modificarla Esta modificación de proteínas puede cambiar por sí misma la interacción proteica.

En definitiva, las interacciones proteína-proteína tienen una importancia capital en prácticamente todos los procesos en una célula viva. La información acerca de esas interacciones mejora nuestro conocimiento sobre las enfermedades y puede proporcionar las bases para nuevos enfoques de los tratamientos terapéuticos.

Entender la forma como interactúan las proteínas y enzimas es un área muy importante a investigar en la bioquímica y en la biología celular. En base a esto

[5] Es una forma de muerte celular, que está regulada genéticamente.

[6] Se les da este nombre a las células inmaduras que se encuentran circulando en sangre o dentro de la médula ósea. La presencia de más del 20% de blastos en la médula ósea son indicativos de una leucemia aguda.

nuevas aéreas de investigación especializada en el tema han surgido, el estudio del "interactoma[7]" viene a responder las preguntar que se plantean científicos luego de estudiar genomas y proteomas. En otras palabras, el genoma nos indica las posibles proteínas que puede sintetizar un organismo en particular, el proteoma determina las proteínas que son expresadas en un momento en particular bajo condiciones establecidas y el interactoma como interactúan las proteínas para otorgar funcionalidad y sincronización de los múltiples procesos de la célula estudiada.

2.3 Minería de Texto

La meta fundamental de la minería del texto es recuperar información que se oculta en texto de manera no estructurada y presentar el conocimiento destilado a los usuarios en una forma sucinta. La ventaja de la explotación minera del texto es que permite a científicos recoger eficiente y sistemáticamente información, acceder, interpretar y descubrir el conocimiento necesario para la investigación.

2.3.1 Fases De La Minería De Texto

A fin de contribuir en este campo aún creciente, es importante sistematizar los métodos que ya están en uso. En la literatura se encuentra que la mayoría de los métodos dividir en cinco diferentes pasos. Los pasos son los siguientes:

- Recolección de textos
- Preprocesamiento
- Análisis de datos
- Visualización
- Evaluación.

2.3.1.1 Recopilación de textos

En el proceso de recopilación de textos es aconsejable que sea de forma automática, generalmente se puede buscar mediante expresiones booleanas, por tanto, es muy fácil de obtener grandes grupos de textos, los que contienen las palabras especificadas, como ejemplo en la literatura biomédica una gran parte de estos textos se encuentra en PubMed una base de datos contiene más de 12 millones de referencias de las publicaciones biomédicas. El paso más básico en esta fase es descargar automáticamente los resúmenes implementando un

[7] Conjunto de todas las interacciones entre proteínas en un contexto dado (célula, organismo, proceso fisiológico, etc.)

programa que descarga resultados de la consulta en una base de datos especialmente diseñados y prepararlos para su uso posterior.

2.3.1.2 Preprocesamiento de textos

El pre-procesamiento es una tarea necesaria para la preparación de los datos y tiene un impacto significativo en el la etapa de análisis de datos y permite que los datos que van a ser utilizados conserven su coherencia.

El propósito del pre-procesamiento de datos es principalmente corregir las inconsistencias de los datos que serán la base de análisis en procesos de minería de texto. El paso fundamental en esta etapa es reconocer los tokens(unidades gramaticales más pequeñas) y etiquetar entidades usando algoritmos de categorización basados en un conjunto de reglas lingüísticas (semánticas y sintácticas), diccionarios técnicos y técnicas de reducción de sentencias. Los algoritmos tienen que tomar en cuenta la diversidad de reglas de puntuación, abreviaturas y otras convenciones según el idioma del texto. Tiene que tomar en cuenta la ambigüedad, pues la mayoría de las palabras pueden recibir un etiquetamiento diferente según el contexto en que estén.

En esta etapa los textos se transforman en algún tipo de representación estructurada que facilite su análisis. Generalmente se usan procesos como:

- Etiquetado de palabras. Marca palabras con etiquetas en función del contexto gramatical de la palabra en la frase, por lo tanto dividir las palabras en los nombres, verbos, etc.
- Tokenización: se refiere a la división del texto en palabras o términos.

Todo lo anterior es importante para el análisis exacto de las relaciones entre las palabras.

2.3.1.2.1 Módulos en el pre-proceso

Según (Zhou & He, 2008) la minería de texto necesita de los siguientes módulos para extraer información.

Módulo de zonificación. Aquí se dividen los documentos bloques básicos para su posterior análisis. Para la construcción de los bloques se identifican las frases, oraciones y párrafos con un valor especial en la temática de consulta. Las herramientas que se usaron compararon los resultados empleando diferentes unidades de texto, como frases, sentencias oraciones, y los resúmenes de la base de datos de artículos médicos PubMed.

Modulo de reconocimiento de proteínas. Antes de la extracción de las interacciones proteína-proteína, es crucial para facilitar la identificación de los nombres de las proteínas, que no es una tarea simple por lo visto anteriormente.

Modulo de extracción de interacciones. A medida que la recuperación de las interacciones proteína-proteína han atraído mucha atención en el campo de la extracción de la información biomédica, se han propuesto muchas alternativas de solución que van desde simples métodos estadísticos que dependen de co-ocurrencias o co-citación de proteínas en los textos a los métodos que emplea un profundo análisis sintáctico y semántico.

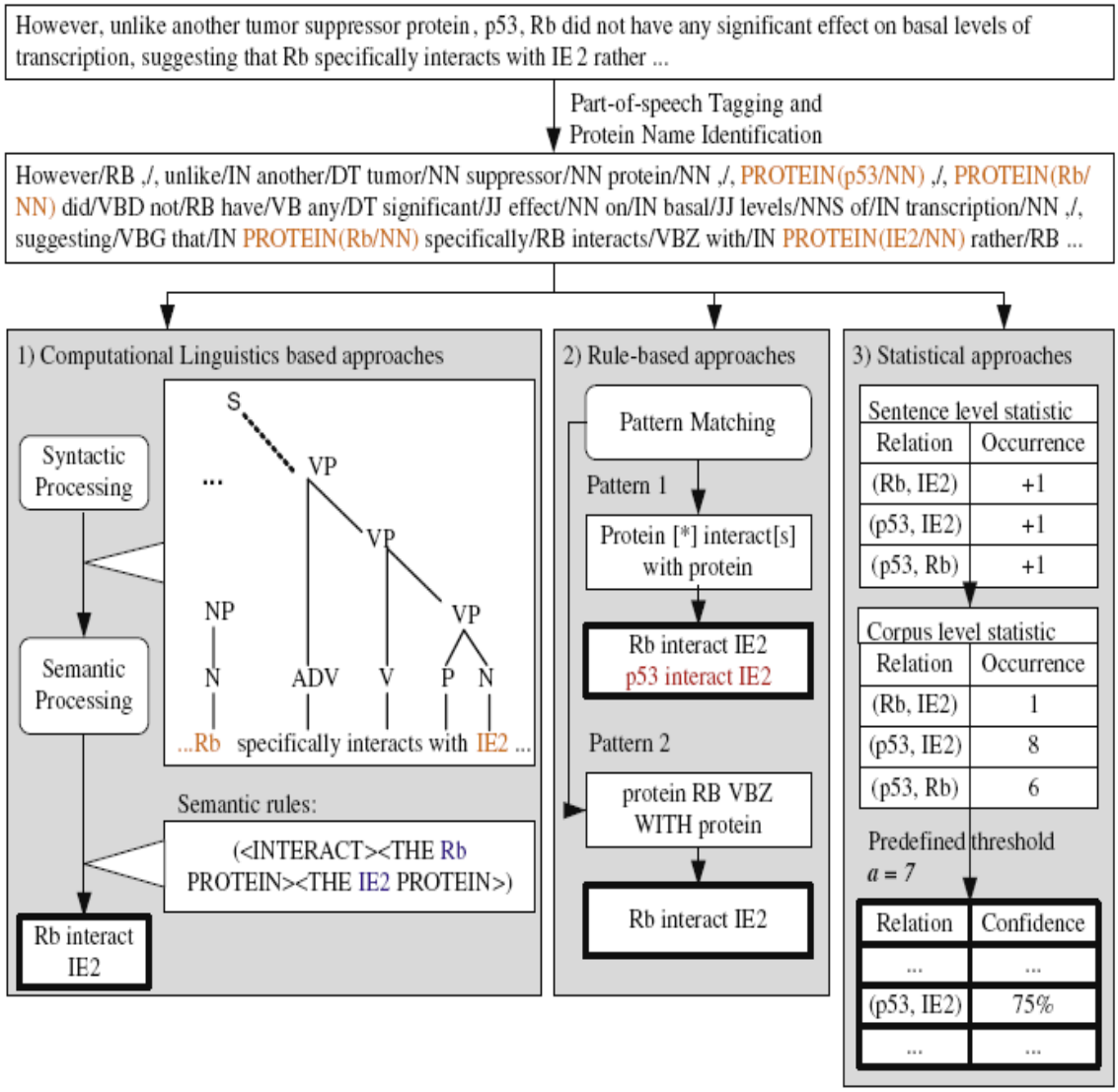


Figura 1 Módulos de extracción de información **(Zhou & He, 2008)**

2.3.1.2.2 Técnicas de preprocesameinto

2.3.1.2.2.1 Métodos computacionales basados en la lingüística

Para descubrir conocimientos de texto no estructurado, es natural emplear la lingüística computacional, como el análisis sintáctico o el análisis semántico para analizar sentencias estructuras. Los métodos de esta categoría definen gramáticas para describir la estructura de las oraciones y el uso de analizadores para traer información sintáctica y dependencias internas dentro de las oraciones. La siguiente figura ilustra el proceso de estos métodos.

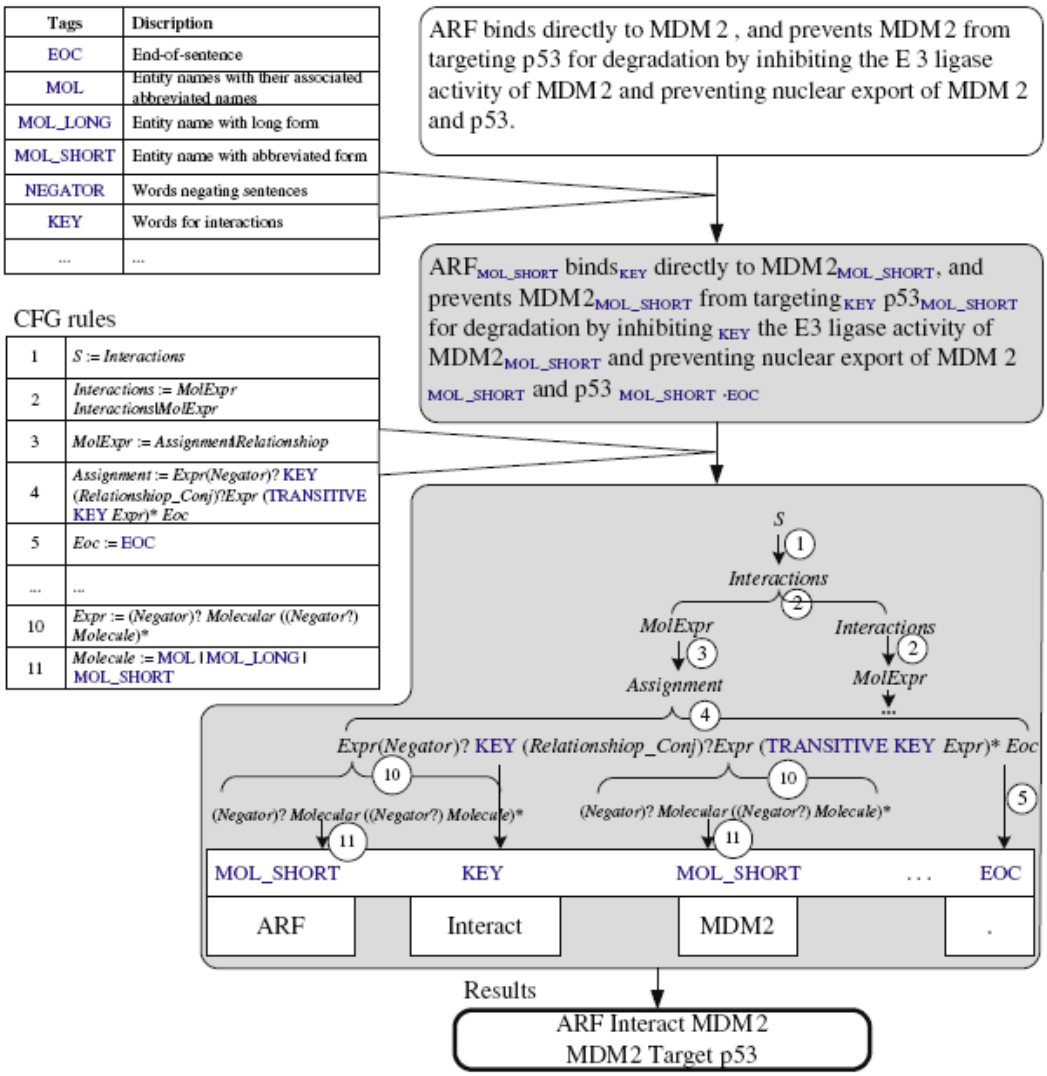


Figura 2 Método basado en Lingüística **(Zhou & He, 2008)**

2.3.1.2.2.2 Métodos basados en reglas.

Estos definen un conjunto de reglas para posibles relaciones textuales, patrones que codifican estructuras similares en la expresión de las relaciones. Cuando se combina con métodos estadísticos, esquemas de puntuación en función de las ocurrencias de los patrones para describir la confianza de la relación que se utilizan normalmente.

2.3.1.2.2.3 Métodos basados en aprendizaje automático y estadísticos.

El aprendizaje automático se refiere a la capacidad de una máquina para aprender de la experiencia para extraer conocimiento a partir de corpus de datos. A diferencia de las dos categorías anteriores estos necesitan un laborioso esfuerzo por definir un conjunto de reglas o gramáticas, Las técnicas de máquina de aprendizaje son capaces de extraer patrones de interacciónes proteína-proteína, sin intervención humana.

Los métodos estadísticos se basan en acontecimientos de la palabra en un cuerpo de texto grande. Las características más significativas o patrones se detectan y se utilizan para clasificar los resúmenes o frases que contienen interacciones de proteínas e identificar las características correspondientes a cada una de las relaciones entre los genes o las proteínas.

2.3.1.3 Análisis De Datos

Este es el paso más diverso y optimizado de los cinco. Muchas de las técnicas de minería de datos son aplicables aquí, debido a que en esta etapa es donde la extracción de la información real sucede. El análisis de datos es muy dependiente de la etapa de pre-procesamiento y la representación del modelo de datos que fue elegido en proceso previo.

En esta etapa se implementan técnicas como redes neuronales, algoritmos bayesianos, modelos de Markov, modelos de Clustering y aprendizaje no supervisado.

2.3.1.4 Visualización

Es una fase clave para representar los resultados que se obtienen de la extracción de información pues lo que no se ve no sirve para nada. En esta etapa se cuenta

con una gran cantidad de herramientas que usan distintas formas de visualizar los resultados obtenidos. Lo más simple es hacer una tabla para que el usuario acceda a la información que necesita, pero existen formas más elaboradas, como visualización de grafos y otros métodos más complejos que conecta a otra tipo de información por medio de enlaces en un objeto.

La otra cuestión de la visualización es, la cantidad de datos para mostrar al usuario. Por lo general, el usuario se enfrenta con los resultados puros sin la metainformación sobre cómo y por qué los resultados fueron recuperados. Este es especialmente importante si los resultados contienen algún tipo de valoración y el usuario quiere saber que era exactamente lo que hizo un resultado superior a otro.

2.3.1.5 Evaluación

Los métodos clásicos de evaluación son las diversas formas de validación cruzada y unidades de prueba y maquinas de aprendizaje supervisado a fin de optimizar sus parámetros.

En el caso de estudio aplicado a la biología, la minería de textos se ha guiado principalmente por los lingüistas computacionales, pues los biólogos están empezando a explorar los métodos, el escenario clásico surge de dos comunidades de investigación que necesitan comunicarse con el fin de aprender cada una de la otra. Esto comienza con el desarrollo de un lenguaje común y criterios de evaluación comunes. Para evaluar un método de la minería de texto, su salida es tanto comparada con un estándar o patrón oro o inspeccionada de forma manual por un experto. Esto produce tres valores importantes: recuperaciones/ extracciones correctas (verdaderos positivos, TP), errores de tipo I (falsos positivos, FP), y errores de tipo II (falsos negativos, FN). A partir de estas, otras medidas se pueden derivar otras que tienen mayor significado.

		Patrón oro		
		Positivo	Negativo	Total
Resultado Test	Positivo	TP	FP	TP + FP
	Negativo	FN	TN	FN +TN
	Total	TP+FN	FP+TN	TP+FP+FN+TN

Tabla 1 *Definición de las medidas* **(Hersh, 2005)**

Dentro de la minería de Textos los más comunes son:

- Recall (índice de recuperación): La fracción de los documentos pertinentes a la clase c que se recuperaron, es también conocida como la sensibilidad, este término tiene mucho que ver con la relevancia.

  $$\text{Recall}_C = \frac{TP_C}{TP_C + FN_C} \quad [1]$$

- Precisión: La fracción de los documentos recuperados que sean pertinentes en la clase c, también conocida como la especificidad.

  $$\text{Precision}_C = \frac{TP_C}{TP_C + FP_C} \quad [2]$$

- F-Resultado: La medida más utilizada para la clasificación de información de recuperación, el reconocimiento de entidades y los métodos de extracción de información. Se define como la media armónica de recall y la precisión. Debido a la importancia relativa de recuperación y precisión entre varía de las tareas, el método con la mejor puntuación de F-no es necesariamente lo mejor para una determinada tarea lo que genera uno de los mayores problemas con estas estimaciones de la calidad.

  $$F_C = \frac{2.\text{Precision}_C \,.\text{Recall}_C}{(\text{Precision}_C + \text{Recall}_C)} \quad [3]$$

Para el índice de recuperación se tiene el inconveniente que debe conocerse el número total de documentos relevantes para la consulta. Sin embargo, es poco probable, tal vez incluso imposible, para tener éxito en la identificación de todos los documentos pertinentes en una base de datos.

El objetivo de estas evaluaciones se orienta principalmente a la investigación de los sistemas en recuperar información confiable, con medidas basadas en la pertinencia porque miden la proporción de relevancia de las relaciones recuperadas de un cuerpo de datos.

Como se puede observar, todas las etapas están muy interrelacionadas, así pues, la primera etapa condiciona el descubrimiento de los patrones que la minería de texto puede realizar que se necesita para analizar y visualizar los resultados para su validación.

2.3.2 Minería de textos aplicada a la Biología

Una vez descifrado la secuencia del genoma humano, el paradigma de investigación ha cambiado dando paso a la descripción de las funciones de los genes y a futuros avances en la lucha contra enfermedades. Este nuevo contexto

ha despertado el interés de la Bioinformática, que combina métodos de las Ciencias de la Vida con las Ciencias de la Información haciendo posible el acceso a la gran cantidad de información biológica almacenada en las bases de datos, y de la Genómica, dedicada al estudio de las interacciones de los genes y su influencia en el desarrollo de enfermedades. En este contexto, la minería de textos surge como un instrumento emergente para el análisis de la literatura científica. Una tarea habitual de la minería de textos en Biología Molecular y Genómica es el reconocimiento de entidades biológicas, tales como genes, proteínas y enfermedades. El paso siguiente en el proceso de minería lo constituye la identificación entre entidades biológicas, tales como el tipo de interacción entre gen- gen, gen-enfermedad, gen-proteína, para interpretar funciones biológicas, o formular hipótesis de investigación (Galves, 2003).

El biólogo necesita estar al tanto de los nuevos descubrimientos que se publican, pero las publicaciones crecen a un ritmo considerable que rebasan las capacidades de asimilación de cualquier ser humano incluso en temas de su dominio, por lo anterior la minería de textos se torna indispensable.

Para extraer conocimiento la minería de texto estriba en procesos como la recuperación de la información (IR), el reconocimiento de entidades (ER), y la extracción de información (IE). La integración de minería de textos con otra fuente de datos genera los mejores resultados en la extracción del conocimiento. La siguiente figura ilustra el estado actual de estos enfoques y la importancia de cada una de ellos.

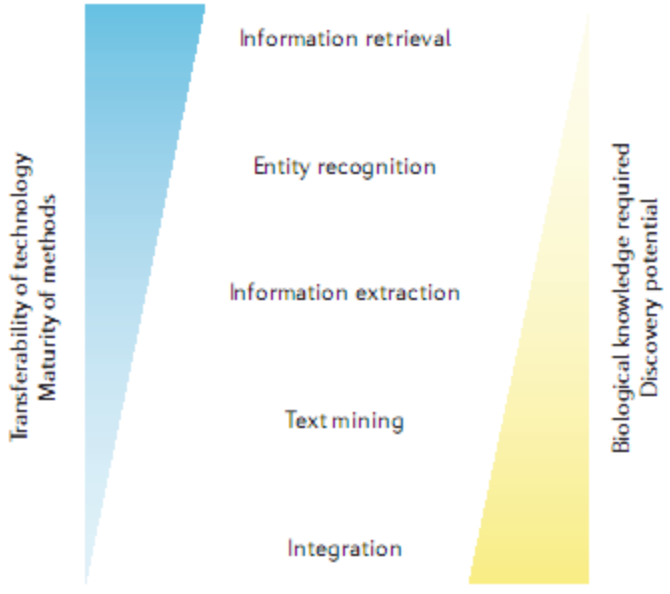


Figura 3 Técnicas asociadas con la minería de textos, (Jensen, Saric2, & Bork, 2006)

Mientras que la recuperación de información, el reconocimiento de entidades y extracción de información se establecen tareas de lingüística computacional, los métodos de otros campos se han transferido a la biomedicina. La minería de

textos y la integración de datos se encuentran todavía en su infancia debido a que pocos métodos se han propuesto. Esto es debido a una visión biológica muy amplia que se requiere para desarrollarlos. Sin embargo, estas herramientas son de gran valor para biólogos porque ayudan a en sus investigaciones proporcionando el mayor potencial para liderar nuevos descubrimientos biológicos.

2.3.2.1 Recuperación de información (RI)

Aunque el enfoque principal de este trabajo es la minería de texto es necesaria una breve revisión de las más importantes metodologías que se utilizan en la minería de texto.

Los sistemas RI se enlazan a una colección de textos teniendo como objetivo identificar los segmentos de texto (ya sea completo artículos, resúmenes, párrafos o frases) que pertenecen a un tema determinado sin ser necesaria su mención explícita. El sistema RI más conocido es PubMed, el cual se basa en dos modelos: el modelo booleano y el modelo vectorial. El modelo booleano permite al usuario recuperar todos los documentos que contienen ciertas combinaciones de términos mediante una operación lógica, por ejemplo, “EVI1”y “Leucemia “. Por el contrario, el modelo vectorial representa cada documento por un vector de expresión, en la que cada término se le asigna un valor de acuerdo a una frecuencia basada en esquema de ponderación. Estos vectores documento posteriormente pueden ser comparados con un vector de consulta que especifica la relativa importancia de cada término de la consulta. Alternativamente, pueden ser comparados entre sí para calcular la similaridad de documento, que se utiliza en PubMed por la función ‘artículos relacionados ‘y otros métodos de clustering de documentos.

Los sistemas RI tienen que ser capaces de superar el alto índice de ambigüedades, sinónimos que la terminología biomédica. Ante este problema estos sistemas incluyen la eliminación de palabras irrelevantes como ‘la’ y ‘lo’, que se producen en casi todos los documentos, y truncan terminaciones de palabras comunes como “-ing ‘y’- ‘s’ para permitir que las diferentes formas de la misma palabra se ajusten a la búsqueda, por ejemplo, “interacts” y “interating” representan una misma búsqueda. PubMed y muchos otros sistemas biomédicos IR también hacen uso de tesauros para automáticamente ampliar la consulta con otros términos asociados. Por ejemplo, la consulta booleana de cualquier proteína podría ampliarse a cualquiera de sus sinónimos o a la función más conocida de esta. Muchos métodos RI avanzados, también usan métodos ER para identificar mejor los documentos que mencionan un determinado gen o proteína. Sin embargo los avances logrados en este campo los actuales sistema IR no son garantía de recuperar con precisión todo tipo de consulta.

Estos sistemas tienen como tarea la identificación de los nombres y símbolos de las entidades biológicas. La identificación de nombres es un paso previo, que permitirá establecer posteriormente las posibles relaciones, lo que puede resultar un trabajo sencillo, pero en realidad el sólo hecho de identificar una palabra en un

texto de manera unívoca es un trabajo dispendioso porque se debe ahondar en temas lingüísticos que abarcan gran complejidad de reglas.

Los sistemas ER tiene dos tareas principales: en primer lugar, el reconocimiento de palabras que se refieren a las entidades y en segundo lugar, la identificación única de las entidades en cuestión.

2.3.2.2 Reconocimiento de las entidades (RE)

Los primeros métodos de RE se basaban en normas de forma manual que establecían características típicas de nombres como letras que van seguidas de números, o la terminación "-asa" -, así como información contextual de las palabras cercanas, como "gen" y "receptor.

En contraste con estos sistemas, muchos RE utilizan el enfoque basado en diccionarios el cual utiliza una lista de nombres de sinónimos de genes que se comparan con los documentos mediante algoritmos que permiten la variación en cómo se escriben los nombres - por ejemplo, 'Cdc28, 'Cdc28', 'Cdc28p "o" CDC-28'. Los enfoques Basados en diccionarios tienen una ventaja crucial sobre los basados en características no sólo reconoce los nombres, sino también identificar los números de accesión de los genes o las proteínas a que se refieren. Muchos sistemas combinan comparación con diccionario con métodos basados en reglas o estadística para reducir el número de falsos positivos. Los mejores métodos de realización de ER en las evaluaciones se basan en curación cuidadosa de listas de nombres de genes para eliminar alias que causan muchos falsos positivos. La principal dificultad en ER surge de la falta de normalización de los nombres. Cada gen o proteína típicamente tiene varios nombres y abreviaturas (por ejemplo, el gen Acf1, con 13 alias (CG1966, ACF, ATP, CAF, acf1, p170/p185, CHRAC, dACF, dCHRAC, ACF1, Acf-1, Acf, CHRAC-175), otras proteínas también son palabras comunes en inglés, por ejemplo, "hairy" o "bad". (Gálvez, 2008).

Una vez identificadas las entidades biológicas, se tienen que resolver los problemas de sinonimia y abreviaturas, que podrían ser unificadas a continuación en alguna forma normalizada. Existen también problemas de homonimia y abreviaturas cuando una misma entidad biológica puede referirse a múltiples entidades o puede ser la abreviatura de varias entidades, como la abreviatura de nombre de gen PSA, que se refiere a los nombres de genes 'Puromycin-Sensitive Aminopeptidase', 'Prostate Specific Antigen', 'PSoriatic Arthiritis', 'Phosphoserine Aminotransferase'. Varios trabajos de minería de textos se han dedicado a resolver estos problemas de ambigüedad (Liu, et al., 2002); (Yu et al, 2002); (Chang et al., 2002) y (Tuason al., 2004). Frente a estas investigaciones, el problema de la normalización de genes es un campo relativamente nuevo e inexplorado.

El reciente desarrollo de métodos para resolver ambigüedad de nombres de genes o proteínas es por lo tanto un avance importante para la RE. Aunque RE es en general como un bloque de construcción para sistemas RI e EI, también puede ser útil por sí solo para vinculación de la literatura cruzada que se relaciona con ciertos genes.

2.3.2.3 Extracción de la Información (IE)

Estos sistemas extraen tipos de datos predefinidos como relaciones entre entidades biológicas, desde nuestra secuencia ejemplo, un IE sistema debería deducir que Cdc28 se une Clb2, que Swe1 es fosforilado por el complejo Cdc28-Clb2 y Cdc5 está involucrado en la fosforilación de Swe1. Dos enfoques fundamentalmente diferentes a la extracción de relaciones de los textos biológicos actualmente se utilizan ampliamente, como son la o-ocurrencia y el procesamiento del lenguaje natural (PLN).

2.3.2.3.1 Co-Ocurrencia

El método más sencillo para IE es identificar entidades que co-ocurren dentro de los resúmenes o sentencias. Como dos entidades podrían ser mencionados juntas sin estar de cualquier manera relacionada, la mayoría de sistemas usa una frecuencia basada en sistemas de puntuación (SCORE) para clasificar las relaciones extraídas. Si dos entidades son reiteradamente mencionadas en conjunto, es probable que de alguna manera estén relacionadas, aunque el tipo de la relación no se conoce. Los métodos de Co-ocurrencia tienden a dar mejor (recall) pero pierde precisión frente a los métodos NLP, y se adaptan bien como partes de herramientas exploratorias debido a su capacidad para identificar las relaciones de casi cualquier tipo.

Los métodos de co-ocurrencia también puede utilizarse para extraer relaciones de cierto tipo solamente, tales como interacciones físicas proteína-proteína, mediante la combinación con un sistema de categorización personalizado de texto para identificar los resúmenes o frases pertinentes.

Sin embargo, las oraciones complejas que contienen múltiples relaciones pueden dar lugar a relaciones adicionales erróneas. Este enfoque también es incapaz de extraer las relaciones de dirección (por ejemplo, si una proteína está involucrada en la fosforilación de otra o viceversa) y tiene dificultad para distinguir entre las relaciones directas e indirectas del Procesamiento de lenguaje natural.

2.3.2.3.2 PLN (procesamiento del lenguaje natural)

Los métodos de NLP combinan el análisis de la sintaxis y la semántica; el Texto es el primer “tokens” para identificar los límites de una frase y la palabra, y un

parcial-etiquetado del lenguaje (por ejemplo, un sustantivo o verbo) se asigna a cada palabra. Un árbol de sintaxis se deriva entonces para cada sentencia para delinear los sintagmas nominales y representar sus interrelaciones. Los métodos RE y simples diccionarios se utilizan posteriormente para marcar semánticamente las entidades biológicas pertinentes (por ejemplo, los genes y proteínas) y otras palabras clave (por ejemplo, la activación, la represión o la fosforilación).

Por último, un conjunto de reglas se utiliza para extraer las relaciones sobre la base del árbol de sintaxis y las etiquetas semánticas. Pocos sistemas de NLP intentan resolver las relaciones anafóricas, por lo que la mayoría de sistemas son por lo tanto incapaces de extraer las relaciones que abarcan múltiples oraciones.

Esto no es gran limitación como podría parecer porque la mayoría de relaciones se mencionan en una sola frase. Existen varios programas para tokenización y *part-of-speech*[8] de textos, la mayoría de los cuales se adaptan fácilmente a los textos biomédicos de reciclaje en un corpus etiquetado manualmente como GENIA. El etiquetado semántico es más complicado, pero puede ser enormemente simplificado utilizando los métodos existentes RE.

Por el contrario, el desarrollo de la gramática y la extracción reglas que pueden analizar correctamente las oraciones y extracto de los hechos sigue siendo un reto. El flujo de trabajo idealizado descrito anteriormente indica que el análisis sintáctico de las oraciones, y su interpretación semántica se llevan a cabo en dos pasos separados. Sin embargo, los analizadores de Inglés más genérico de bajo rendimiento si aplican directamente a los textos biomédicos debido a la terminología técnica que contienen y, en particular, el uso de frases largas y complejos sustantivos. Los mejores resultados se pueden obtener mediante primera marcado de los sintagmas nominales. Sin embargo, muchos sistemas biomédicos NLP han combinado el analizador sintáctico y las reglas de extracción semántica en un personalizado analizador parcial que se dirige específicamente a sólo las partes pertinentes de las sentencias y directamente extrae los datos. El principal inconveniente de este enfoque es que un gran número de reglas de extracción son necesarias para cubrir las muchas formas ligeramente diferentes de expresar una cierta relación. Estas reglas pueden ser desarrolladas manualmente o aprendidas automáticamente a partir de un corpus.

2.3.2.3.3 Inconvenientes en la extracción de información

La literatura biomédica presenta algunos problemas para la extracción de información, entre los que se destacan:

Vocabulario: El tamaño del vocabulario es muy grande, por ejemplo los sistemas que traten el problema de reconocimiento de nombres de proteínas o genes, se

[8] También llamado **etiquetado gramatical:** asigna (o etiquetar) a cada una de las palabras de un texto su categoría gramatcal.

tienen que enfrentar a cientos de miles y decenas de miles de entidades respectivamente. Adicionalmente, cada día se reportan nuevos descubrimientos que hacen que el vocabulario crezca constantemente. En el caso de las otras palabras importantes que se relacionan con las interacciones existe una amplia cantidad de términos que se deben analizar para identificar una relación, las siguientes tablas muestran un pequeño conjunto de palabras de un número más amplio que existe que dan la idea de la existencia de interacciones en un texto.

abrogation	acetylation	activity/-ation	apparatus
binding	cluster	complex	control
conversion	destabilization	downregulation	effect
expression	hyperexpression	induction	inhibitor/-ion
interaction	ligand	modulation	obstruction
phosphorylation	regulator/-ion	repressor/-ion	stabilization
stimulation	suppressor/-ion	synthesis	upregulation

Tabla 2 *nombres que identifican interacciones.* **(Blaschke, C. et al., 2005)**

acetylate	contact	enhance	interfere	repress
activate	contain	exhibit	link	respond
affect	control	form	modulate	sever
associate	derive	fuse	phosphorylate	stabilize
bind	destabilize	include	potentiate	stimulate
block	dimerize	induce	recognize	suppress
comprise	downregulate	inhibit	recruit	upregulate
conjugate	encode	interact	regulate	yield

Tabla 3 *Verbos comunes que indican interacciones.* **(Blaschke, C. et al., 2005)**

Falta de estándares para la nominación de las entidades: Debido a que muchos autores de diferentes áreas trabajan sobre los mismos tópicos y reportan los resultados de acuerdo al entendimiento que tengan de su particular estudio.

Sinonimia: Derivada de la falta de estándares para la nominación de los nombres temporales que se asignan en la literatura a entidades recién descubiertas, esto hace que una entidad sea referenciada de muchas maneras incluso en el mismo texto, por ejemplo en el ámbito de los genes, en el diccionario LocusLink se reportan 80000 genes de diferentes especies pero cuando se incluyen todos sus sinónimos conocidos el número asciende a 200000.

Polisemia: Es un fenómeno que se presenta cuando una palabra o entidad tiene diferentes significados que dependen del contexto. Los sistemas de IE deben ser

capaces de desambiguar términos cuando sea necesario.

Terminología técnica específica al dominio: Terminología que no encaja dentro de reglas gramaticales, ni sintácticas de ningún lenguaje, este característica hace que

la precisión y el índice de recuperación de los sistemas bajen porque se identifican falsos positivos.

Anáforas: Las referencias mediante pronombres o abreviaciones a entidades introducidas con anterioridad en los artículos son muy comunes en literatura biomédica, las anáforas representan un reto especial que es tratado en conferencias especializadas.

Catáforas: otro fenómeno que suele presentarse con mucha frecuencia en textos biomédicos son las catáforas que representan el fenómeno contrario a las anáforas, es decir, anticipa un conjunto de sentencias o palabras que hacen referencias a una entidad que se expresará enseguida. Las anáforas y las catáforas conforman lo que se denomina problema de resolución de la retórica.

Es común que la información que se refiere a una entidad este dispersa en el texto por tanto se hace necesario un mecanismo que permita combinar información de diferentes sentencias; de la misma manera existen zonas especificas del texto que no aportan información, tales zonas deben ser identificadas y descartadas.

2.3.2.4 Integración en las herramientas

Aunque la minería de textos pueden ser utilizados para descubrir relaciones, el enfoque de data mining que integran la literatura con otros tipos de datos tienen un mayor potencial para hacer descubrimientos biológicos. Como un ejemplo de cómo este podría lograrse, las relaciones que se aplican a una determinada proteína de interés puede ser extraído de la literatura, seguido de búsquedas de secuencias similares a la transferencia estas relaciones en proteínas. De esta manera, métodos de minería de texto se podría utilizar para hacer inferencias que se basan en relaciones de múltiples especies, y por lo tanto conectar a las comunidades de investigadores que trabajan en diferentes modelos de organismos.

La mayoría de las primeras aproximaciones utilizan métodos ER o base de datos de referencias cruzadas para recuperar los resúmenes de las bases de datos que se asocian con uno o más genes por ejemplo, una familia de proteínas o un grupo de genes que son co-expresados en un experimento de microarrays.

Entonces, estos resúmenes se pueden utilizar para identificar importantes sobrerrepresentación de las palabras claves dentro del texto, o anotando términos MeSH, los cuales pueden contribuir a la caracterización de los genes en cuestión. Por otra parte, los resúmenes se pueden utilizar para evaluar la coherencia de clúster o construir una red de asociación funcional de los genes de sus co-ocurrencia en abstracts.

A través de su capacidad de reunir a muchos tipos de datos, las redes tienen el potencial para formar la base para integración de textos y datos. Hay varios web

basada en herramientas que facilitan el acceso a las redes de proteínas que son basados en IE y experimentos de alto rendimiento. Estos han demostrado ser valiosos como instrumentos de exploración que permiten a los investigadores ver muchos tipos de información para un conjunto de proteínas de interés. Además, son útiles para proporcionar una visión estructurada de otro tipo de datos de alto rendimiento. Por ejemplo, los datos de expresión se pueden asignar a las redes de interacción proteína y puede visualizar la forma en que la síntesis de complejos de proteínas es regulada en el nivel de transcripción. Estas redes también pueden combinarse con otros tipos de datos para proporcionar puntos de vista en las bases moleculares de una enfermedad. Por ejemplo, redes de proteínas basadas en la bibliografía se han integrado con estudios de mapeo de ligamiento para identificar genes candidatos para la enfermedad de Alzheimer dentro de una región genómica en la base de sus interacciones con los genes que ya se sabe que tiene un papel causal en la enfermedad.

2.4 REDES BAYESIANAS

Las redes bayesianas aportan conceptos importantes y necesarios para la consecución de los objetivos de este proyecto.

Las redes Bayesianas han cobrado suma importancia actualmente en aplicaciones médicas, es una técnica eficiente para representar el conocimiento y para implementarlas en situaciones donde el conocimiento es incierto, principalmente porque explican influencias entre muchos datos, lo que facilita la interpretación, inferencia, predicción, aprendizaje sobre relaciones de dependencia y causalidad y la combinación del conocimiento con los datos. Gracias a estas ventajas que no dan otras técnicas se pueden manejar datos incompletos, se evita el costo para los algoritmos y también el sobre ajuste de los datos.

En la actualidad el uso de modelos causales para representar el conocimiento se ha hecho muy popular lo que ha aumentado el interés de los científicos en la construcción de estos modelos para comprender y predecir fenómenos. En este tema las redes bayesianas llevan la bandera porque logran combinar causalidad y conocimiento a priori con los datos, los cuales están disponibles de manera cruda, porque son de difícil interpretación para un modelo, por ello una fase importante es la construcción los modelos, y para estos además del conocimiento del experto del dominio quien generalmente puede especificar relaciones de influencia, se necesitan datos procesados y muchos experimentos.

Otros nombres con el que los que se pueden referir a las redes bayesianas son redes de creencia, redes probabilística, redes causales y otros autores la denominan mapas de conocimiento.

Para comprender un poco más las redes bayesianas es necesario tratar unas definiciones previas de teoría de grafos y teoría de la probabilidad:

2.4.1 Definiciones previas

Arco. Es un par ordenado (X, Y). Esta definición de arco corresponde a lo que en otros lugares se denomina arco dirigido. En la representación gráfica, un arco (X, Y) viene dado por una flecha desde X hasta Y.

Grafo dirigido: Es un par G = (N, A) donde N es un conjunto de nodos y A un conjunto de arcos definidos sobre los nodos.

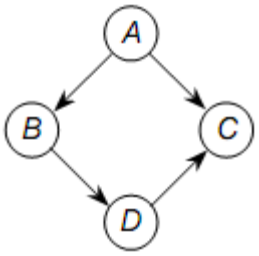


Figura 4 Grafo dirigido **(Diez, 2005)**

Grafo no dirigido: Es un par *G = (N, A) donde N es un conjunto de nodos y A un* conjunto de arcos no orientados (es decir, pares no ordenados (*X, Y)) definidos sobre* los nodos.

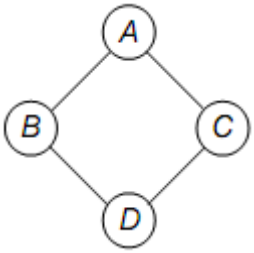


Figura 5 Grafo no dirigido **(Diez, 2005)**

Camino: Es una secuencia ordenada de nodos (*Xi1,..., Xir) tal que*

$\forall$ j = 1,..., r-1, ó bien el arco Xj $\rightarrow$ Xj+1 $\in$ A o bien el arco Xj+1 $\rightarrow$ Xj $\in$ A.

Camino dirigido: Es una secuencia ordenada de nodos (*Xi1,..., Xir) tal que para todo j*= 1,..., *r-1 el arco Xj* $\rightarrow$ *Xj+1* $\in$ *A.*

***Padre: X** es un padre de Y si y sólo si existe un arco X → Y. Se dice también que Y es hijo de X. Al conjunto de los padres de X se representa como pa(X), y al de los hijosde X por S(X).*

2.4.1.1 Tipos de conexiones Causales

Existen tres posibles tipos de conexiones en términos de su capacidad para transmitir información dada o no dada una evidencia.

Conexión en serie:

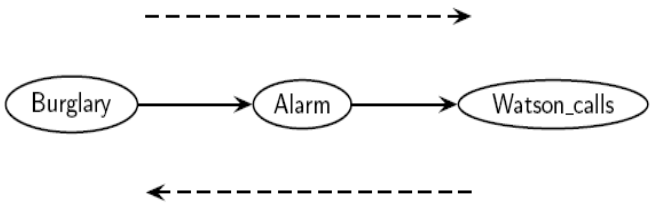


Figura 6 Conexión en Serie **(Uffe & Anders, 2005)**

Conexión divergente:

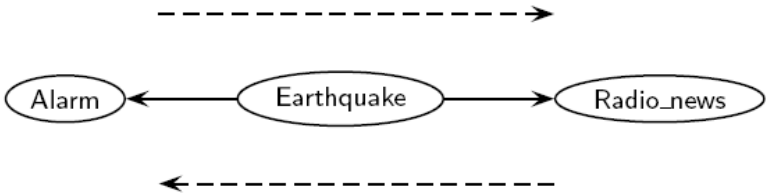


Figura 7 Conexión divergente **(Uffe & Anders, 2005)**

Conexión convergente:

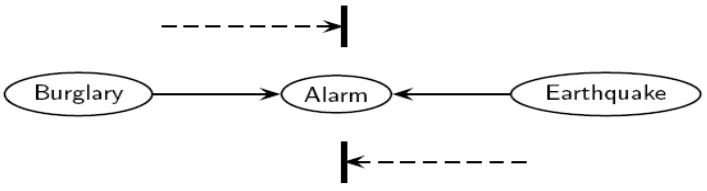


Figura 8 Conexión Convergente **(Uffe & Anders, 2005)**

2.4.2 Teorema de Bayes

El Teorema de Bayes ofrece un método estadístico para calcular una probabilidad condicional en circunstancias de dependencia. Este teorema es de gran utilidad para evaluar una probabilidad a posteriori partiendo de probabilidades simples, y así poder revisar la estimación de la probabilidad apriori de un evento que se encuentra de un estado o en otro.

El Teorema de Bayes, dentro de la teoría probabilística, proporciona la distribución de probabilidad condicional de un evento "A" dado otro evento "B" (probabilidad posteriori), en función de la distribución de probabilidad condicional del evento "B" dado "A" y de la distribución de probabilidad marginal del evento "A" (probabilidad simple o apriori).

$$P\left(A/B\right) = \frac{P\left(AnB\right)}{P\left(B\right)} \qquad [4]$$

2.4.3 Definición formal de las Redes Bayesianas

Formalmente, una red bayesiana es un es un grafo acíclico[9] dirigido (DAG) en la cual cada nodo representa una variable de un dominio y cada arco una dependencia probabilística, en la cual se especifica la probabilidad condicional de cada variable dados sus padres.

Las redes bayesianas o probabilísticas se fundamentan en la teoría de la probabilidad y combinan la potencia del teorema de Bayes [10] con la expresividad semántica de los grafos dirigidos; las mismas permiten representar un modelo causal por medio de una representación gráfica de las independencias o dependencias entre las variables que forman parte del dominio de aplicación.

Una red Bayesiana es construida por la adquisición de conocimiento de un modelo cualitativo. Representa conocimiento genérico, tiene una parte cualitativa, es un grafo acíclico donde los nodos son los atributos y una parte cuantitativa que son las probabilidades asociadas con los atributos; se usan cuando la

[9] Grafo que no contiene ciclos, es decir que no contiene caminos que vayan de un nodo y regresen al mismo.

[10] El teorema de Bayes, enunciado por Thomas Bayes, en la teoría de la probabilidad, es el resultado que da la distribución de probabilidad condicional de una variable aleatoria *A* dada *B* en términos de la distribución de probabilidad condicional de la variable *B* dada *A* y la distribución de probabilidad marginal de sólo *A*.

incertidumbre se asocia con un resultado y puede expresarse en términos de una probabilidad. Este método cuenta con un dominio del conocimiento codificado y ha sido usado para los sistemas de diagnóstico por medio de una representación compacta de distribuciones de probabilidades conjuntas utilizando independencia condicional.

Una relación entre los grafos y la teoría de probabilidades, incertidumbre, complejidad y aprendizaje automático. En ellas podemos identificar los siguientes componentes:

1. Nodos : representan variables
2. Arcos : relaciones de dependencia causales entre variables
3. Tablas de probabilidades :indican la influencia entre los nodos

La red Bayesiana se puede ver como un conjunto (X, D, P) donde:

- X es el conjunto de variables del dominio que se quiere representar. Pueden ser continuas o discretas.
- D es el grafo acíclico dirigido (DAG) cuyos nodos están etiquetados con los elementos de X. Siendo los arcos dirigidos los indicadores de la relación de influencia y en algunos casos relación causal.
- P es la distribución conjunto sobre las variables (X).

Luego tenemos

Sea X = {x1, x2,.., xn} conjunto de variables en una estructura S, que codifica un conjunto de certezas de independencia condicional acerca de las variables X,

P = conjunto de distribución de probabilidades locales asociadas con cada variable.

Xi o nodo o variable, p(ai) son los padres de xi en S. La falta de arcos indica independencias condicionales entre variables. En general, para la estructura S, la distribución de las probabilidades conjuntas de X es:

$$P(X) = \prod_{i=1}^{n} P(X_i|pa_i) \qquad [5]$$

Las distribuciones de probabilidades locales P, son distribuciones en los términos del producto indicado en la ecuación.

Una red bayesiana es un grafo acíclico dirigido, las uniones entre los nodos tienen definidas una dirección en el que los nodos representan variables aleatorias y las

flechas representan influencias causales; el que un nodo sea padre de otro implica que es causa directa del mismo.

Se puede interpretar a una red bayesiana de dos formas:

- Distribución de probabilidad: Representa la distribución de la probabilidad conjunta de las variables representadas en la red.

- Base de reglas: Cada arco representa un conjunto de reglas que asocian a las variables involucradas. Dichas reglas están cuantificadas por las probabilidades respectivas.

2.4.4 Regla de la cadena

La regla de la cadena sostiene que la probabilidad conjunta puede ser calculada como:

$$P(X_1, X_2, \dots \dots, X_n) = \prod_{t=1}^{n} P\left(X_t | X_1, X_2, \dots \dots X_{t-1}\right) \quad [6]$$

Los asertos de independencia condicional junto con las tablas de probabilidad condicional nos permiten obtener la tabla de probabilidad conjunta de todas las variables a partir de las tablas de probabilidad condicional de cada variable en función de sus padres; de esta forma, aplicando la regla de la cadena conjuntamente con la propiedad de independencia condicional se obtiene:

$$P(X_1, X_2, \dots \dots X_n) = \prod_{t=1}^{n} P(X_t | X_1, X_2, \dots \dots, X_{t-1}) = \prod_{t=1}^{n} P\left(X_t | \prod X)_t\right) \quad [7]$$

El siguiente ejemplo muestra el proceso para calcular la probabilidad conjunta de varias variables conocida la estructura gráfica de la red y sus respectivas probabilidades condicionales en función de sus padres:

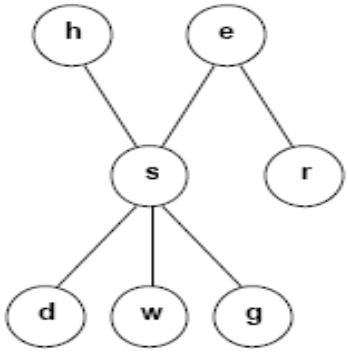


Figura 9 Red bayesiana que contiene el conjunto de nodos {d, e, h, r, s, w, g} **(Díez, 1993)**

Aplicando la regla de la cadena y que cada nodo es independiente de sus predecesores conocidos sus padres se obtiene que:

$P(h,e,r,s,d,w,g) = P(h)P(e|h)P(r|h,e)P(s|h,e,r)P(d|h,e,r,s)P(w|h,e,r,s,d)$ ...

$P(h,e,r,s,d,w,g) = \cdots P(g|h,e,r,s,d,w)$

$P(h,e,r,s,d,w,g) = P(h)P(e)P(r|e)P(s|h,e)P(d|s)P(w|s)P(g|s)$ [8]

La expresión anterior calcula la probabilidad conjunta de todos los nodos que componen la red a partir de las probabilidades condicionales de cada nodo en función de sus nodos padres. Dichas independencias condicionales son importantes porque simplifican la representación del conocimiento (menos parámetros) y el proceso de razonamiento o inferencia (propagación de probabilidades).

Separación. La variable *Z separa las variables X e Y si estas dos ultimas son* condicionalmente independientes dada *Z.*

La propiedad fundamental de una red bayesiana es la separación direccional (llamada d-separation), que se define así:

Separación direccional. Dado un grafo dirigido acìclico conexo y una distribución de probabilidad sobre sus variables, se dice que hay separación direccional si, dado un nodo X, el conjunto de sus padres, pa(X), separa condicionalmente este nodo de todo otro subconjunto Y en que no haya descendientes de X. Es decir:

$\mathrm{P}(\mathrm{x}/\mathrm{pa}(\mathrm{x}),\bar{\mathrm{y}}) = \mathrm{P}(\mathrm{x}/\mathrm{pa}(\mathrm{x}))$ [9]

2.4.5 Ejemplo de red Bayesiana

Con el siguiente ejemplo (Pearl, 1990) se ilustra mejor todo lo descrito anteriormente, tenemos dos vecinos en la casa, Juan y María, que han prometido avisar a la policía si oyen la alarma antirrobo instalada en una casa.

Juan y María podrían no llamar aunque la alarma sonara: por tener música muy alta en su casa, o confundirla con el teléfono, por ejemplo Incluso podrían llamar aunque no hubiera sonado: por confusión por ejemplo la alarma salta normalmente con la presencia de ladrones Pero también cuando ocurren pequeños temblores de tierra.

En este contexto se puede expresar esta situación mediante la red bayesiana, con las probabilidades expresadas.

En el ejemplo, no hay relación directa en el hecho que Juan o María llamen, y el que se produzca el temblor. También se puede decir que robo y temblor son causas directas de que se active la alarma y causas de que llamen los vecinos, pero no quiere decir que estos detecten el temblor o el robo directamente.

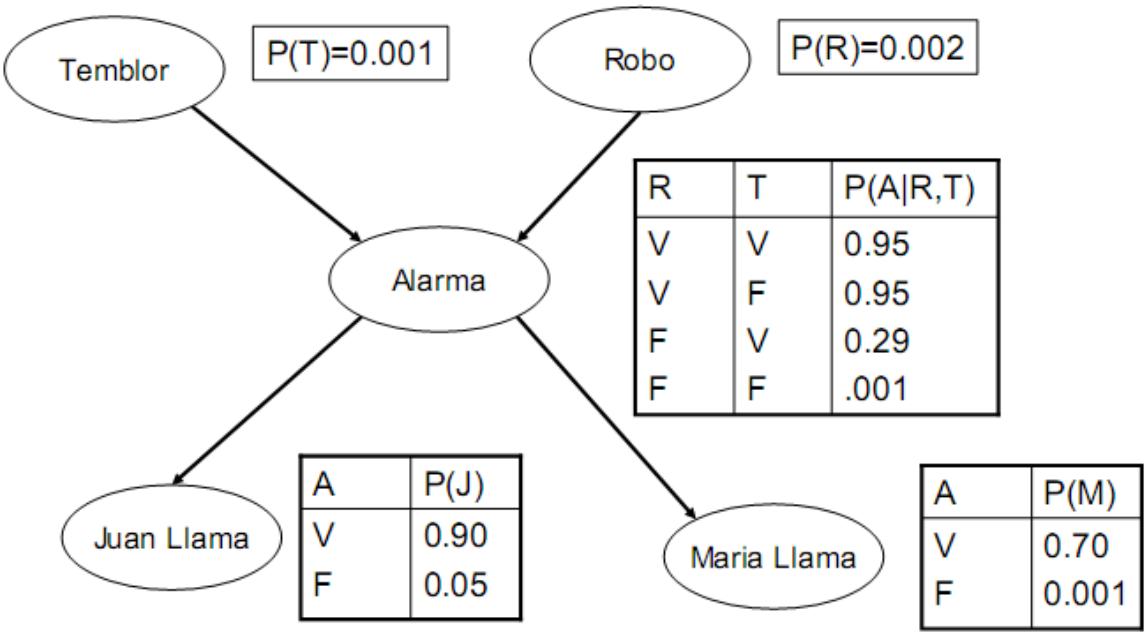


Figura 10 Representación de una situación por medio de red Bayesiana **(Russell & Norvig, 2004)**

En la red no se hace referencia directa, por ejemplo a las causas de que Juan o María ano escuchen la alarma, estas están implícitas en las tablas de probabilidades.

2.4.6 Construcción de una red bayesiana

Para construir la estructura de una red bayesiana se parte de unos datos, proporcionados por el experto, muchas veces estos datos son crudos y corresponde al ingeniero hacer un análisis previo para poder procesarlos o pre-procesarlos.

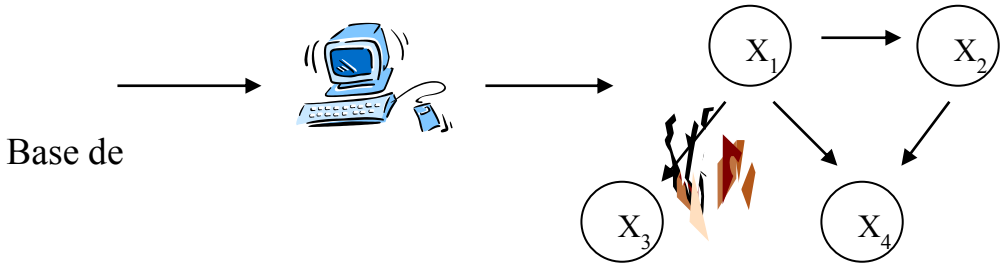


Figura 11 Tratamiento de las redes Bayesianas (Carrillo Calvet)

Para la construcción de una red bayesiana, es necesario realizar varias tareas hasta conseguir una estructura final dispuesta a funcionar dentro del sistema experto. Actualmente existen dos formas de construir una red bayesianas, que se pueden denominar como automática y manual (36); también se podría incluir como otro proceso de construcción la combinación de ambos tipos.

El proceso manual, construye la red bayesiana a partir de la ayuda de un experto humano que conozca a fondo el problema que se quiere modelar. Así a través de esta ayuda, el ingeniero del conocimiento, establecerá primero la estructura de la red causal (fase cualitativa), y posteriormente añadirá las probabilidades condicionales (fase cuantitativa) de los nodos creados. La construcción manual de la red bayesiana será la estructura central del diseño de la aplicación desarrollada, por eso se detallarán en profundidad los pasos posteriormente.

El proceso automático consiste en tomar una base de datos en la que todas las variables que nos interesas estén representadas y que contenga un número de casos suficientemente grande. Aplicando entonces alguno de los algoritmos que se han desarrollado recientemente para esta tarea, se obtienen los enlaces y las probabilidades condicionales que definen la red bayesiana. Sin embargo, en muchos problemas reales, es muy difícil contar con una base de datos suficientemente grande y detallada para la construcción de la red. De la construcción automática se hablará brevemente de los algoritmos que se pueden aplicar para dar una ligera idea de las posibilidades en el desarrollo, ya que estos algoritmos se apoyan en un desarrollo matemático bastante complejo para poder contemplarlos en el presente proyecto.

Ambas formas, aunque de distinta manera, implican en su proceso de construcción básicamente tres tareas:

1. Identificar las variables y sus valores.

2. Identificar las relaciones entre las variables, completando la definición del grafo que representa el modelo.

3. Obtener las probabilidades asociadas a cada nodo del grafo.

2.4.6.1 Construcción manual

La construcción manual de modelos gráficos probabilísticos no es una tarea trivial y no existen unos criterios definidos que se puedan aplicar constantemente ante cualquier problema. Aún así, en esta tarea se cuenta normalmente con la experiencia previa del experto humano, si existiese en su caso, para aventurar las relaciones entre las variables; pero normalmente el sentido común y la experiencia propia del ingeniero del conocimiento serán los aspectos más importantes a la hora de la construcción manual de una red bayesiana.

Los elementos implicados en la construcción manual de redes bayesianas así como las distintas fases necesarias que se describirán a continuación están basados en la estructura presentada en la Tesis Doctoral, Explicación en redes bayesianas causales.

2.4.6.2 Construcción automática

El conjunto de técnicas que se emplean para la construcción de las redes bayesianas de manera automática, son conocidas como algoritmos de aprendizaje, que permitirán extraer toda la información necesaria.

Según (Pearl, 2009) pionero en redes bayesianas, existen dos fases de aprendizaje, cuando se trabajan con redes bayesianas, que se puede denominar respectivamente aprendizaje estructural y aprendizaje paramétrico.

Estas dos fases se pueden resumir como:

Aprendizaje paramétrico: a partir de la estructura de la red, se obtiene las probabilidades a priori de los nodos raíz y las probabilidades condicionales de las demás variables requeridas a través del uso de bases de datos.

Aprendizaje estructural: obtener la estructura de la red bayesiana a partir de bases de datos, obteniendo las relaciones de dependencia e independencia entre las variables existentes. Los algoritmos que aprenden la estructura de la red bayesiana se engloban generalmente dentro de una de las categorías siguientes

- Algoritmos que se basan en un procedimiento que busca la mejor estructura en el espacio de posibles soluciones, midiendo la calidad de cada red candidata mediante funciones de evaluación. Estos son algoritmos que se caracterizan por el tipo de función y por el procedimiento de búsqueda.
- Algoritmos basados en detección de independencias, que toman como entrada el conjunto de relaciones de independencia condicional y generan la red que mejor representan estas relaciones.
- Algoritmos híbridos que se basan en la combinación de ambas metodologías.

En el aprendizaje de redes bayesianas, será casi requisito principal para poder realizar la tarea de aprendizaje a partir de datos, disponer de bases de datos muy extensas en las que estén especificado el valor de cada variable en cada uno de los casos.

La combinación de ambas posibilidades, permite orientar al experto y al ingeniero del conocimiento para afianzar o corregir su percepción del dominio. Se puede optar por obtener el modelo de forma manual, a través de la ayuda de expertos humanos y aplicar alguno de los algoritmos de aprendizaje para la obtención de las probabilidades. Por otro lado, también se puede aprender la red a partir de una base de datos y posteriormente realizar una depuración refinando la estructura y los parámetros con la ayuda de expertos humanos.

2.4.7 Inferencia

El problema de la inferencia en redes Bayesianas se plantea como la obtención de la distribución a posteriori de ciertas variables de la red dado que se ha observado el valor que toman otras variables. Esto es lo que se conoce como “propagación de probabilidades”. Diversos métodos de propagación han sido desarrollados. Sin embargo, se ha demostrado que este es un problema NP-duro (COOPER) y por lo tanto intratable a nivel computacional para redes de tamaño suficientemente grande (Robles V, 2003).

A partir de una red ya construida, y dados los valores concretos de algunas variables de una instancia, podrían tratar de estimarse los valores de otras variables de la misma instancia aplicando razonamiento probabilístico.

Este razonamiento probabilístico sobre las redes bayesianas consiste en propagar los efectos de las evidencias (variables conocidas) a través de la red para conocer las probabilidades a posteriori de las variables desconocidas. De esta forma se

puede determinar un valor estimado para dichas variables en función de los valores de probabilidad obtenidos.

En general, una red puede usarse para calcular la distribución de probabilidad para cualquier subconjunto de variables dados los valores de cualquier subconjunto de las restantes.

Además de estimar la probabilidad de ciertos eventos (la variable de consulta), las Redes bayesianas permiten:

- Estimar que variables de evidencia hay que observar para obtener información útil.
- Hacer análisis de sensibilidad: determinar que variables tienen más peso en las probabilidades de las variables consultadas.
- Explicar al usuario los resultados de una inferencia probabilista.

Tipos de evidencia.

- Evidencia dura (hard). Conocimiento determinista: P(A)=1 ó P(A)=0. Al asignar evidencia dura al nodo se le llama instanciación.
- Evidencia parcial (soft). Conocimiento probabilístico (distinto a 0 y a 1). Incluye a las probabilidades a priori y a las actualizadas tras instanciarse alguna variable.

2.4.8 Ventajas de las redes bayesianas

- Ganancia de conocimiento en un dominio y sus relaciones causales.
- Relaciones condicionales claramente delimitadas por los arcos del grafo.
- La Red Bayesiana es una buena representación modular y gráfica del conocimiento
- Solo pocas probabilidades conjuntas necesitan calcularse
- Las Redes Bayesianas son intuitivas: de fácil interpretación
- Utilización de un algoritmo de aprendizaje distribuido localmente
- Una representación bien elaborada puede disminuir el número de parámetros exponencialmente
- La relación entre todas las variables: permite inferir valores que faltan
- Permite manejar evidencias sin el problema del crecimiento exponencial.

2.4.9 Desventajas de las redes bayesianas

- La escogencia inicial de las variables no es una tarea simple

- La definición de la estructura do modelo no es evidente
- No es fácil saber cuál es la mejor red, tampoco las mejores variables
- Es preciso conocer los valores a priori iníciales
- Las relaciones causa efecto no son siempre evidentes

2.4.10 Aplicaciones de las redes bayesianas

- Diagnóstico de cáncer de seno.
- Diagnóstico de cáncer cérvico -uterino.
- Evaluación del potencial de marcadores genéticos para el diagnóstico y diferenciación de tipos de cáncer.
- Evaluación de trayectorias escolares de alumnos universitarios.

3 CASO DE ESTUDIO

Para extraer interacciones de proteínas de los textos científicos se puede usar únicamente un sistema IR, pero por sí sólo no arroja más que una lista de documentos que contienen la palabra de consulta, sin dar un valor agregado a la información. Para la proteína de interés EVI1 los sistemas IR recuperan todos los artículos que contengan el nombre de la proteína, en este caso las publicaciones sobre EVI1 han sido pocas, concentrándose la gran mayoría desde los años 1990 hasta el 2000.

En la Figura 10 se puede ver cuánto ha evolucionado el número de artículos publicados acerca de EVI1.

Entre estas herramientas se probaron algunas que permiten obtener un ranking de los temas de consulta, por ejemplo Medie, MedlineRanker y Pescador, realizando unas pruebas para recuperar artículos específicos para un posterior estudio.

El siguiente paso de las herramientas de minería de texto para extraer información de interacciones consiste en identificar las entidades, esta etapa sirvió como filtro para identificar las herramientas que posteriormente serían objeto de estudio para la evaluación.

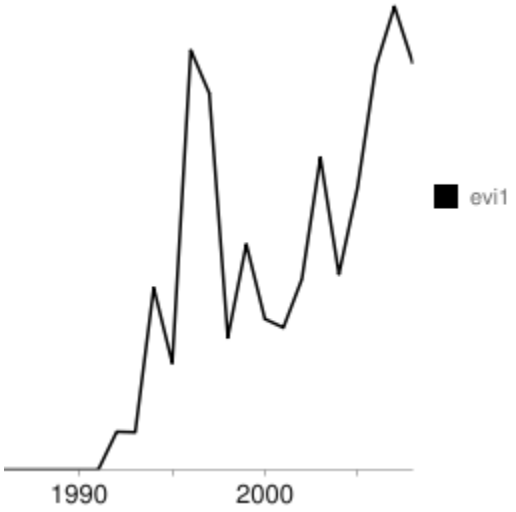


Figura 12 Figura 10 Articulos publicados con el tópico EVi1. fuente Mltrends.

Las herramientas que usan la metodología ER, se basan en las RI, y por ello proporcionan un valor agregado al manejo de información, estas herramientas identificaron entidades del tipo de Evi1, y con este resultado se puede obtener un grupo de proteínas que posiblemente interactúen con EVI1, estas herramienta son de gran ayuda para el científico quien intuitivamente puede inferir de los resultados nuevas interacciones o proponer hipótesis, pero la información obtenida no es suficiente y se necesita de una herramienta más compleja que permita obtener mejor desempeño.

Aunque en esta investigación se hallaron herramientas interesantes intermedias entre estas tecnologías, las cuales hacen un ranking y por metodologías de clúster se entrenan con los mismos datos y muestran un listado de palabras claves fuertemente relacionada en los textos y algunas de ellas, incluso separan proteínas, de lo cual se consiguen resultados interesantes al mezclarlas con cualquiera de las anteriormente mencionadas. Después de estos pasos se procedió a usar las herramientas que integraban más funcionalidades para la posterior evaluación. En el apéndice, está la lista de herramientas más conocidas y en la siguiente tabla las que se estudiaron.

3.1 ANÁLISIS Y CLASIFICACIÓN DE LAS HERRAMIENTAS

HERRAMIENTA	Metodologías	Sitio
STRING	Integration	http://string-db.org/
POLYSEARCH	Integration	wishart.biology.ualberta.ca/**polysearch**
IHOP	ER, IE	www.**ihop**-net.org/
PPFINDER	ER, IE	http://liweilab.genetics.ac.cn/tm/
FACTA	ER, IE	http://text0.mib.man.ac.uk/software/facta/a.cgi
NOVOSEEK	ER, IE	http://www.novoseek.com
GENIA	ER,	http://text0.mib.man.ac.uk/software/geniatagger/
MedLineRanker	ER, Ranker	http://cbdm.mdc-berlin.de/tools/medlineranker/
MINOTAUR	ER,IR	www.bioinf.manchester.ac.uk/dbbrowser/minotaur/index.html
MEDIE	ER	http://www-tsujii.is.s.u-tokyo.ac.jp/medie/
ALIBABA	ER	http://alibaba.informatik.hu-berlin.de/
PUBGENE	Integration	www.**pubgene**.org/
LITINSPECTOR	Integration	http://www.litinspector.org/
PESCADOR	Analizador	http://cbdm.mdc-berlin.de/~pescador/index.php

Tabla 4 *Herramientas de minería de texto utilizadas*

3.1.1 STRING *(*Search Tool for the Retrieval of Interacting Genes/Proteins*)*

La base de datos cubre actualmente 2.590.259 proteínas a partir de 630,

STRING lee automáticamente los abstracts de pubmed y otras fuentes, y también información curada que se ha introducido en las bases de datos, donde curadores expertos han computarizado textos biológicos.

Método:

Las interacciones incluyen asociaciones (funcionales) directas (físico) e indirectas. Se derivan a partir de cuatro fuentes:

- Contexto genómico

- Experimentos de alto rendimiento de procesamiento
- Co-expresión(conservado)
- organismos Conocimiento previo.

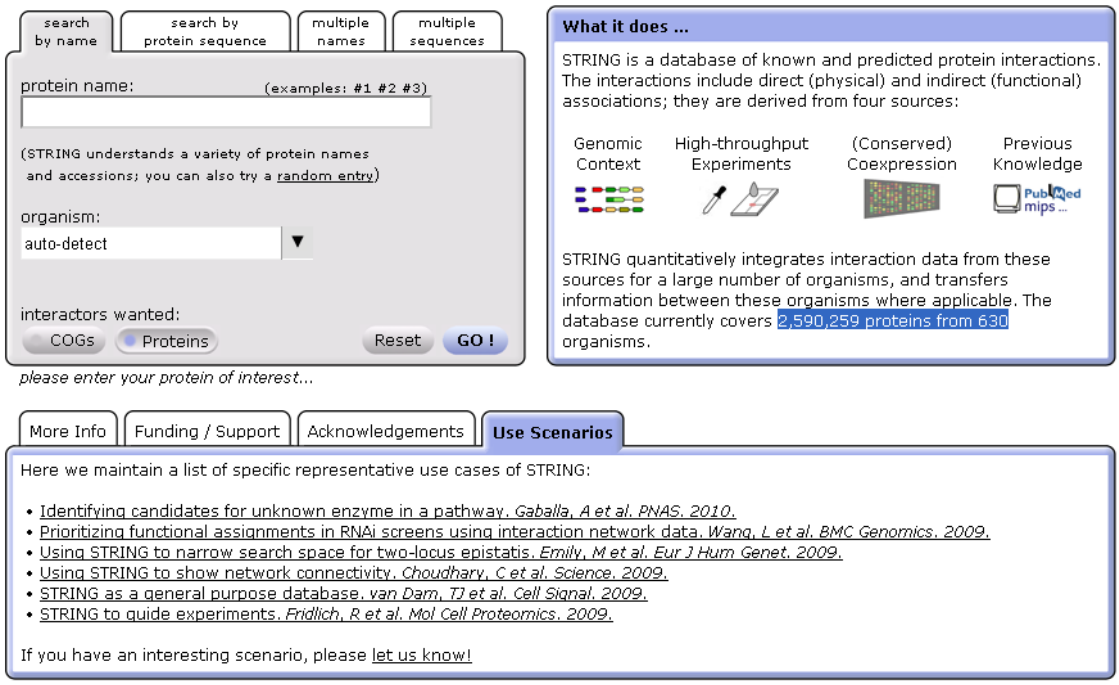


Figura 13 Formato consulta STRING.

En el primer cuadro se hacen las consultas, sean estas simples o múltiples, esta última opción proporciona una ventaja para los investigadores, porque pueden entrar a consulta parejas de proteínas o incluso listas para mostrar relaciones entre ellas y encontrar nodos escondidos o proteínas desconocidas que conecten a las de consulta, de las cuales no se tenía conocimiento, y de las que posteriormente se puede investigar ingresando a las resúmenes que muestra el modulo de minería de texto de STRING. Los restantes cuadros muestran una breve información de la herramienta, cómo son fuentes de información, colaboradores, publicaciones acerca de STRING, etc.

STRING permite hacer consulta por organismo, esto es muy importante, pues muchas herramientas trabajan con organismos más simples de los que se les conoce el genoma completamente. El organismo de estudio es el hombre por lo que se selecciona en el menú inicial Homo Sapiens como lo muestra la figura.

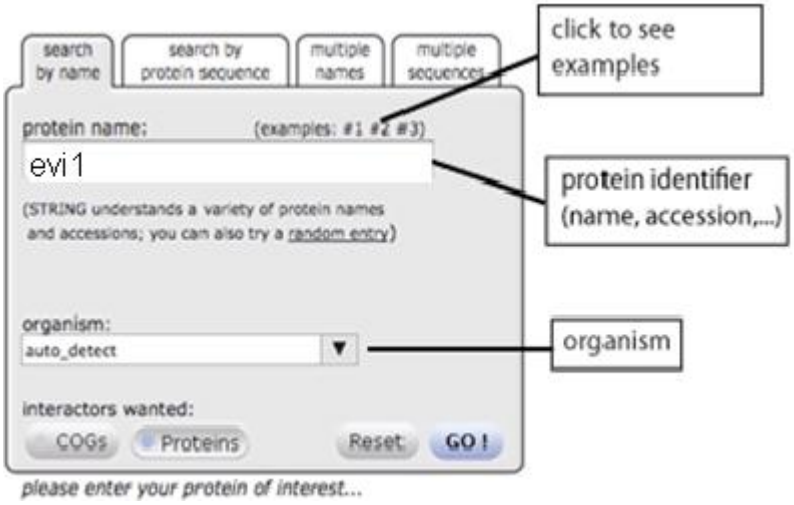


Figura 14 Consulta EVI1 en STRING

Seguida la consulta se genera una red de las interacciones que más conocidas, que han sido curadas y que más se mencionan en la literatura científica. El resultado lo muestra en tres partes: un grafo, el resumen de predicción y los parámetros.

Interpretación de la salida:

En la figura 18 se muestra la red y sus diferentes interacciones que arroja la búsqueda por medio de la herramienta STRING

Estas líneas representan la existencia de los siete tipos de evidencia usados en predecir las asociaciones.

- La línea roja indica la presencia de evidencia de la fusión.
- La línea Verde evidencia la vecindad.
- La línea azul evidencia la coocurrencia.
- La línea púrpura evidencia experimental.
- La línea amarilla evidencia text mining.
- La línea azul clara evidencia de la base de datos.
- La línea negra evidencia del coexpression

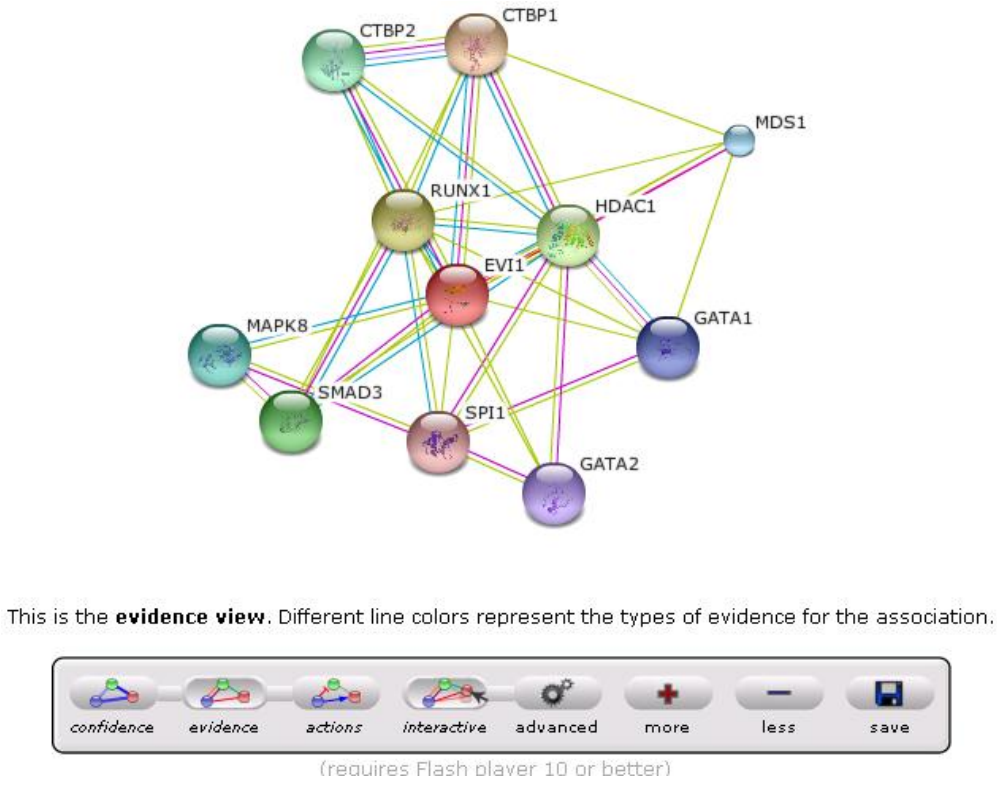


Figura 15 Gráfico de la red de la proteína evi1.

STRING usa un score combinado de todos los parámetros evidenciados en cada asociación predicha, este valor se halla calculando la probabilidad conjunta de las probabilidades de todos los canales de diferentes pruebas, bien sean estas puntuaciones que se han asignado a vecindades, Text mining, co-ocurrencia, etc.

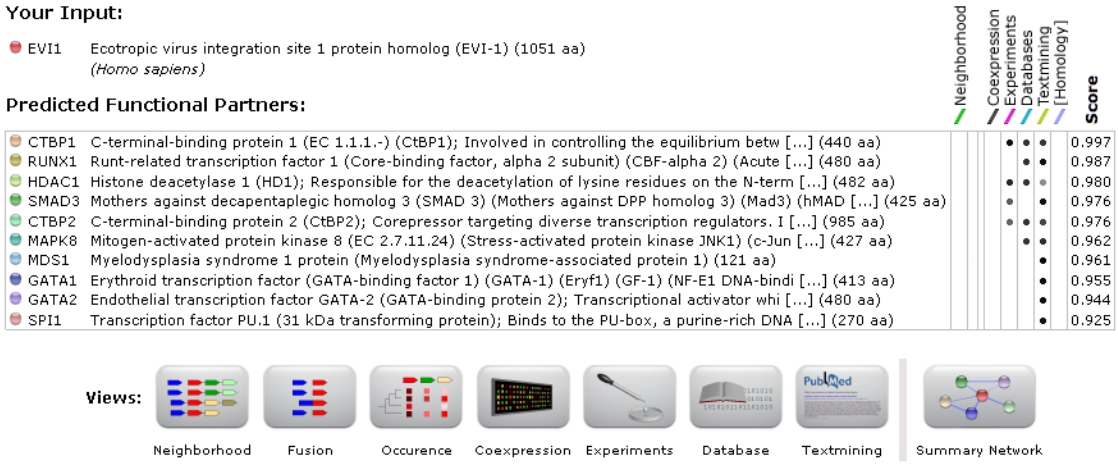


Figura 16 Clasificación de los score de la proteína evi-1

STRING identifica las proteínas y palabras claves que guarda en un diccionario, así en el texto va identificando relaciones. STRING contiene información derivada de la literatura de dos tipos: información derivada de manera automática, y la que se ha introducido manualmente. Este último es importado de diferentes bases de datos externas, donde curadores expertos han leído y computarizado los textos de biología. Estos datos son accesibles en STRING en la vistas de base de datos y experimentos que están debajo del resumen de predicción en el mismo navegador donde está la opción text mining.

Sin embargo, debido a la enorme cantidad de conocimientos biológicos publicados (incluidos los textos más antiguos), procesamiento de text mining accede a una gran cantidad de conocimientos, pero también es muy difícil para el software de supervisión de STRING no cometer errores en este proceso.

El modulo de text mining usa un sistema RE para etiquetar las palabras claves en especial las proteínas.

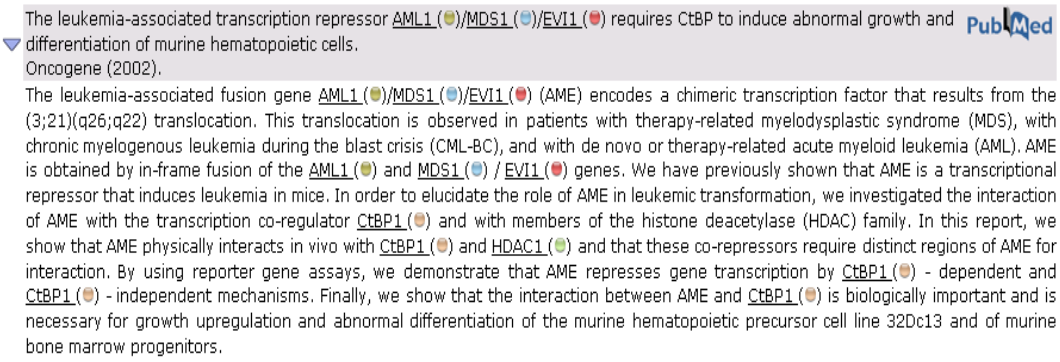


Figura 17 Muestra del modulo de minería de texto que usa STRING.

Ventajas:

- Es libremente accesible y se pone al día regularmente
- Es una base de datos de las interacciones sabidas y previstas de la proteína.
- Muestra resultados en una tabla descargable que ayudara en la construcción de la matriz de adyacencia.

Desventajas:

- Podría presentarse sinónimos de los genes, lo cual haría más difícil abarcar en totalidad la existencia de todas las interacciones.
- Debido a la alta cantidad de conocimiento biológico el procesamiento de text mninig accede a una gran errores principalmente se generan por falsos positivos y negativos falsos.

3.1.2 IHOP (información Hyperlinked sobre las proteínas)

El uso de genes y proteínas como hipervínculos entre oraciones y resúmenes de la información en PubMed se puede convertir en un recurso navegable. IHOP (Información sobre Hipervínculos de Proteínas) es una línea servicio que ofrece una red de genes guiada como una forma natural de el acceso a millones de resúmenes en PubMed y trae todas las ventajas de Internet para investigar la literatura científica. Navegando a través de oraciones relacionadas entre sí dentro de esta red se está más cerca de la intuición humana que el uso de búsquedas de palabras clave convencionales y permite paso a paso la adquisición de control de la información. Por otra parte, la literatura en la red puede ser superpuesta a los datos experimentales para la interacción y facilitar el análisis simultáneo del conocimiento nuevo y el conocimiento existente.

La red que presenta en la actualidad IHOP contiene más de cinco millones de frases relacionadas con 40000 genes del Homo sapiens, Mus musculus, Drosophila melanogaster, Caenorhabditis elegans, Danio rerio, Arabidopsis thaliana, Saccharomyces cerevisiae y Escherichia coli.

IHOP construye una red del gene convirtiendo la información en MEDLINE en un recurso navegable usando genes y proteínas como enlaces de hipertexto entre las oraciones y los resúmenes.

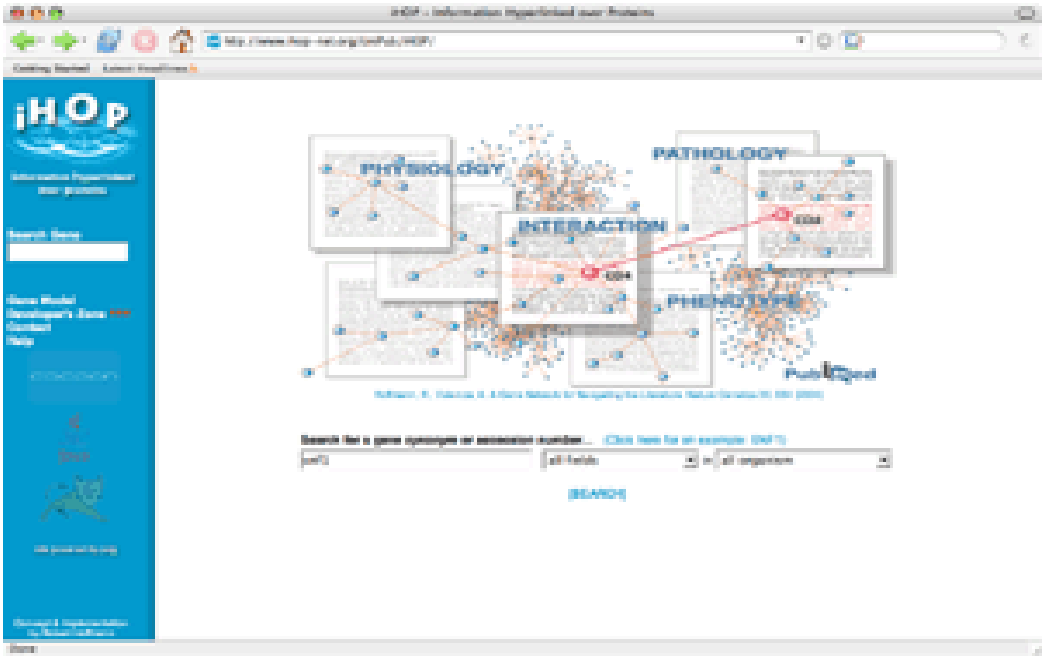


Figura 18 Visualización de la herramienta IHOP

METODO

Figura 19 Respuesta de IHOP a la consulta Evi 1.

Las oraciones interesantes se pueden recoger en un modelo del gene. Por ejemplo, al analizar las oraciones para el gene A, puede ser que parezca una asociación interesante o inesperada con el gene B. La oración encontrada mencionará un aspecto. La lista de interacciones sacadas de esta consulta se ilustran en el anezo B.

Interpretación de la salida

El modelo del gene almacena estas oraciones y representa los resultados de su investigación. Además, todas las asociaciones entre las oraciones recogidas se representan como grafo.

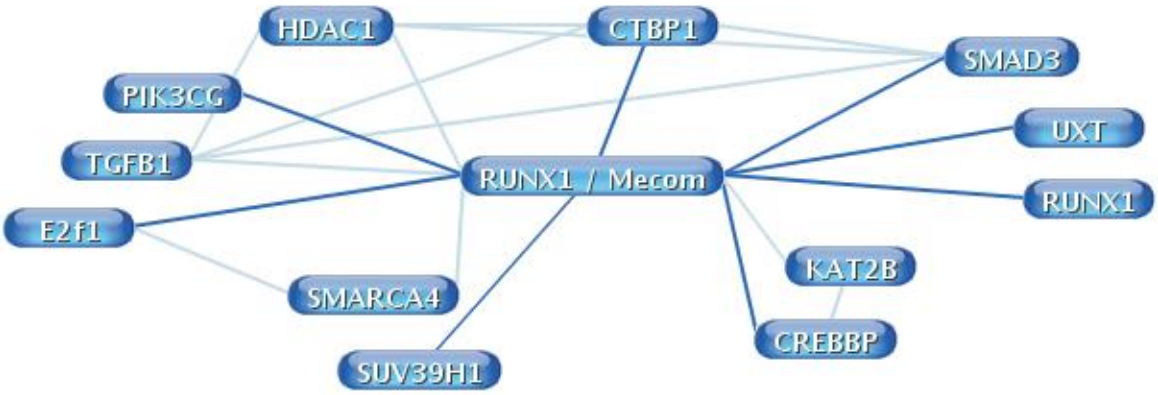


Figura 20 Visualización de las interacciones de evi-1 por IHOP

Ventajas

- IHOP tiene la Ventaja sobre representaciones puramente gráficas en que los investigadores pueden guardar control permanente sobre el origen y la confiabilidad de la información, puesto que se mueven entre las oraciones a partir de la fuente original.
- A pesar de mostrar errores de nombres, ambigüedad y otros, la información obtenida se puede considerar bastante ilustrativa en cuanto a interacciones encontradas.
- Busca en los textos desde la estructura misma de las sentencias u oraciones mencionando aspectos de la relación entre el gen A y el gen B, pero no necesariamente el más importante.
- Todas las asociaciones entre las oraciones recopiladas se representan como gráfico. Busca en los textos desde la estructura misma de las sentencias u oraciones y menciona un aspecto de la relación entre el gen A y el gen B, pero no necesariamente la más importante.
- Presenta una información de carácter general, como el símbolo, el nombre y el organismo de un gene.

Desventajas

- Dificultades para resolver ambigüedades, los sistemas automáticos, como iHOP, por lo tanto exhibirán siempre ciertos errores.

- Los expertos reconocerán símbolos incorrectos simplemente identificados explorando una oración, pero el sistema muestra errores, no es capaz de reconocer estos símbolos, además lleva un proceso muy manual para agregar todos los abstracts que muestren las interacciones pertinentes.
- Los expertos deben reconocer símbolos incorrectos al explorar una oración.
- Lleva un proceso muy manual para agregar todos los abstracts que muestren las interacciones pertinentes.

3.1.3 PUBGENE

PubGene está diseñado para presentar información sobre los genes, las proteínas y las palabras claves relacionadas en una forma organizada e intuitiva.

Las etapas que trabaja pubgene para averiguar sobre un gen o proteína son:

1. Consulta el nombre del gen o la palabra clave.
2. Enlaza "la literatura de redes" que muestra los genes cocitados.
3. Referencias cruzadas de los se incorporan en la lista los sinónimos.
4. Recupera fuentes de resúmenes;
5. Anota redes con ontología, temas médicos o términos químicos.
6. Recupera términos.

Pub Gene cataloga no sólo genes individuales, sino pares de genes. Es decir, las listas de registros que PubGene coidentifica citando genes. La Concitación sugiere una relación biológica entre los genes implicados.

Método:

PubGene utiliza una interfaz gráfica de acceso a bases de datos que cataloga la ocurrencia de los genes y frases de la identificación de genes en la literatura científica. Bases de datos se actualizan cada dos semanas a través de búsquedas en MEDLINE. Las búsquedas se realizan utilizando una lista de sinónimos para cada gen que incluyen el gen principal de identificación (ID) (que se enumeran en Entrez Gene) y otros símbolos y frases que se usan para identificar el gen y su producto proteínico. Por otra parte, el interfaz del grafo puede presentar una relación de artículos, lo que permite al investigador buscar todos los registros que citan el gen.

PubGene crea herramientas para los investigadores, los médicos y el público en general. El Gráfico de redes proporcionan una visión general de cómo los conceptos se relacionan entre sí. Estas redes son un producto de algoritmos patentados, que constantemente busca textos científicos y datos de la secuencia para extraer la información pertinente a determinados conceptos biomédicos. Los

resultados son una referencia cruzada para construir bases de datos que le permiten visualizar de forma rápida y navegar hacia la información que se necesita.

PubGene cataloga no sólo genes individuales, sino pares de genes. Es decir, las listas de registros que PubGene coidentifica. La co-citación sugiere una relación biológica entre los genes implicados por ejemplo (Nature Genetics 2001 28:21-8, PMID 11326270). Revista, fecha y PMID.

Descubrimiento de las relaciones entre los genes

PubGene utiliza cocitación para crear redes de identificadores de genes, lo que permite la posibilidad de descubrir relaciones entre dos genes a través de un gen de intermediación. Además de las listas de artículos, PubGene puede buscar por palabras claves utilizando identificadores de genes, o usar palabras claves para encontrar genes (MeSH, GO y los identificadores químicos). Múltiple identificadores de genes se pueden presentar en un momento para revisar las relaciones entre estos.

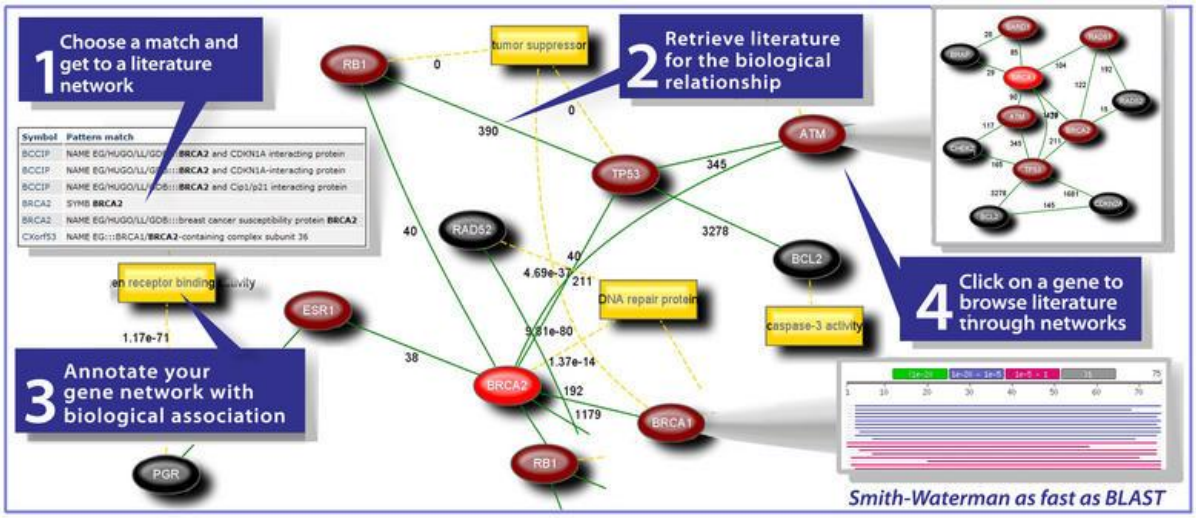


Figura 21 Las diferentes etapas de PubGene para mostrar las características de los genes o proteínas.

Relaciones entre los genes

PubGene utiliza cocitación para crear redes de identificadores de genes, lo que permite la posibilidad de que el descubrimiento de las relaciones entre dos genes a través de un gen de intermediación. Además de las listas de artículos, PubGene puede buscar por palabras clave utilizando identificadores de genes, o usar palabras clave para encontrar genes (MeSH, GO y los identificadores químicos).

Múltiple identificadores de genes se pueden presentar en un momento para revisar las relaciones entre estos.

- Las búsquedas se realizan utilizando una lista de sinónimos para cada gen que incluyen el principal gen identificación ID (que se enumeran en Entrez Gene) y otros símbolos y frases que se usan para identificar el gen.

-Presenta la co-citacion sugiriendo una relación biológica entre los genes implicados.

-Además de las listas de artículos, PubGene puede buscar por palabras claves utilizando identificadores de genes, o usando palabras claves para encontrar genes (MeSH, GO y los identificadores químicos.

Los resultados pueden verse en el anexo F, y en las figura 20 y 21.

Interpretación de la salida

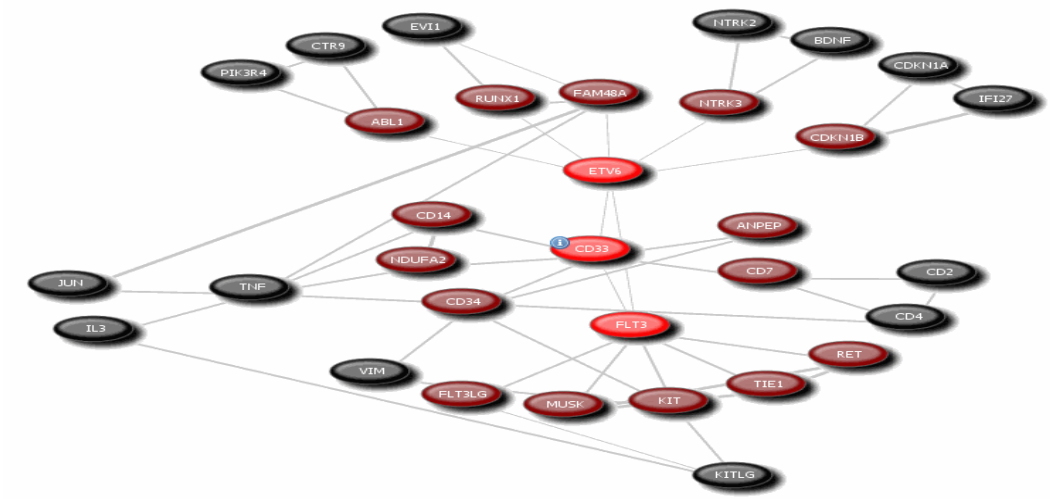


Figura 22 Red de la proteína de evi1 para la herramienta pubgene.

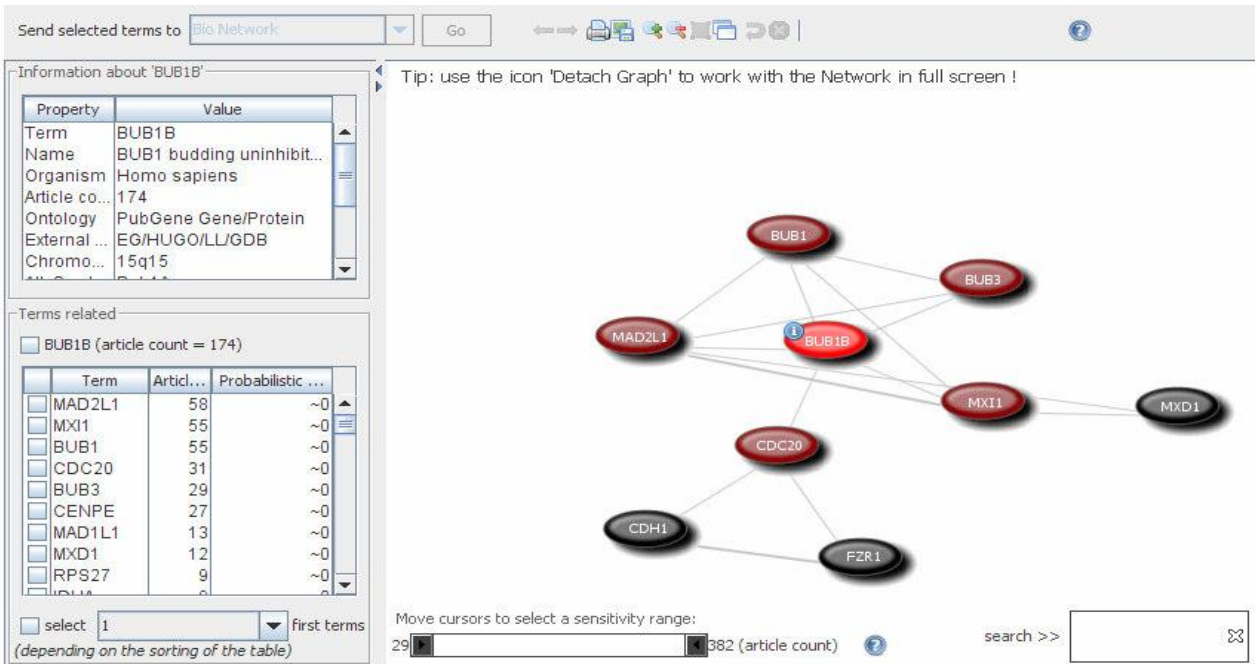


Figura 23 Figura 21 Resultado de la consulta de pubgene

La información textual y la red:

Información básica sobre el gen se muestra en la parte superior Izquierda de la pantalla. Este cuadro es un cuadro de desplazamiento en dos columnas tituladas como propiedad y el valor. La información en la columna de la Propiedad describe la información indicada en la columna de valor. Estos tipos son:

- Termino: reconocido símbolo gen o una palabra clave correspondiente al término de la consulta (en algunos casos traducida del texto de la consulta).
- Nombre: nombre asociado con el término.
- Número de artículos: el número de registros de Medline, que contiene el término de consulta o uno de sus sinónimos al menos una vez (devuelve 0 en la secuencia de modo).
- Factores externos DB: bases de datos utilizadas para decidir el símbolo oficial, el nombre y comunes sinónimos del término (bases de datos HUGO= Organización del Genoma Humano, EG = Entrez Gene, el BGF = Base de datos del genoma).
- Vecinos: número de vecinos literatura (genes o proteínas cocitada al menos una vez con el término de consulta) o la secuencia vecinos (los genes o las proteínas cuya secuencia similitud al término de la consulta está por encima de corte).

La lista de área de trabajo del gen se presenta en una tabla estructurada en 7 columnas. La primera columna contiene el símbolo de la identificación del gen. A la izquierda de cada gen es un símbolo de casilla de verificación que permite al usuario seleccionar uno o varios de los genes. Los genes seleccionados se puede utilizar en una consulta a cualquier de las funciones PubGene utilizando los botones de la radio y la barra Enviar por encima de la tabla. Las otras columnas muestran el nombre completo, Organismos HOMOLOGENE, el número de artículos citando el gen, número de vecinos (otros genes que aparecen en los artículos junto con el gen), la Variante símbolos (sinónimos) para designar el gen y la opción de eliminar la entrada de genes de la lista.

Ventajas

- Posibilita la búsqueda de más interacciones por nodos estratégicos.
- PubGene provee información de investigaciones actualizadas y completas, sobre cada uno de los genes, sus proteínas y sus relaciones con otros genes.
- El gráfico interfaz puede presentar una relación de artículos, lo que permite al investigador buscar todos los registros que citan el gen.
- Cerca de diez mil nuevos registros son añadidos cada semana
- Posibilita la búsqueda de más interacciones por nodos estratégicos.

Desventajas

- Las búsquedas son confundidas por el hecho de que muchos genes son conocidos por varios sinónimos, así como algunos sinónimos son ambiguos asignado a varios genes.
- Las búsquedas son confundidas por el hecho de que muchos genes son conocidos por varios sinónimos, así como sinónimos ambiguos asignado a varios genes.
- Existen diferentes nomenclaturas para las proteínas. Los genes entre especies no es coherente, por ejemplo, los genes estudiados en C. elegans puede no ser fácilmente reconocido por los identificadores de sus equivalentes humanos.

3.1.4 LITINSPECTOR

Herramienta de búsqueda automática de genes y vías de transducción de señales en la minería de datos del NCBI PubMed

Método

Permite la entrada de un gen, texto libre, sinónimo, palabra o frase; la consulta puede ser filtrada para sólo los resúmenes para los que también se definen las categorías de palabras claves (de tejidos, enfermedad, o vía) los que fueron

identificados. LitInspector permite la entrada sinónimos de genes y números de identificación de genes o texto libre. Al menos un gen o de texto libre sinónimo palabra o frase que se ha presentado. La consulta, además, puede ser filtrado para sólo los resúmenes para los que también se definen las categorías de palabras clave (de tejidos, la enfermedad, o vía) fueron identificados.

Figura 24 Abctracs etiquetados por Litinspector

Interpretación de salida

Ref.	Pathway Component	Signaling Pathway
14	TGF BETA	tgf beta signaling (BioCarta STKE KEGG)
9	SMAD	mothers against dpp homolog signaling (BioCarta STKE KEGG)
2	JNK	jun n terminal kinase signaling (BioCarta STKE STKE KEGG)
1	BMP	tgf beta signaling (BioCarta STKE KEGG)
1	ERBB2	epidermal growth factorreceptor family member erbb2 (her 2/neu) signaling (BioCarta BioCarta STKE)
1	FAK	focal adhesion kinase 1 signaling (KEGG)
1	JAK	janus kinase signaling (STKE KEGG)
1	PI3K	phosphatidylinositol signaling (BioCarta STKE KEGG)
1	STAT	signal transducer and activator of transcription signaling (BioCarta STKE KEGG)
1	TGF BETA RECEPTOR	tgf beta signaling (BioCarta STKE KEGG)
1	WNT	wingless type signaling (BioCarta STKE KEGG)

Figura 1 *Tabla de las interacciones con la herramienta litinspector*

Ventajas

- Presenta un solo sinónimo automáticamente considerando todos los sinónimos de este gen.
- Este software tiene una mayor exactitud a diferencia de IHOP que tiene muchos términos ambiguos.
- Una de las ventajas de búsqueda de LitInspector es la presentación de un solo sinónimo automáticamente considerando todos los sinónimos de este gen.

- Este software tiene una mayor exactitud a diferencia de IHOP que tiene muchos términos ambiguos.

Desventajas

- PubGene proporciona pocos ejemplos de documentos para cada búsqueda.
- La evaluación de resultados completos no es posible.
- Una evaluación de resultados completos no es posible. Esta evaluación debiera hacerse usando 10 resúmenes disponibles.

3.1.5 ALIBABA

Es una herramienta de búsqueda de las entidades s biológicas y sus relaciones en resúmenes de PubMed. Visualiza los resultados a través de redes gráficas o grafos. La información que proporciona es relacionada principalmente con interacciones de proteína, las relaciones de la enfermedad y la especificidad de tejido de genes. Es capaz de incluir las vías de KEGG y permite realizar búsquedas utilizando los identificadores de UniProt. Encuentra rápidamente información sobre las proteínas que interactúan, los genes con implicaciones en las enfermedades, etc. Ali Baba permite buscar las proteínas con sólo preguntar por ID UniProt en lugar de escribir una larga lista de sinónimos. Ali Baba es capaz de incluir las vías de KEGG, y espera mejorar agregando nuevas bases de datos en un futuro próximo. # Ali Baba enlaces toda la información a la literatura de base y bases de datos - esto le proporciona información detallada sobre determinados aspectos.

Interpretación de salida

Al buscar información utilizando motores de búsqueda convencionales, por ejemplo, PubMed, los usuarios ven los datos y un resumen a la vez y "oculto" con el texto en lenguaje natural. Alibaba es una herramienta interactiva para la representación gráfica de los resúmenes en los resultados de búsqueda.

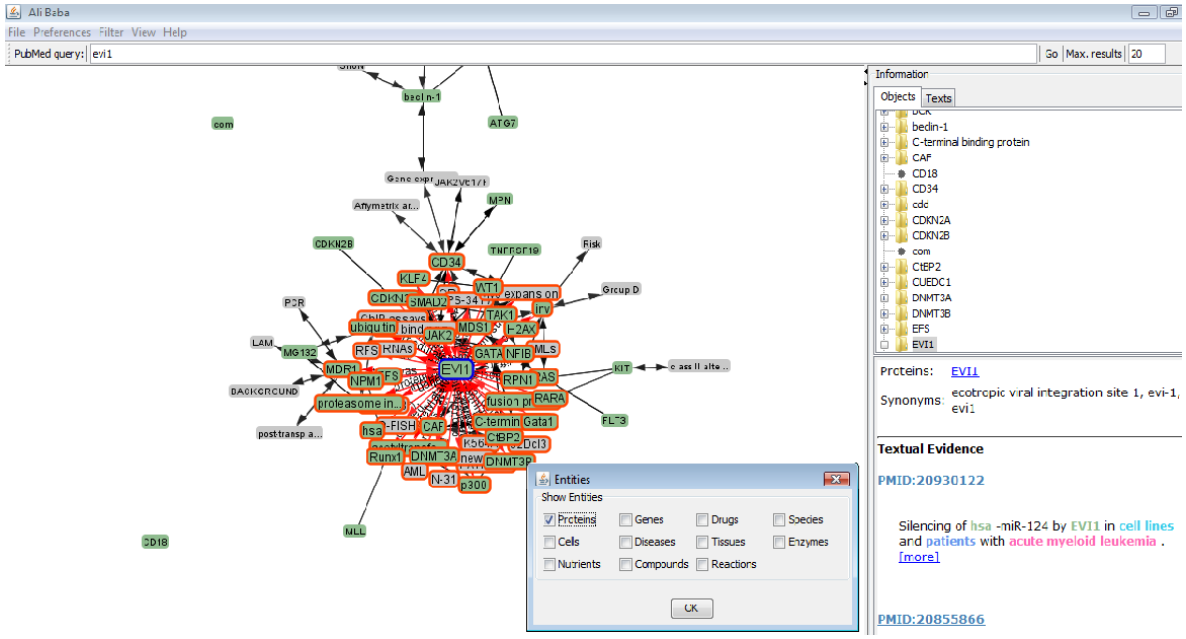


Figura 25 Resultados consulta Alibaba

Se analiza el conjunto de resúmenes que se ajusten a una consulta EVI1 de PubMed y presenta la información extraída de los objetos biomédicos y sus relaciones como una red gráfica. Alibaba extrae de las asociaciones entre las células, las enfermedades, las drogas, las proteínas, las especies y los tejidos. Varias opciones de filtro permiten una búsqueda más específica. Por lo tanto, los investigadores pueden captar redes complejas que se describen en varios artículos.

Ventajas

- Este software construido en JAVA se puede descargar de la red.
- Además de redes de proteínas involucra otro tipo de redes, como de enfermedades asociadas, genes, drogas asociadas en los artículos para tratar las enfermedades.
- Permite parametrizar la búsqueda

Desventajas

- Presenta ambigüedades, y muchos datos faltantes.

3.1.6 FACTA

FACTA: Finding Associated Concepts with Text Analysis. Es un motor de búsqueda asociado con herramientas de análisis de textos publicados en Medline para encontrar las asociaciones de una consulta de atención a conceptos como la co-ocurrencia de genes y proteínas, las enfermedades, los síntomas, medicamentos y compuestos basados en el análisis de co-ocurrencia. Proporciona en una pantalla los conceptos que co-ocurren en los documentos. Al hacer la consulta por EVI1 se presentaron estos resultados del anexo e y en la siguiente figura.

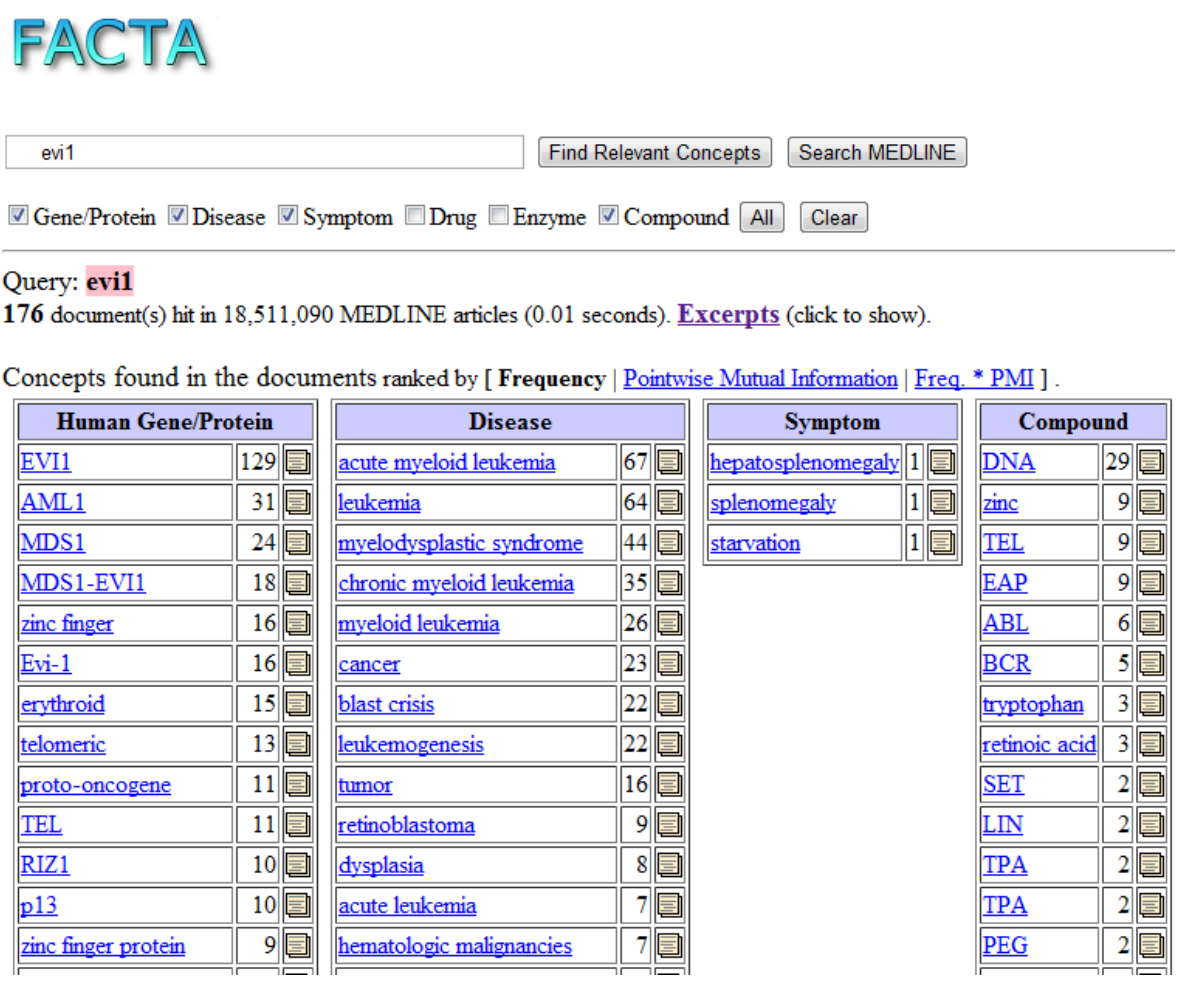


Figura 26 Resultados Facta

Los conceptos se presentan al usuario en un formato tabular y clasifican con base a las estadísticas de co-ocurrencia. A diferencia de los sistemas existentes que proporcionan una funcionalidad similar, FACTA no sólo muestra índice de las

palabras sino también los conceptos mencionados en los documentos, que permite al usuario emitir una consulta flexible (por ejemplo, palabras clave o combinaciones booleanas libre de palabras claves) y recibir los resultados de inmediato, aun cuando el número de los documentos que coinciden con la consulta sea muy grande. El usuario también puede ver fragmentos de MEDLINE para obtener evidencia textual de las asociaciones entre los términos de la consulta y los conceptos etiquetados.

Desventaja

- Presenta relaciones sin especificar si son directas o no.
- Los resultados no son descargables en texto plano.

3.1.7 POLYSEARCH

Permite la consulta de proteínas asociadas con enfermedades y otras entidades, haciendo una búsqueda de términos pares, por ejemplo, genes contra genes, genes contra enfermedades como se muestra en la figura.

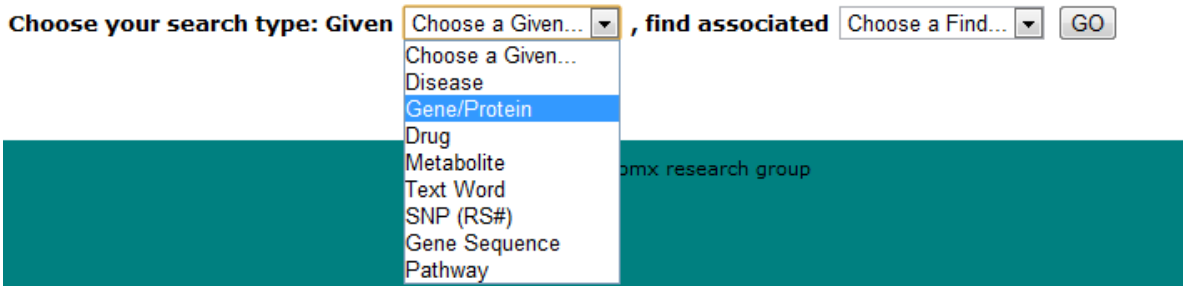


Figura 27 Formato Consulta Fase I Polysearch

Dado que muchas consultas en la genómica, la proteómica o metabolómica implican este tipo de búsquedas exhaustivas, PolySearch soporta más de 50 diferentes clases de consultas en casi una docena de tipos diferentes de texto, bases de datos científicos abstractos o bioinformática. La consulta típica con el

apoyo de PolySearch es "Dada X, encontrar todos los Y's", donde X o Y pueden ser las enfermedades, los tejidos, los compartimentos de células, genes / proteínas nombres, SNPs, mutaciones de las drogas y metabolitos.

PolySearch también explota una variedad de técnicas de minería de texto y recuperación de información para identificar, destacar y clasificar resúmenes informativos, párrafos o frases como puede verse en la siguiente figura.

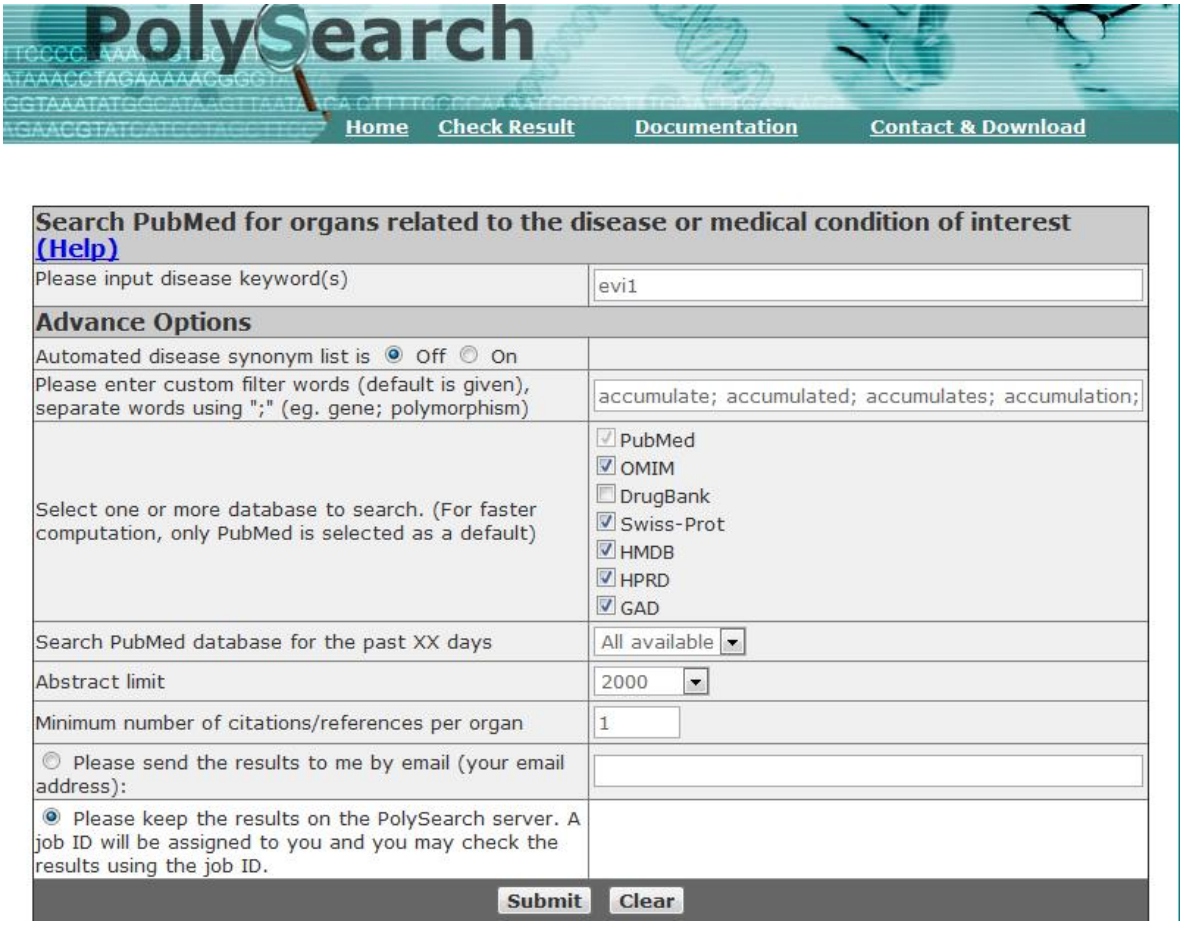


Figura 28 Formato consulta Polysearch Fase II

Polysearch usa minería de texto basado en diccionarios y sinónimos, los cuales son adecuadamente normalizados y mantiene nueve diferentes listas tesauros o sinónimo de genes humanos, proteínas humanas, enfermedades humanas, los medicamentos aprobados, metabolitos, proteínas y vías de genes, tejidos humanos, órganos humanos y la localización subcelular. Estos tesauros o

compendios son, obviamente, fundamentales para muchas de las consultas amplias.

Z Score	Relevancy Score	Gene/Protein Name	Synonyms	PubMed Hits	OMIM Hits
6	560 (0,0,4,92)	RUNT RELATED TRANSCRIPTION FACTOR 1	AML 1; AML1; AML1 EVI 1; AML1c; AMLCR 1; AMLCR1; Acute myeloid leukemia 1 protein; CBF alpha 2...	–	5 (0,0,4,92)
4.9	465 (0,2,11,18)	MDS1	MDS 1; MDS1; MDS1 EVI1; MDS1/AML1 fusion gene; Myelodysplasia syndrome 1 protein; Myelodysplasia syndrome associated protein 1; PRDM 3; PRDM3...	–	6 (0,2,11,18)
3.3	325 (1,2,2,15)	PRDM16	MDS1/EVI1like gene 1; MEL 1; MEL1; PFM13; PR domain containing 16; PR domain zinc finger protein 16; PR domain containing protein 16...	–	2 (1,2,2,15)
2.4	245 (0,0,0,49)	CGD	CGD; CGD91 PHOX; CYBB; Cytochrome B 245 heavy chain; Cytochrome B(558) beta chain; Cytochrome b(558) subunit beta; Cytochrome b 245 beta polypeptide; Cytochrome b558 subunit beta...	–	1 (0,0,0,49)
2	215 (0,0,7,8)	EAP	APA; EAP; Aminopeptidase A; CD249; CD249 antigen; Differentiation antigen gp160; ENPEP; Glutamyl aminopeptidase...	–	3 (0,0,7,8)
1.6	180 (1,1,1,1)	inv	INV; INVS; Inversin; Inversion of embryo turning homolog; NPH 2; NPH2; NPHP 2; NPHP2...	–	1 (1,1,1,1)
1.6	180 (1,1,1,1)	AP1	AP1; Activator protein 1; JUN; Proto oncogene c jun; Protooncogene c jun; Transcription factor AP 1; V jun avian sarcoma virus 17 oncogene homolog; c Jun...	–	1 (1,1,1,1)
1.6	180 (1,1,1,1)	erythropoietin	EP; EPO; Epoetin; Erythropoietin; Erythropoietin precursor; Epoetins; Erythropoietins; Erythropoietin precursors	–	1 (1,1,1,1)
1.3	150 (0,0,5,5)	ECOTROPIC VIRAL INTEGRATION SITE 1	AML1 EVI 1; MDS1 EVI1; PRDM 3; PRDM3; EVI 1; EVI1; Ecotropic viral integration site1; Ecotropic virus integration 1 site protein...	–	2 (0,0,5,5)

Figura 29 Resultados Polysearch

3.1.8 NOVOSEEK

Novoseek es un motor de búsqueda biomédica para la búsqueda de conocimientos publicados en la literatura biomédica. Novasen indexa literatura biomédica desde Pubmed, libres documentos completos, publicaciones de revistas de acceso abierto en PubMedCentral. Y otras fuentes, utilizando una tecnología de minería de texto que permite la identificación de los principales términos biomédicos. Para realizar esta identificación ambigua, Novoseek tiene en cuenta datos externos disponibles, así como información contextual. Como resultado de esta tecnología de indexación, novoseek es capaz de recuperar todos los documentos donde se menciona un término no importa qué sinónimos se utiliza y desecha los documentos donde se utiliza el término con un sentido no deseado.

Como anteriores herramientas de este estudio Novoseek extrae información de múltiples conceptos biomédicos clave, no importa si son las enfermedades, medicamentos, compuestos químicos, los síntomas o los genes.

Hace un etiquetado de los conceptos biomédicos en el texto y filtra los resultados de una manera rápida y fácil y accede a los documentos.

Los resultados obtenidos en la consulta hecha en esta herramienta se muestran en el anexo C y en la siguiente figura:

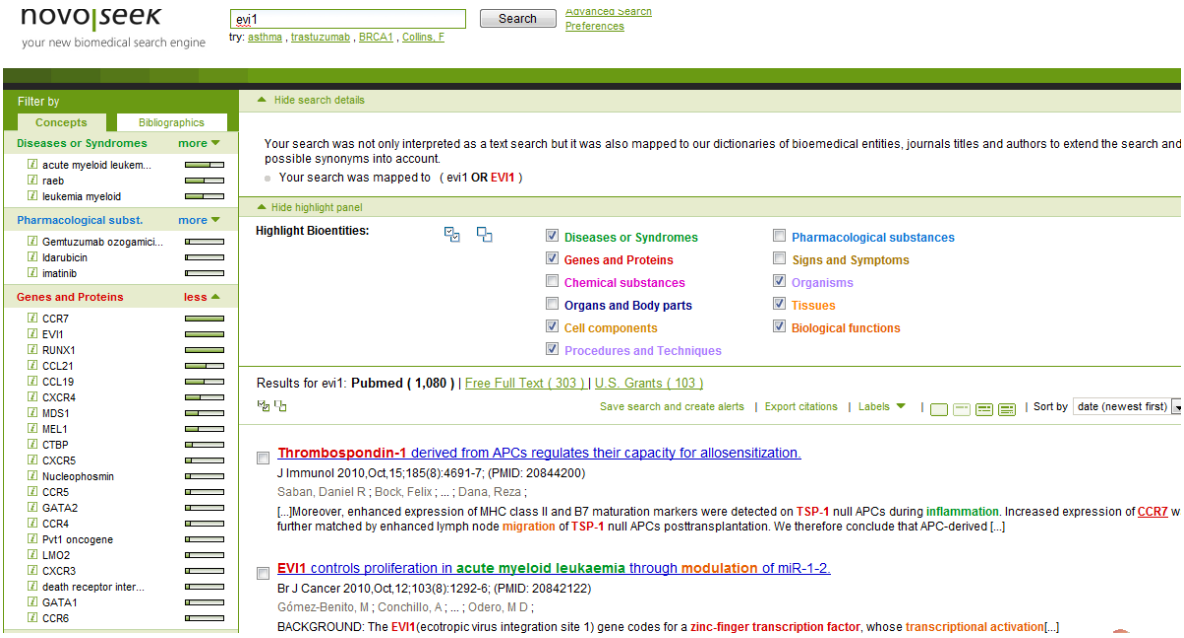


Figura 30 Resultados Novoseek

3.1.9 PPFINDER

Es una herramienta basada en web, que extrae interacciones entre proteínas de humanos de los resúmenes en PubMed con base a su co-ocurrencias.

El objeto de estudio de este trabajo es construir una red de interacciones para la proteína EVI1, utilizando diferentes herramientas que usan los enfoques antes mencionados, para un mejor desempeño de ellas se combinaran algunas.

Es evidente que el simple trabajo de identificar y extraer documentos relacionados que contengan la entidad EVI1 lleva una complejidad inherente, la cual va aumentando en cada escaño del proceso, porque además de que se debe reconocer las entidad EVI1, se suma el trabajo de reconocer otras entidades del mismo tipo, en este caso “proteínas” y extraer múltiples tipos de relaciones de todas estas entidades.

Como se indicó anteriormente, la minería de texto se basa en filosofías de IE, IR y EI para extraer relaciones o interacciones, muchas herramientas de estudio

cumplen con estas filosofías, pero otras deben integrarse para lograr el resultado deseado.

El primer paso relacionado con el pre-proceso sugiere una búsqueda de la proteína en cuestión en los artículos científicos almacenados en un repositorio, grandes bases de datos, u otras herramientas capaces de conectar con varios repositorios lo cual generará resultados más óptimos.

Todas las herramientas de este estudio realizan sus búsquedas en diferentes repositorios responsables de los datos aunque el usuario visualice un sólo repositorio porque existen procesos independientes de sincronización e integración de los datos.

3.2 RESULTADOS

En esta etapa se realizó el procedimiento de búsqueda de las herramientas que permitieran extraer el conocimiento de interacciones con la proteína de estudio. Se analizaron diversas herramientas tanto de IE, como de minería de texto, y se escogieron para un posterior estudio las que proporcionaban un número de interacciones considerable, es importante destacar que muchas herramientas investigadas aportan una precisión importante, es decir buena calidad de información, pero también se necesita una cantidad importante, un equilibrio entre calidad y cantidad, pues con poca información es difícil predecir o comprender un fenómeno, se necesita un número considerable de información siendo consciente de que se encontraran muchos falsos positivos o errores de otro tipo, pero en ésta área tan difícil de predecir es vital tener material para procesar.
La tabla 5 muestra los resultados de las consultas, presentando el número de interacciones halladas, estas interacciones pueden ser de dos tipos directas e indirectas, los resultados que se muestran fueron los que obtuvieron los mejores índices en la evaluación, al escoger la combinación de parámetros y umbrales adecuados.

El criterio de evaluación se baso en una serie de métricas que se confrontaron con un patrón oro, proporcionado por un conjunto de proteínas seleccionada por 94 artículos que el grupo IMM escogió para el estudio ya que es información que en la mayoría de los casos ha sido curada, es importante resaltar que este patrón oro se ajusta a las necesidades o intereses del experto y que en este orden de ideas es posible que muchas herramientas con reconocida trayectoria no cumplan con los requerimientos para el estudio pertinente, y también una colección de artículos puede restringir la evaluación, y generar más datos faltantes para confrontar, y efectivamente el patrón de comparación mostraba no contenía muchos de los resultados que daban las herramientas, lo que origina un sobreajuste en los datos.

HERRAMIENTA	Interacciones	observaciones
STRING	54	ID, muestra grafo, etiqueta texto, metainformación
POLYSEARCH	32	II, ID, consulta por enfermedad, etiqueta texto
IHOP	26	ID, se usaron distintos grupos de 50 artículos, etiqueta textos
PPFINDER	21	II,
FACTA	24	II, consulta por enfermedad, etiqueta textos
GOPUBMED	48	II, etiqueta textos, se complementa con GOgene, metainformación
GOGENE	48	Se complementó con GOpubmed, meta información
NOVOSEEK	25	II, etiqueta texto, consulta por enfermedad
ALIBABA	41	ID, etiqueta texto, muestra grafo, meta información
PUBGENE	25	ID,II, muestra Grafo, metainformación
LITINSPECTOR	28	II,ID, etiqueta Texto

Tabla 5. *Comparación herramientas*

Muchas Herramientas proporcionaron una larga lista de proteínas que aparecían en la consulta con Evi1, una del orden de 600 proteínas, daba por hecho que Para el interés de estudio proporcionaba muchos datos irrelevantes porque el patrón fue de 70 proteínas. En la red de Evi1 hay mucho por descubrir, pero se necesita un espacio de búsqueda más pequeño para el coste de algoritmos de aprendizaje que se puedan implementar. Se evaluaron 10 herramientas con la combinación de parámetros que proporcionaron los mejores resultados en cada una de ellas, también se usaron dos herramientas alternas que se integraron a algunas de las evaluadas obtener mejores resultados.

HERRAMIENTA	Artículos Recuperados	FP	FN	TP	TN	Recall	Precisión	F-Mesure
LITINSPECTOR	>50	5	12	16	14	0,57	0,76	0,653061
ALIBABA	50	22	5	22	11	0,81	0,50	0,619718
POLYSEARCH	100	12	4	23	6	0,85	0,66	0,741935
STRING	>50	13	4	32	5	0,89	0,71	0,790123
IHOP	400	7	13	17	9	0,57	0,71	0,62963
PUBGENE	130	12	6	13	6	0,68	0,52	0,590909
PPFINDER		7	5	13	13	0,72	0,65	0,684211
NOVOSEEK	1077	5	15	14	15	0,48	0,74	0,583333
FACTA	176	7	15	17	8	0,53	0,71	0,607143

Tabla 6 *Evaluación de herramientas*

3.2.1 Análisis de resultados

La tabla muestra los mejores resultados de cada herramienta escogiendo la combinación de parámetros y umbrales que proporcionen mejores resultados. Como el resultado que se obtuvo de cada herramienta fue heterogéneo en cuanto a la cantidad de interacciones extraídas, se trató de mantener un tamaño equivalente para comparar con el patrón establecido, para tratar de cubrir todos los ejemplos positivos y ajustar los datos faltantes que en el caso del patrón establecido fue muy distinto a los resultados que por aparte cada herramienta establece en sus evaluaciones. Por ejemplo muchas herramientas reportan precisiones del orden del 90 %, pero se debe tener en cuenta que para el estudio de ellas los patrones de comparación que se utilizaron eran más amplios.

Sistemas presentados por otros grupos de rendimiento de precisiones más del 90%, pero que necesitan los grupos de patrones muy grande. Pues en el estudio pertinente el patrón para cada herramienta tiende a cubrir muy pocos ejemplos. Para un mejor perfonmance se debe ajustar los patrones de comparación con un aumento en los artículos investigados por los expertos. Sin embargo este ejercicio podría afectar sensiblemente el índice de recuperación,

Para obtener otra aproximación, se usaron herramientas alternas que hacían un ranking de los documento s escaneados y escogían por validación cruzada y métodos de clúster un nuevo conjunto de entidades para el nuevo análisis, mejorando en algunas herramientas su desempeño.

Sin embargo la herramienta STRING se considero la más completa porque continuó con los mejores resultados ante las combinaciones de las otras herramientas con las herramientas alternas que se usaron.

No obstante a lo anterior se presume que los patrones que se establecieron con un número determinado de artículos pueden ser optimizados con el uso de las mismas herramientas que se usaron para ranking, pues se debe tener en cuenta que fueron hechos manualmente lo que implica un alto costo de tiempo y un margen de error más grande.

Para este evento se usaron las herramientas medline Ranker y pescador, se encontraron nuevas relaciones que permitieron ampliar el patrón de comparación al ingresar los PIMD que proporcionaba la primera lista a estas herramientas se obtuvieron los siguientes resultados.

HERRAMIENTA	Artículos Recuperados	FP	FN	TP	TN	Recall	Precisión	F-Mesure
LITINSPECTOR	>50	3	10	22	12	0,69	0,88	0,77193
ALIBABA	50	15	7	26	12	0,79	0,63	0,702703
POLYSEARCH	100	11	4	24	6	0,86	0,69	0,761905
STRING	>50	11	4	34	5	0,89	0,76	0,819277
IHOP	400	7	13	19	9	0,59	0,73	0,655172
PUBGENE	130	12	6	13	6	0,68	0,52	0,590909
PPFINDER		7	5	14	12	0,74	0,67	0,7
NOVOSEEK	1077	5	15	20	10	0,57	0,80	0,666667
FACTA	176	6	15	18	8	0,55	0,75	0,631579

Tabla 7. *Resultados al mejorar el patrón de comparación.*

El resultado anterior evidencia la importancia de los patrones de comparación, mientras más minuciosa sea la tarea para establecerlos mejores resultados se obtienen lo que permite mejores ajustes y bases para establecer nuevas hipótesis.

El uso de estas métricas de rendimiento estándar se estableció por los investigadores para ajustar las medidas a cualquier sistema, pero se debe tener en cuenta la complejidad inmersa en los sistemas biológicos por los problemas señalados en este trabajo para la extracción de interacciones. Para ajustar el sistema a cualquier medida. La mayoría de las aplicaciones necesitaría de alta precisión. No es el caso de los sistemas biológicos, por eso se necesitan analizar otras medidas. Si se obtiene un índice de recuperación alto quiere decir que la predicción de falsos positivos de las interacciones puede ser identificado y resuelto por los expertos.

De los anteriores resultados confrontados con el patrón oro proporcionado por los expertos se destacaron las cuatro primeras, pero el análisis debe ser más profundo.

En cuanto a resultados que permitan inferir nuevas relaciones se han clasificado algunas sentencias, extraidas de diferentes herramientas de minería de texto, de

las investigadas se encontró que muchas usa sistemas ER, entre ellas, Ihop, novoseek, Alibaba, String y Litinspector.

Se ha escogido como ejemplo algunas sentencias extraídas de los Abtracs para ilustrar algunas relaciones que pueden ayudar a el planteamiento de nuevas hipótesis, a continuación el resultado recuperado por STRING.

3.2.2 Sentencias con información de interraciones

In this report, we show that EVI1 ( ) physically interacts with BRG1 ( )-J Biol Chem (2003).

Ectopic expression of BRG1 ( ) is able to repress the E2F1 ( ) -J Biol Chem (2003).

Overexpression of Evi-1 ( ) increased Pbx1 ( ) expression in hematopoietic stem/progenitor cells. -Oncogene (2009).

An analysis of the Pbx1 ( ) promoter region revealed that Evi-1 ( ) upregulates Pbx1 ( ) transcription. -Oncogene (2009).

In this report, we show that AML1 ( )/MDS1 ( )/EVI1 ( )physically interacts in vivo with CtBP1 ( ) and HDAC1- Oncogene (2002).

Evi-1 ( ) represses TGF-beta signaling by direct interaction with Smad3 ( ) -Blood (2001).

Here we show that Evi1 ( ) interacts with the histone methyltransferase SUV39H1 ( ).-FEBS Lett (2008).

Evi1 ( ) also interacts with another histone methyltransferase, G9a.- FEBS Lett (2008).

A unique AML1 ( ) (CBF2A) rearrangement, t(1;21)(p32;q22), observed in a patient with acute myelomonocytic leukemia. -Cancer Genet Cytogenet (2001).

we subsequently showed that EVI1 ( ) interaction with GATA1 ( ) blocks proper erythropoiesis.- Cancer Res (2009).

We report here that EVI1 ( ) interacts with PU.1 ( ) and represses the PU.1 ( )-Cancer Res (2009).

EVI1 ( ) and MDS1 ( ), in 3q26 are generated, resulting in the formation of a chimaeric transcription factor.- Pathol Biol (Paris) (1997).

AML1 ( )(RUNX1 ( )) can initiate a myelodysplastic syndrome (MDS) that progresses to acute myelogenous leukemia (AML) in association with overexpression of Evi1 ( ).-Blood (2008).

Fusion of ETV6 ( ) to MDS1 ( ) / EVI1 ( ) as a result of t(3;12)(q26;p13) in myeloproliferative disorders. - Cancer Res (1997).

MDS1 ( ) and EVI1 ( ) are rearranged by the t(3;21)(q26;q22) and by the t(3;12)(p13;q22). As a result of the translocation, they are expressed as fusion genes either with AML1 ( )or with TEL ( ). - Leukemia (1997).

The ETV6 ( ) / TEL ( ) gene has been reported to fuse to PDGFRbetab MDS1 ( ) / EVI1 ( ), BTL, ACS2, STL, JAK2, ABL, CDX2, TRKC, AML1 ( ), and MN1. - Blood (2000).

MDS1-EVI1 fusions and ETV6 ( ) - EVI1 ( ) fusions, respectively, occur. The Ribophorin I ( )-EVI1 fusion in particular may be a common occurrence in t(3;3).- Cancer Res (1997).

We identified a novelRUNX1 ( ) partner gene, MDS1 ( )/EVI1 ( )-like-gene 1 (PRDM16), in an AML patient with t(1;21).- Genes Chromosomes Cancer (2005).

Repression of RUNX1 ( ) activity by EVI1 ( )-Cancer Res (2007).

Our findings suggest that Evi-1 ( ) promotes hematopoietic stem/progenitor expansion at the embryonic stage through up-regulation of GATA-2 ( ) and repression of TGF-beta signaling.- Cancer Sci (2008).

AML1 ( ) - Evi1 ( ), AML1 ( )-MDS1-Evi1, AML1 ( ) - EAP fusion transcripts and Evi1 ( ) gene were detected in mRNA level, but no AML1 ( ) - Evi1 ( ) fusion transcript.- Cancer Sci (2008).

RUNX1 ( ) - EVI1 ( ) has oncogenic potentials through dominant - negative effect over wild-type RUNX1 ( ), inhibition of Jun kinase (JNK ( )) pathway, stimulation of cell growth via AP-1, suppression of TGF-beta - mediated growth inhibition and repression of C/EBPalpha.- Cancer Sci (2008).

but also networks connecting several genes with more than one fusion partner (e.g. ETV6 ( ) / RUNX1 ( ) (AML1), RUNX1 ( ) / CBFA2T1 (ETO), ETV6 ( ) / EVI1 ( ), RUNX1 ( ) / EVI1 ( ), ETV6 ( ) / ABL, BCR / ABL).- Cytogenet Cell Genet (2000).

CBFA2 ( ) forms a fusion gene with ETO and MDS1 ( ) / EVI1 ( ) in translocations in myeloid leukemia and with ETV6(TEL) in the t(12;21) common in childhood pre-B acute lymphoblastic leukemia.- blood(1998)

3.3 CONSTRUCCIÓN DE LA RED BAYESIANA

La mejor alternativa para el estudio fue STRING, *(Search Tool for the Retrieval of Interacting Genes/Proteins)* porque ofrecía una integración de varias metodologías, extrayendo información de los documentos que posteriormente convirtió en conocimiento y proporcionando un listado para ser analizado posteriormente por los expertos para formular nuevas hipótesis, ya sea intuitivamente con una revisión escueta o a través de un estudio concienzudo de las interacciones obtenidas.

Muchas de las interacciones que se hallan en la literatura están curadas por expertos, lo mismo sucede con las bases de datos de proteínas, pero como es sabido el genoma humano es complejo y toda esta información es cambiante. Hay relaciones que no están curadas y aparecen en la literatura esperando ser investigadas, y hay otras relaciones que pueden inferirse de un conjunto de interacciones publicadas, y es aquí donde las herramientas de minería de texto cumplen un papel destacado. Queda para estudio investigar sobre la veracidad de las interacciones reportadas por este experimento que no están aun aceptadas. Es muy probable que ya sean halladas y las herramientas de minería no las hayan detectado.

Pero además de las inferencias que se pueden con seguir por minería de texto, existe la posibilidad de mejorar las predicciones alimentando una red bayesiana con los datos obtenidos por este experimento. Es decir el primer paso de la

minería de texto en el que se centró este trabajo sirvió para introducir datos más estructurados y con alta confidencialidad para la construcción de una red bayesiana, que permite aun depurar los resultados a través de procesos de aprendizaje.

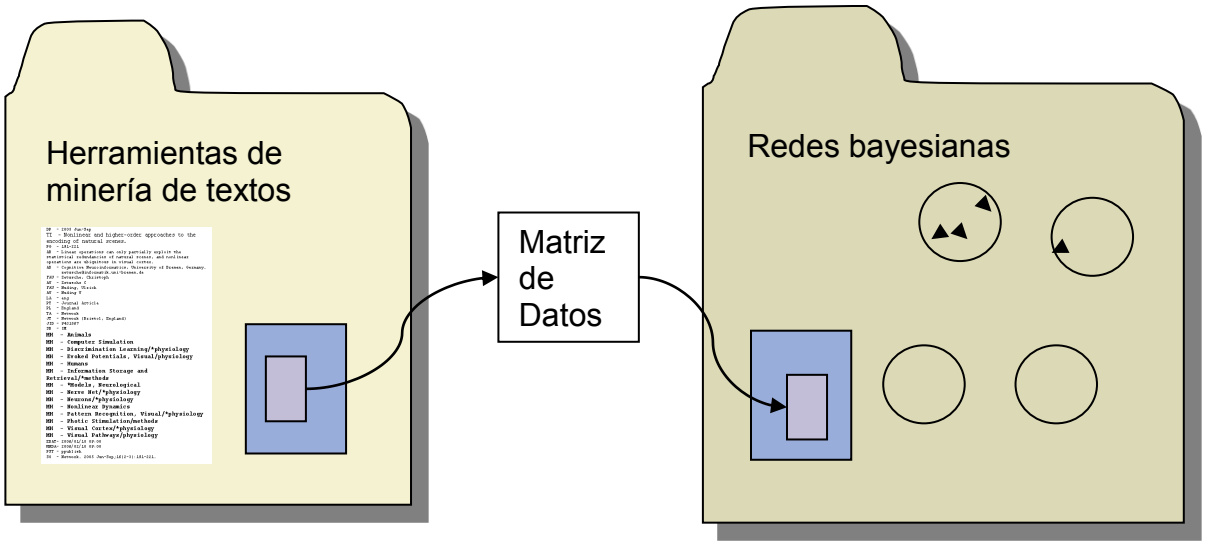


Figura 31 Proceso de construcción de la Red Bayesiana.

3.3.1 Modelamiento de la Red Bayesiana

Para crear la estructura de la red bayesiana es necesario extraer las relaciones de las interacciones. Después de analizar varias herramientas que especifican estas relaciones se escogido con ayuda del experto la herramienta STRING, porque además de extraer información de los abstracts que lee automáticamente de Pubmed y otras fuentes también la obtiene de diversas fuentes como son las bases de Datos científicas (detallando muchos aspectos implícitos en ellas que son de interés) y los experimentos reportados.

El paso a seguir será convertir esas relaciones en datos que puedan ser interpretados por la red bayesiana, por lo cual se hizo un algoritmo en Matlab que permitió deducir una matriz de adyacencia con los datos extraídos de la herramienta de minería de texto.

STRING arroja datos crudos de las interacciones de proteína y las almacena en un archivo plano que tiene la información almacenada en varias columnas, luego con ese archivo se delimitan los campos de interés que en este caso se reduce a los nombres de las proteínas relacionadas, luego se procesan esos datos para construir la matriz de adyacencia.

3.3.2 Construcción de la matriz de adyacencia

1. Se crea una matriz cero, cuyas columnas y filas representan los *nodos* del grafo.
2. Por cada arista que une a dos nodos, se suma 1 al valor que hay actualmente en la ubicación correspondiente de la matriz.

Se construyó un algoritmo, transformando los datos extraídos en texto plano en una matriz de adyacencia para construir la red bayesiana implementando una herramienta para crear y visualizar la red bayesiana, en el anexo B se muestra un conjunto de las herramientas más conocidas para redes bayesianas, se escogió el toolbox BNT de MATLAB primero porque es una herramienta libre que tiene muchas funcionalidades a pesar que no tiene interfaz, pero puede combinarse con otros visualizadores de redes, en el ejercicio se utilizó la herramienta Biograph porque tiene una interfaz más amigable.

Posteriormente se creó un algoritmo para transformar los resultados extraídos de texto plano en una matriz de adyacencia.

3.4 VISUALIZACIÓN DE LA RED

Se creo un algoritmo que tomó de las bases de datos los resultados obtenidos en texto plano filtrando las columnas que identificaban las relaciones y convertirlas en una matriz de adyacencia equivalente para crear una estructura de red bayesiana por medio del toolbox de matlab (BNT) Bayes Net Toolbox.

De muchas búsquedas que se hicieron variando los parámetros se determinó la salida de la herramienta STRING que aparece en el Anexo D, posteriormente a este archivo de texto se le aplicó el agoritmo que se construyó y se obtuvo una matriz de abyacencia que permitió crear la estructura de red bayesiana.

En la figura 32 se muestra una salida de este algoritmo, el grafo completo tiene más de 50 proteínas. En esta figura se pueden destacar algunas proteínas que no han sido consideradas por los investigdores.

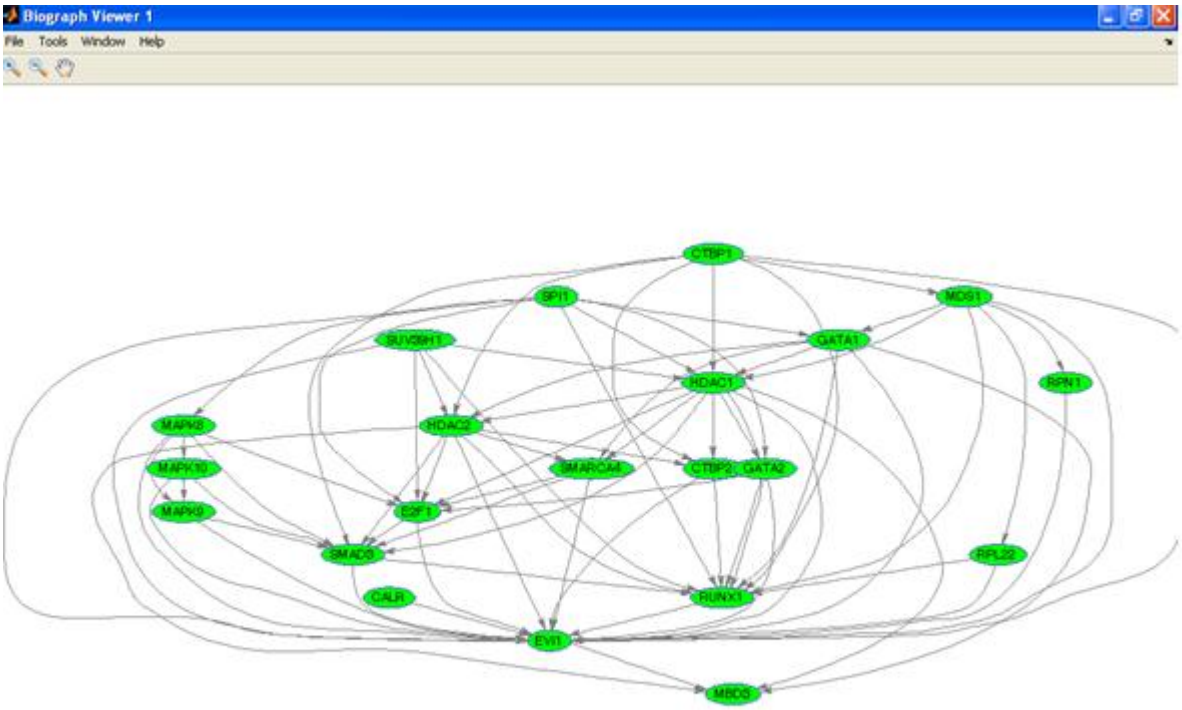


Figura 32 Representación de la Red Bayesiana usando BNT de Matlab

4 CONCLUSIONES

- Se comprobó que la minería de textos proporciona elementos útiles para la investigación y recopilación automática de información para convertirla en conocimiento que permita a los investigadores deducir nuevas hipótesis o interacciones entre proteínas, lo que podría hacer que la medicina sea más predictiva al comprender de manera global las proteínas implicadas en las enfermedades.

- La representación del conocimiento que se logra por medio de la minería de texto optimiza la labor del biólogo y los curadores porque proporciona información organizada y comprensiva que difícilmente puede vislumbrarse con otros enfoques de la ciencia.

- La integración de herramientas de minería de texto permite un mejor resultado, reduciendo los errores que en el campo de investigación de interacciones es considerable debido a la complejidad inherente en este tipo de relaciones.

- Este acercamiento con el área de la biología, que bien puede llamarse biología de sistemas permite unir campos que han trabajado separados, para reducir el coste en los procesos de investigación del grupo IMM (Instituto de medicina molecular de la UIS) no sólo en lo relacionado con las interacciones entre proteínas sino también en otros trabajos de investigación de diferente índole, porque aún es mucho lo que se le puede explotar a la minería de textos.

- En la evaluación de herramientas es importante resaltar que debido a los patrones de comparación instaurado por el experto los resultados no son aplicables globalmente sino a los experimentos de interés.

- Es necesario superar la barrera de la terminología técnica, para obtener resultados que mejoren el desempeño de las herramientas y permitan crear nuevas herramientas.

- Se proporcionó elementos que mejoran el acceso a fuentes de información de bases de datos manejadas por la comunidad científica, para que sea aprovechada en forma eficiente y competitiva por aquellos grupos o individuos la necesiten.

- Los resultados de las evaluaciones fueron mucho más bajos que lo que las herramientas mostraban en sus artículos, esto se debe en parte a que los

investigadores utilizan diferentes colecciones de prueba para evaluar sus sistemas, dando lugar también a la obtención de diferentes resultados según los corpus textuales en los que se aplican los experimentos.

- Las redes bayesianas son un instrumento útil para representar el conocimiento, permite hacer inferencias de manera fácil.

- Los datos obtenidos en el caso de estudio restringen el espacio de probabilidades de una red bayesiana mejorando el desempeño de algoritmos de inferencia y aprendizaje que se apliquen posteriormente.

5 RECOMENDACIONES

- Se espera en un futuro se vinculen nuevos trabajos en el área para clasificar de mejor manera los datos y predecir con certeza cómo reaccionan ciertas sustancias frente a otras sustancias no estudiadas.

- Se recomienda involucrar a grupos de investigación en la extracción de interacciones de proteínas de la literatura científica Y aprovechar los resultados de la minería de texto para implementar modelos de redes bayesianas que permitan una mejor representación del conocimiento extraído para de esta forma optimizar las predicciones al aplicar algoritmos de aprendizaje.

- Debido al carácter multidisciplinar en el campo es necesario que se incremente la interacción de expertos en biología y medicina molecular con ingenieros de conocimiento para formar equipos de investigación que faciliten la integración de las disciplinas, el fortalecimiento de profesionales con conocimientos en ambos ramos y el desarrollo de proyectos de investigación con resultados en tiempo menor de lo que usualmente lleva a los expertos hacer las inferencias o investigaciones de manera convencional.

- Aunque es una tarea compleja que involucra conocimientos más robustos en biología molecular se puede optimizar mejor los resultados de la extracción de las interacciones diseñando algoritmos que puedan utilizar información para comprender el tipo de interacción de las mismas.

6 BIBLIOGRAFIA

Ananiadou, S et al. (2010). Event extraction for systems biology by text mining the literature. *Trends in Biotechnology, 28*(7), 381-390.

Ananiadou, S. et al. (2006). Text mining and its potential applications in systems biology. *Trends Biotechnology, 24*(12), 571–579.

Aplicada, C. d. (2004). *CIMA.* Recuperado el Enero de 2009, de http://www.cima.es/labs/gen%C3%A9tica/resumen/1

Blaschke Cristian, H. L. (2002). *Oxford Journals*.

Blaschke, C. et al. (2005). Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics 6, S16, 6*(S16).

Blaschke, C., & Valencia, A. (2001). The Potential Use of SUISEKI as a Protein Interaction discovery tool. *Genome Informatics*(12), 123–134.

Bressán. (2003). *Almacenes de datos y mineria de datos*. Recuperado el 22 de Mayo de 2007

Carrillo Calvet, H. (s.f.). *UNAM*. Recuperado el 11 de Agosto de 2010, de www.dynamics.unam.edu/ptid/.../MINERIA_simposium080606.ppt

Cohen K. Bretonnel, Y. H. (2005). Text Mining Tools that Work. *Biology*.

Cohen, A., & Hersh, W. R. (Febrary de 2006). A survey of current work in biomedical text mining. *Nature Reviews Genetics, 7*, 119-129.

Dana, P. (2003). *Thesis Submitted for the Degree of Doctor of Phicosophy.*

Díez, F. (1993). *Parameter adjustment in Bayes Networks, The Generalizad noisy or - Gate, Proceedinngs of The Ninth Conference on Uncertainty in Articial Intellegence.* Addison Wesley.

Diez, F. J. (2005). *Introducción al razonamiento aproximado.* Madrid: UNED.

Doorn-Khosrovani, B. v. (2008). Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica*.

Draft. (2009). Evaluation in information Retrieval. *Cambridge*.

Febles, J. P. (s.f.). Aplicacion de la Mineria de datos a la Bioinformatica. *25*.

Friedman, C. et al. . (2001). GENIES: a natural-language processing system for the extraction of molecular.

Gálvez, C. (2008). Minería de Textos: La nueva Generación de análisis de literatura científica en biología celular y genómica. *Enc. Bibli: R. Eletr. Bibliotecon., 1*(25), 1-14.

García, D. (2010). *Desarrollo de un entorno de.usuario para aplicación de. redes bayesianas dinámicas a problemas de fusión de información.* Madrid: Universidad Carlos III de Madrid. Departamento de Informática.

Hartemink, A. (2001). Principled Computational Methods for the Validation and Discovery of Genetic Regulatory Networks. (M. I. Technology, Ed.)

Hena Jose, T. V. (2007). Extraction of Protein Interaction Data: A comparative Analysis Of Methods in Use.

Hersh, W. (2005). Evaluation of medical Text Mining system: Lesson learned from information retrieval. *Briefings in bioinformatics, 6*(4), 344-356.

Hirschman, et al. (2002). Accomplishments and challenges in literature data mining for biology. *Bioinformatics review, 18*(12), 1553–1561.

Hoffman Robert, K. M. (2005). Text Mining for Matabolic Pathways, Signaling cascades, and protein Networks. *Sciences STKE*.

Huang, M. et al. (2004). Discovering patterns to extract protein-protein interactions from full texts. . *Bioinformatics, 20*, 3604–3612.

Jensen, L. J., Saric2, J., & Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *NATURE REVIEWS|Genectis, 7*, 119-129.

Jhon, M. N. (2006). Text mining for Biology and Biomedicine. *Elsevier*.

Jimenez Garcdiaq Luis Felipe, M. H. (s.f.). *Biologia Celular ay Molecular.* Pearson.

Jose Hernandez Orallo, M. J. (2006). *Introduccion a la Mirneria de Texto .* Pearson.

Jose Ramirez Orallo, M. J. (s.f.). Introduccion a la Mineria de Texto . Pearson.

Kim, J.D. et al. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics, 9*(10).

Korb Kevin B., N. A. (2004). *Bayesian Artificial Intelligence.* Chapman & Hall.

Krallinger, M. et al. (2008). Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge.

Laricchia-Robbio1, L. (2008). Inducible expression o EVI1 in human myeloid cells causes phenotypes consistent with its role in myelodysplastic syndromes. *Journal of Leukocyte Biology by the Society for Leukocyte Biology*.

M., G. S. (2003). Prediction of Protein - Protein Interaction Network . *Bioinformatic*.

Mathivanan Suresh. (2006). An Evaluation of Human Protein -Protein Interaction Data in the public Domain. *BMC Bioinformatics*.

McNaught, J., & Black, W. (2006). Information extraction. En *Text Mining for Biology and Biomedicine.*

Odero, M. D. (s.f.). *CIMA- Centro de Investigación Médica Aplicada.* Recuperado el Junio de 2009, de http://www.cima.es

Odor Kanae, k. J.-D. (2008). *BMC Bioinformatics*.

paulino, G. (2007 -2008). *Interacciones proteina -proteina ligando*. Recuperado el Martes de Agosto de 2009, de http://novacripta.cbm.cam.es

Pearl, J. (2009). *Causality Models, Reasoning and inference* (Segunda edición ed.). Cambridge : Cambridge University Press, 2009.

Pe'er, D. (2005). Bayesian network analysis of signaling networks: a primer. *Science STKE*.

Rinaldi Fabio, S. (2007). Mining of relation s between proteins over biomedical scientific literatura using a deep linguistic approach. *ELSEVIER*, 127-136.

Russell, S., & Norvig, P. (2004). *Inteligencia Artificial Un Enfoquie Moderno* (Segunda Edición ed.). Mexico: Prentice Hall.

Rzhetsky, A. et al. (2009). Getting started in text mining: part two. *Plos Comput. Biol., 5*(1).

Sachs, K., Perez, O., Pe'er, D., & LauffenburgeR, D. (2005). Causal protein signaling networks derived from multiparameter single-cell data. . *Science*, 523-529.

Saric, J. et al. (2004). Extracting Regulatory Gene Expression Networks from PubMed.

Tan Pang-Ning, S. (2005). *Introduction to Data Mining.* Pearson Adddison Wesley.

Uffe, B. J., & Anders, L. (2005). *Probabilistic Networks — An Introduction toBayesian Networks and Influence Diagrams.* Aalborg University.

Yan Jun, F. B. (2008). Protein Complex Identification by Supervised graph local clustering. *Bioinformatics*, 250-258.

Zhou, D., & He, Y. (2008). Extracting interactions between proteins from the literature. *Biomedical Informatics*, 1-15.

7 ANEXOS

ANEXO A: LISTA DE COMPARACIÓN DE PROTEÍNAS

#	REFERENCIA	PROTEINA	RELACION
1	Roy, 2010	Caspasa III	↑ Indirecta, Apoptosis inducida* H_2O_2
2	Roy, 2010	Anhidrasa carbónica III	↓ Directa. Inhibe apoptosis *paclitaxel
3	Bingemann, 2010		Modulador de la esta. Acido retinoica
4	Goyama, 2010	Suv39 1 HMT	Acoplan Regulación de la transcripción R.
5	Goyama, 2010	Gga HMT	Unión inmortalización Or
6	Shimaba, 2009	Pbx1	↑ Directa HSC – Leukogenesis
7	Komeno, 2009	P53	↓ Indirecta revisión
8	Larichia – Robbio, 2009	Gata 1	↑ Indirecta. Bloqueo eritropoyesis
9	Larichia – Robbio, 2009	Pu 1	Unión. Inhibe la activación promotor miel
10	Senyuk, 2009	Stat 3	↑ Indirecta
11	Takanata, 2009	Ski	Interactuan – Indirecta – señalización TGFB
12	Marri, 20009	EGT – 43	De la familia Evi 1

13	Qiu, 2009	Calreticulina	Regulacion trnascripcion. En cardiogenesis
14	Catttaneo, 2009	SOV 3gH1	Acoplan regulación transcripcional
15	Cattaneo, 2009	Gga	Acoplan regulación transcripcional
16	Spensberger, 2008	Mbd3b	Interaccion Inhibe la fucnion des de la histona
17	Modlich, 2008	PRDMTG	↑ De ambas suficiente para iniciar leucemia
18	Sato, 2008	Gata 2	↑ Promueve la expansión progenitor
19	Sato, 2008	TGF-B	↓ Madre hematopoiticas en embrión
20	Qiu, 2008	Calreticulina	↓ Directa en Cardiogenesis
21	Li, 2007	Cg	↑ Analisis del promotor de Cg afin por Evi 1
22	Tokita, 2007	CLEBP alpha	$\frac{\quad\vert}{\quad} f_{(x)}$ Reclutamiento de la histona de via
23	Tokita, 2007	CtBP	Union CtBP y la discrupcion de unión al DNA ↑
24	Carrella, 2007	Mn1	Coexpresa Pacientes AML conn sobre de Evi1
25	Rimanni, 2007	EGt43	Homologo del Oncogen Evi 1 ↑
26	Jin, 2007	Hoxa	Evi 1 coexpresada con Trib altera la
27	Jin, 2007	Meis 1	↑ Expresion de estos y pertuban la ≠ mieloide
28	Piccin, 2007	Vitamina D3	↓ Pituitaria envuelta en eritro
29	Piccin,2007	TGF-B1	↓ Leucogenesis diabetes eritroleucemia
30	Laricchia - Robbio	Gata 1	Union Bloquea la union de Gata 1 a secuencias DNA

31	Liu, 2006	PAL 1	⊣ Indirecta inhibiendo señalamiento del TGF-B
32	Liu, 2006	P13K	ON Inhibe proceso de apoptosis mediadra
33	Liu, 2006	AKT	ON por TGFB envolviendo PIBK y AKT
34	Boyd, 2006	Sox 4 (proviral)	Cooperan Sox4 activan Evi1 (Repeticiones LTR)
35	Kilbey, 2005	CDK2	↓ Ind. Evi1 bloquea endomitosis en megakariocitos ↓
36	Kilbey, 2005	Ciclina A	↓ Ind. Inhibiendo $a_{(x)}$ Catalica de CDK2 de CicliA
37	Yatsula, 2005	Gadd45g	
38	Yatsula, 2005	Gata 2	Genes cuyo promotor presenta
39	Yatsula, 2005	Zfpml Fog 2	Sitios de union al dominio
40	Yatsula, 2005	Skil (SnoN)	N-terminal de Evi1, la mayoria
41	Yatsula, 2005	Klf5(BTEB2)	De estos sitios es unida por Evi.
42	Yatsula, 2005	Don	Tipo silvestre
43	Yatsula, 2005	Map3K14 (Nik)	
44	Katoh, 2005	FGF20	Promotor presenta doblesito de union para Evi1
45	Katoh, 2005	FGF11	Promotor presenta sitio de union para Evi 1
46	Yuasa, 2005	Gata 2	↓ Indirecta promotor Gata sitios de union para Evi 1
47	Alliston, 2005	Smad 1	⊣ Represion de induccion de vias BMP/Smad 1

48	Alliston, 2005	Smad 2	⊣ Activina /Smad 2
49	Alliston, 2005	Smad 7	⊣ Evi1 actua como correpresos de las Smads
50	Buonamici, 2005	IFN – Alpha	⊣
51	Buonamici, 2004	EPOR	↓ En la ≠ terminal eritroide y
52	Buonamici, 2004	C-MPI	↓ Formacion de plaquetas
53	Chi, 2003	E2F1	↑ Activa el promotor de E2F1
54	Chi, 2003	BRG1	Interactua Promoviendo el crecimiento
55	Yokoi, 2003	TERC	Blanco de Evi1 .Codifica el componente RNA de la terome
56	Takahashi, 2002	PLZ F	↑ Promotor PLZ F presenta sitio de union a Evi1
57	Vinatzer, 2001	Histona de acetilasa	Interactuan
58	Chakraborty, 2001	CtbP1	Interactuan - correguladores que se asocia
59	Chakraborty, 2001	Histona de acetilasa	Interactuan Evi1 para regular la expresion
60	Chakraborty, 2001	CBP (HAT)	Interaccion con Evi1 realza
61	ChaKraborty, 2001	P/CAF (HAT)	Interaccion acetilacion de Evi 1 Coactivadores
62	ShimiZu, 2002	CD34	Evi 1 Expresado especifical en OCD34(+)
63	Shimizu, 2002	Trombopoyetina	Alta expression Evi1 o con trombopoeilina
64	Shimizu, 2002	Eritropoyetina	Bajos niveles de Evi1 con + EPO
65	Shimizu, 2002	GM-CSF	Bajos niveles de Evi1 + GM
66	Joosten, 2002	Cb 2	Colaboran. Para leukogenesis

67	Palmer, 2001	CtBP	Union co-represora ayuda a la $a_{(x)}$ Evi 1
68	Izutsu, 2002	CtBP	Corepresor, interactua con Evi 1 para inhibir TGFB
69	KuroKawa,2000	JMK	EVI 1 muerte por selectiva de JNK
70	Daimon , 1999	CHTP 1 / TLO (1)	Regio del promotor presenta sitio de union a Evi 1
71	Roberts, 1999	Smad 3	
72	Sitalio, 1999	TGFB	del señalamiento
73	Kilbey, 1999	Ciclina A	Acortar la fase G1 del ciclo O
74	Kilbey, 1999	Cdk 2	Y reduce los requirimientos de mitogenos
75	Kilbey, 1999	Ciclina E	Y suero para la entrada a la fase S
76	Kilbey,1999	P27	Disregula control dela fase CT1
77	Kilbey, 1999	PRb	Incremento en niveles de Rb hiper
78	Kim, 1998	ITPR2	Promotor presenta sitios de union a Evi 1
79	KuroKawa, 1998	TGF-B	
80	KuroKawa, 1998	Smad 3	
81	Soderholm, 1997	Gata – 1	
82	Ogawa, 1996	AP-1	Indirecta En celulas NIH3T3
83	SuzuKawa, 1994	Riboforina 1	Media la activacion transcripcion de Evi1 en Inv(3)(q21q26)
84	Tanaka, 1994	AP-1	Incrementa la expresion

85	Tanaka, 1994	C-fos	↑ Activa promotor C-fos
86	Tanaka, 1994	C-sun	Incrementa expresion endogena.
87	Khanna – Gupta,1996	G-CSF	⊣ Siempre y cuando los niveles de Evir1 sean altos como cuando es insercion retro ⊣ Difrerenciacion y supervivencia.
88	Perlans, 1996	Gata 1	⊣ Compitiendo por sitios de union.
89	OhyashikCi, 1995	Gata 1	Cuando se presenta sobre-expresion de Evi1, se presenta a su vez expresion en estos 3 genes enpacientes.
90	OhyashikCi, 1995	Gata 2	
91	OhyashikCi, 1995	SCL	
92	Kreider, 1993	Gata 1	
93	Morishita, 1992	Mieloperoxidasa	Impide la expresion

ANEXO B. RESULTADOS DE IHOP

Symbol	Name	Organism	Nr of sentences with MECOM	
RUNX1	runt-related transcription factor 1 (acute myeloid leukemia 1; aml1 oncogene)	Homo sapiens	32	
MDS1	myelodysplasia syndrome 1	Homo sapiens	20	
SUV39H1	suppressor of variegation 3-9 homolog 1 (Drosophila)	Homo sapiens	12	
PRDM16	PR domain containing 16	Homo sapiens	10	
ETV6	ets variant gene 6 (TEL oncogene)	Homo sapiens	7	
CTBP1	C-terminal binding protein 1	Homo sapiens	6	
HDAC1	histone deacetylase 1	Homo sapiens	6	
RPN1	ribophorin I	Homo sapiens	6	
SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin,...	Homo sapiens	6	
TGFB1	transforming growth factor, beta 1 (Camurati-Engelmann disease)	Homo sapiens	6	
UXT	ubiquitously-expressed transcript	Homo sapiens	6	
CEBPA	CCAAT/enhancer binding protein (C/EBP), alpha	Homo sapiens	4	
CREBBP	CREB binding protein	Homo sapiens	4	
GATA1	GATA binding protein 1 (globin transcription factor 1)	Homo sapiens	4	
MBD3	methyl-CpG binding domain protein 3	Homo sapiens	4	
SPI1	spleen focus forming virus (SFFV) proviral integration oncogene spi1	Homo sapiens	4	
CDK6	cyclin-dependent kinase 6	Homo sapiens	3	
KAT2B	K(lysine) acetyltransferase 2B	Homo sapiens	3	
E2F1	E2F transcription factor 1	Homo sapiens	2	
EPO	Erythropoietin	Homo sapiens	2	
MSI2	musashi homolog 2 (Drosophila)	Homo sapiens	2	
SETBP1	SET binding protein 1	Homo sapiens	2	
THPO	Thrombopoietin	Homo sapiens	2	
AKT1	v-akt murine thymoma viral oncogene homolog 1	Homo sapiens	1	

Symbol	Name	Organism	Nr of sentences with MECOM	
BAALC	brain and acute leukemia, cytoplasmic	Homo sapiens	1	
C3orf27	chromosome 3 open reading frame 27	Homo sapiens	1	
CCNE1	cyclin E1	Homo sapiens	1	
CDK2	cyclin-dependent kinase 2	Homo sapiens	1	
EBP	emopamil binding protein (sterol isomerase)	Homo sapiens	1	
ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma...	Homo sapiens	1	
FGF20	fibroblast growth factor 20	Homo sapiens	1	
FOS	v-fos FBJ murine osteosarcoma viral oncogene homolog	Homo sapiens	1	
FOSB	FBJ murine osteosarcoma viral oncogene homolog B	Homo sapiens	1	
FOXA2	forkhead box A2	Homo sapiens	1	
GATA2	GATA binding protein 2	Homo sapiens	1	
GP6	glycoprotein VI (platelet)	Homo sapiens	1	
IL10	interleukin 10	Homo sapiens	1	
IL6	interleukin 6 (interferon, beta 2)	Homo sapiens	1	
ITGB3	integrin, beta 3 (platelet glycoprotein IIIa, antigen CD61)	Homo sapiens	1	
JUND	jun D proto-oncogene	Homo sapiens	1	
LPP	LIM domain containing preferred translocation partner in lipoma	Homo sapiens	1	
MYC	v-myc myelocytomatosis viral oncogene homolog (avian)	Homo sapiens	1	
PAX4	paired box gene 4	Homo sapiens	1	
PF4	platelet factor 4	Homo sapiens	1	
PIK3CA	phosphoinositide-3-kinase, catalytic, alpha polypeptide	Homo sapiens	1	
PRDM2	PR domain containing 2, with ZNF domain	Homo sapiens	1	
PTK2	PTK2 protein tyrosine kinase 2	Homo sapiens	1	
SKIL	SKI-like	Homo sapiens	1	

Symbol	Name	Organism	Nr of sentences with MECOM	
SMAD3	SMAD, mothers against DPP homolog 3 (Drosophila)	Homo sapiens	1	
TERC	telomerase RNA component	Homo sapiens	1	
TGIF1	TGFB-induced factor homeobox 1	Homo sapiens	1	
ZEB1	zinc finger E-box binding homeobox 1	Homo sapiens	1	
NCKAP1L	NCK-associated protein 1-like	Homo sapiens	11	
BCR	breakpoint cluster región	Homo sapiens	3	
GOLGA2LY2	golgi autoantigen, golgin subfamily a, 2-like, Y-linked 2	Homo sapiens	3	
PIK3CG	phosphoinositide-3-kinase, catalytic, gamma polypeptide	Homo sapiens	3	
TTTY17C	testis-specific transcript, Y-linked 17C	Homo sapiens	3	
TTTY4C	testis-specific transcript, Y-linked 4C	Homo sapiens	3	
INHBA	inhibin, beta A (activin A, activin AB alpha polypeptide)	Homo sapiens	2	
MDCR	Miller-Dieker syndrome chromosome region	Homo sapiens	2	
TTTY3B	testis-specific transcript, Y-linked 3B	Homo sapiens	2	
TTTY6B	testis-specific transcript, Y-linked 6B	Homo sapiens	2	
APPBP1	amyloid beta precursor protein binding protein 1	Homo sapiens	1	
FGF12	fibroblast growth factor 12	Homo sapiens	1	
HNF4A	hepatocyte nuclear factor 4, alpha	Homo sapiens	1	
HNMT	histamine N-methyltransferase	Homo sapiens	1	
PAFAH1B1	platelet-activating factor acetylhydrolase, isoform Ib, alpha subunit 45kDa	Homo sapiens	1	
POU3F3	POU class 3 homeobox 3	Homo sapiens	1	
TFCP2	transcription factor CP2	Homo sapiens	1	
YWHAE	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein,...	Homo sapiens	1	

ANEXO C Resultados Herramienta Novoseek

CCR7
EVI1
RUNX1
CCL21
CCL19
CXCR4
MDS1
MEL1
CTBP
CXCR5
Nucleophosmin
CCR5
GATA2
CCR4
Pvt1 oncogene
LMO2
CXCR3
death receptor inter...
GATA1
CCR6

ANEXO D. Resultados herramienta STRING

node1	node2	fusion	cooccurence	homology	coexpression	experimental	knowledge
E2F1	SMAD3	0.000	0.000	0.000	0.000	0.000	0.000
SPI1	GATA1	0.000	0.000	0.000	0.000	0.844	0.000
CTBP1	HDAC2	0.000	0.000	0.000	0.000	0.835	0.900
SPI1	HDAC1	0.000	0.000	0.000	0.000	0.839	0.000
PBX1	RUNX1	0.000	0.000	0.000	0.000	0.000	0.000
CTBP2	RUNX1	0.000	0.000	0.000	0.000	0.000	0.800
SPI1	MAPK8	0.000	0.000	0.000	0.000	0.611	0.000
HDAC1	HDAC2	0.000	0.515	0.975	0.000	0.999	0.900
SPI1	RUNX1	0.000	0.000	0.000	0.000	0.000	0.800
MDS1	EVI1	0.015	0.000	0.000	0.000	0.000	0.000
GATA1	HDAC2	0.000	0.000	0.000	0.000	0.000	0.900
SMAD3	EVI1	0.000	0.000	0.000	0.000	0.636	0.000
CTBP1	EVI1	0.000	0.000	0.000	0.000	0.944	0.800
SPI1	GATA2	0.000	0.000	0.000	0.000	0.844	0.000
SUV39H1	EVI1	0.000	0.000	0.000	0.000	0.000	0.000
HDAC2	CTBP2	0.000	0.000	0.000	0.000	0.000	0.800
SMAD3	RUNX1	0.000	0.000	0.000	0.000	0.577	0.900
SMARCA4	E2F1	0.000	0.000	0.000	0.000	0.000	0.000
HDAC2	E2F1	0.000	0.000	0.000	0.000	0.000	0.000
GATA2	RUNX1	0.000	0.000	0.000	0.000	0.000	0.000
HDAC2	SMAD3	0.000	0.000	0.000	0.000	0.000	0.900
SMARCA4	EVI1	0.000	0.000	0.000	0.000	0.000	0.000
CALR	EVI1	0.000	0.000	0.000	0.000	0.000	0.000
CTBP1	RUNX1	0.000	0.000	0.000	0.000	0.000	0.800
SUV39H1	E2F1	0.000	0.000	0.000	0.000	0.000	0.900
GATA1	HDAC1	0.000	0.000	0.000	0.000	0.606	0.900
GATA2	EVI1	0.000	0.000	0.000	0.000	0.000	0.000
SPI1	EVI1	0.000	0.000	0.000	0.000	0.000	0.000
MDS1	RPN1	0.000	0.000	0.000	0.000	0.000	0.000
MDS1	HDAC1	0.000	0.000	0.000	0.000	0.636	0.000
MAPK8	MAPK10	0.000	0.515	0.976	0.000	0.000	0.900
SUV39H1	RUNX1	0.000	0.000	0.000	0.000	0.636	0.000
MAPK10	EVI1	0.000	0.000	0.000	0.000	0.000	0.800
GATA1	EVI1	0.000	0.000	0.000	0.000	0.000	0.000
SPI1	E2F1	0.000	0.000	0.000	0.000	0.000	0.900
HDAC1	RUNX1	0.000	0.000	0.000	0.000	0.000	0.800
HDAC1	EVI1	0.000	0.000	0.000	0.000	0.848	0.800
CA3	EVI1	0.000	0.000	0.000	0.000	0.000	0.000

CTBP1	CTBP2	0.000	0.000	0.970	0.000	0.611	0.800
PBX1	EVI1	0.000	0.000	0.000	0.000	0.000	0.000
MDS1	RUNX1	0.000	0.000	0.000	0.000	0.000	0.000
HDAC1	MBD3	0.000	0.000	0.000	0.000	0.636	0.900
MAPK8	SMAD3	0.000	0.000	0.000	0.000	0.636	0.000
GATA1	SMARCA4	0.000	0.000	0.000	0.000	0.567	0.000
RUNX1	EVI1	0.000	0.000	0.000	0.000	0.000	0.800
CTBP1	HDAC1	0.000	0.000	0.000	0.000	0.935	0.900
MAPK8	EVI1	0.000	0.000	0.000	0.000	0.000	0.800
HDAC1	SMAD3	0.000	0.000	0.000	0.000	0.000	0.900
GATA1	RUNX1	0.000	0.000	0.000	0.000	0.000	0.000
CTBP2	EVI1	0.000	0.000	0.000	0.000	0.606	0.800
MAPK8	E2F1	0.000	0.000	0.000	0.000	0.000	0.000
E2F1	EVI1	0.000	0.000	0.000	0.000	0.000	0.000
EVI1	MBD3	0.000	0.000	0.000	0.000	0.094	0.000
RPN1	EVI1	0.000	0.000	0.000	0.000	0.000	0.000
MAPK10	SMAD3	0.000	0.000	0.000	0.000	0.000	0.000
HDAC2	SMARCA4	0.000	0.000	0.000	0.000	0.790	0.000
SUV39H1	HDAC2	0.000	0.000	0.000	0.000	0.853	0.000
SUV39H1	HDAC1	0.000	0.000	0.000	0.000	0.942	0.000
GATA1	MBD3	0.000	0.000	0.000	0.000	0.000	0.900
HDAC2	EVI1	0.000	0.000	0.000	0.000	0.000	0.800
HDAC2	MBD3	0.000	0.000	0.000	0.000	0.000	0.900
SMARCA4	SMAD3	0.000	0.000	0.000	0.000	0.000	0.000
HDAC1	E2F1	0.000	0.000	0.000	0.000	0.833	0.900
HDAC2	RUNX1	0.000	0.000	0.000	0.000	0.000	0.800
HDAC1	CTBP2	0.000	0.000	0.000	0.000	0.000	0.800
HDAC1	SMARCA4	0.000	0.000	0.000	0.000	0.567	0.000

ANEXO E. Resultados de la herramienta Facta

Proteína	# Artículos
EVI1	129
AML1	31
MDS1	24
MDS1-EVI1	18
Evi-1	16
zinc finger	16
erythroid	15
telomeric	13
proto-oncogene	11
TEL	11
RIZ1	10
p13	10
GATA-1	9
EAP	9
zinc finger protein	9
CML	8
lymphoid	8
CD34	8
MEL1	7
BCR	6
GATA-2	6
transcription repressor	6
histone deacetylase	6
ribophorin I	6
ABL	6
granulocyte colony-stimulating factor	5
IL-3	5
cryptic	5
CBF	4
unknown function	4
PIK3CA	4
FLT3	4
PRDM1	4

MDS	4
SNO	4
CEBPA	4
FISH	4
CD33	4
p36	4
MLL	4
p53	4
transforming growth factor-beta	4
peroxidase	4
CD7	3
repressor protein	3
p15	3
granulocyte	3
translocated to	3
breakpoint cluster region	3
WT1	3
methyltransferase	3
Myosin heavy chain	3
interferon-alpha	3
ETO	3
transcriptional activator	3
CD13	3
transforming protein	3
c-kit	3
TGF-beta	3
DNA-binding protein	3
chimeric protein	3
megakaryocyte	3
MYC	3
CSF	3
BAALC	3
RUNX2	2
thymidine kinase	2
PU.1	2
c-fos	2
B23	2
GAL4	2

transferrin receptor	2
p16	2
BRCA1	2
chromosome 21	2
p190	2
c-fms	2
CtBP1	2
Smad3	2
tissue plasminogen activator	2
thrombopoietin	2
DNA binding protein	2
FGF	2
ribosomal protein L22	2
RUNX3	2
SUV39H1	2
cytokine	2
c-myc	2
IL-6	2
stem cell factor	2
PML	2
RNase	2
SET	2
PCAF	2
IFN	2

ANEXO F resultados de Pubgene

Proteínas	Articulos publicados
MDS1	21
GATA1	14
CD34	10
IL3	9
MYB	9
MYC	9
SMAD3	9
CD33	5
RUNX1	5
CD7	4
CTBP1	4
FLT3	4
P53	4
CEBPA	3
CSF1R	3
EPOR	3
ETV6	3
GATA2	3
JUN	3
KLF1	3
PERM	3
SCF	3
SMAD4	3
BAALC	2
CASP1	2
CASP3	2
CCNA2	2
CD4	2
CDK2	2
CSF2	2
CTBP2	2
DCOR	2
FOXQ1	2
GR6	2
INI7	2
MAFK	2

NPM	2
NXT1	2
PAI1	2
PCAF	2
PERT	2
SKIL	2
SMAD2	2
SMAD7	2
TCHP	2
TPO	2
TPOR	2
TRFE	2
ACTN4	1

ANEXO G: Lista de herramientas para redes bayesianas y sus propiedades

Nombre	Src	API	Exec	Cts	GUI	Params	Struct	Utility	Free	Undir	Inference
AgenaRisk	No	Si	W,U	Cx	Si	Si	No	No	$	D	JTree
Analytica	Java	Si	W,U,M	Cd	No	No	Si	No	0	D	sampling
Banjo	Java	Si	W,U,M	Cd	No	No	Si	No	0	D	None
Bassist	C++	Si	U	G	No	Si	No	No	0	D	MH
BayesBuilder	No	No	W	D	Si	No	No	No	0	D	-
BayesiaLab	No	No	-	Cd	Si	Si	Si	No	$	CG	JTree, G
Bayesware Discoverer	No	No	W,U,M	Cd	Si	Si	Si	No	$	D	-
B-course	No	No	W,U,M	Cd	Si	Si	Si	No	0	D	-
Belief net power constructor	No	Si	W	D	Si	Si	CI	No	0	D	-
BayesBlock	Python/C++	Si	-	Si	No	Si	No	No	0	-	VMP
Blaise	Java	Si	-	Si	No	Si	No	No	0	-	-
BNT	Matlab/C	Si	W,U,M	G	No	Si	Si	Si	0	D,U	Muchos
BNJ	Java	-	-	D	Si	No	Si	No	0	D	JTree, IS
BNL	Matlab	-	-	D	No	No	No	No	0	D	JTree
BUGS	No	No	W,U	Cs	Si	Si	No	No	0	D	G
Causal Discoverer	No	No	W	-	-	No	Si	No	0	D	-
Clspace	Java	No	W,U	D	Si	No	No	No	0	D	Varelim
CRFtoolbox	Matlab/C	Si	-	No	No	Si	No	No	0	U	-
DBNbox	Matlab	-	-	Si	No	Si	No	No	$	D	Varios

DeriveIt	No	-	-	-	-	Si	Si	-	$	D	JTree, Gibbs
Elvira	Java	Si	W,U,M	Cd,Cx	Si	Si	Si	Si	0	D	JTree, varelim, ls
GDAGsim	C	Si	W,U,M	G	No	No	No	No	0	D	-
GeNle and SMILE	SMILE	Si	W,U,M	Cs	Si	Si	Si	Si	0	D	JTree, IS
GMRFSim	C	Si	W,U,M	G	No	No	No	No	0	U	-
GMTk	No	Si	U	D	No	Si	Si	No	0	D	JTree
gR	-	-	-	-	-	-	-	-	0	-	-
Grappa	-	-	-	D	No	No	No	No	0	D	JTree
HdBCS	C++	-	-	G	No	Si	Si	No	0	U	SL
Hugin Expert	N	Si	W	G	Si	Si	CI	Si	$	CG	JTree
Hydra	Java	-	-	Cs	Si	Si	No	No	0	U,D	-
Infer.NET	C#	Si	Si	Si	No	Si	No	No	0	Si	VMP, EP, G
JAGS	Java	Si	-	Si	No	Si	No	No	0	Si	G
Java Bayes	Java	Si	W,U,M	D	Si	No	No	Si	0	D	Varelim, JTree
LibB	No	Si	W	D	No	Si	Si	No	0	D	SL
MIM	No	No	W	G	Si	Si	Si	No	$	CG	JTree
Mocapy++	C++	Si	W,U,M	G	No	Si	No	No	0	D	G
MSBNx	No	Si	W	D	Si	No	No	Si	0	D	JTree
Netica	No	Si	W,U,M	G	Si	Si	No	Si	$	D	JTree
Bayes net Learner	No	No	W,U	D	No	Si	Si	No	0	D	SL
PMT	Matlab/C	-	-	D	No	Si	No	No	0	D	-
PNL	C++	-	-	D	No	Si	Si	No	0	U,D	JTree
Pulcinella	Lisp	Si	W,U,M	D	Si	No	No	No	0	D	-
RISO	Java	Si	W,U,M	G	Si	No	No	No	0	D	Polytree
Sam lam	No	No	W,U	G	Si	Si	No	Si	0	D	-
LADR	No	Si	-	Cd	No	No	Si	No	$	D	None
Tetrad	No	No	W,U	G	No	Si	CI	No	0	U,D	SL
UC Irvine	Si	No	W,U	D	No	No	No	No	0	U,D	-
UnBBayes	Java	-	C	D	Si	No	Si	No	0	D	JTree
Vides	Java	Si	W,U	Cx	Si	Si	No	No	0	D	VMP
WinMine	No	No	W	Cx	Si	Si	Si	No	0	U,D,	SL
XBAIES 2.0	No	No	W	G	Si	Si	No	Si	0	CG	JTree

A continuación se explicarán las cabeceras de los atributos deL ANEXO F, que recoge el conjunto de características de los paquetes de software:

Src = No: no se incluye el código fuente. En otro caso indica el leguaje de programación.

API = No, significa que el programa no puede ser integrado en nuestro código propio. En otras palabras, que sólo puede ser ejecutado como módulo aislado.

Exec = Ejecutable bajo W = Windows (95/98/NT/XP), U = Unix, M = Mac o C = cualquier máquina con un compilador.

Cts = nudos continuos (latentes) soportados: G = (condicionalmente) nudos gaussianos soportados analíticamente, Cs = nudos continuos soportados por muestreo, Cd = nudos continuos soportados por discretización, Cx = nudos continuos soportados por algún método no especificado, D = nodos discretos soportados.

GUI = si se incluye Interfaz Gráfica de Usuario.

Params = si se permiten parámetros con aprendizaje.

Struct = representa el tipo de estructura de aprendizaje, CI = significa que usa tests condicionales de independencia, No = no se disponen de estructuras y Si = si se permiten.

Utility = indica si se permiten nudos de decisión.

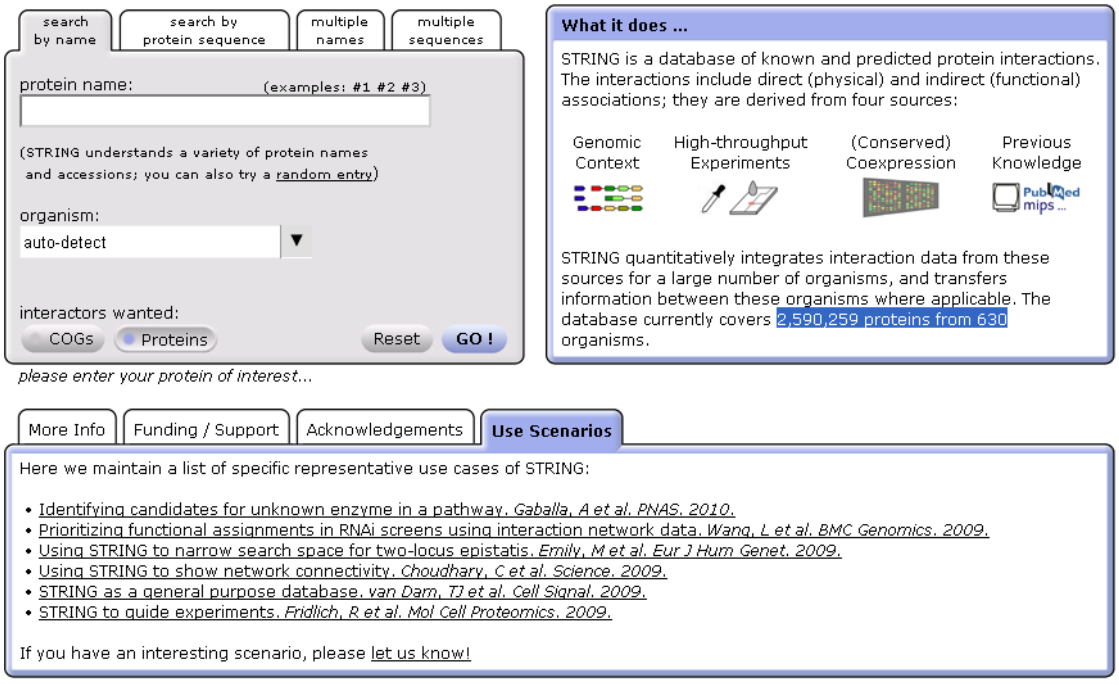
Free = 0 indica software gratuito (aunque posiblemente para uso académico). $ = software comercial (la mayoría tiene versiones gratuitas que están restringidas o reducidas de varios modos; por ej., el tamaña del modelo es limitado, los modelos no se pueden guardar o no hay API).

Undir = representa el tipo de grafo soportado, U = solamente grafos no orientados, D = grafos orientados, UD = ambos tipos grafos (orientados y no orientados) y G = grafos de cadena (mezclados orientados/no orientados).

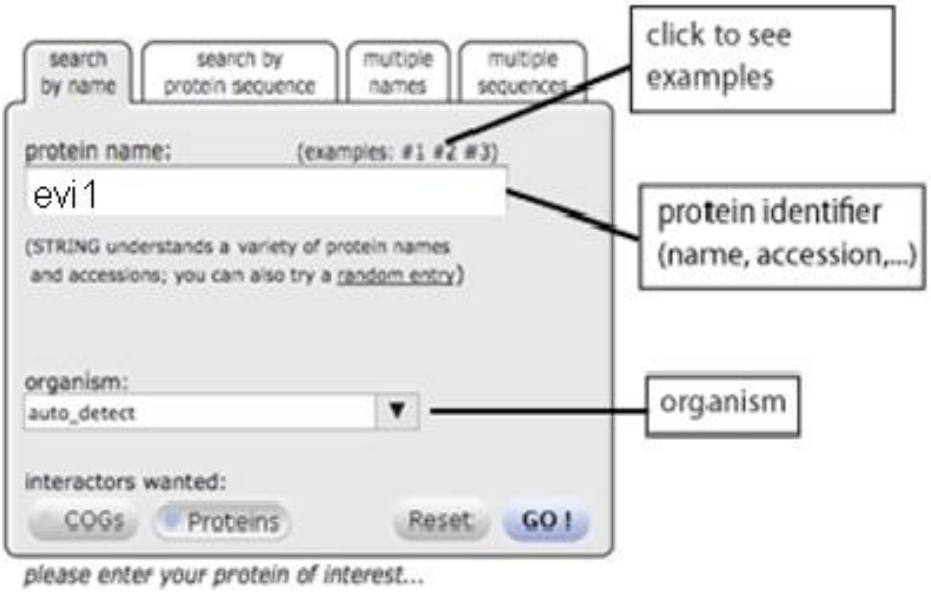
Inference = que tipo de algoritmo de inferencia es usado, jTree = junction tree, varelim = variable (bucket) elimination, MH = Metropols Hastings, G = muestreo de Gibbs, IS = importance sampling, sampling = algún otro método de MonteCarlos, polytree = algoritmo de PEARL restringido a un grafo sin ciclos, VMP paso de mensajes variacional, EP = propagación expectativa, none = no se soporta la inferencia (el programa es diseñado solamente para la estructura de aprendizaje de datos completamente observados).

ANEXO H. Guía de uso STRING

Se accede a STRING con el siguiente link http://string.embl.de/ , la página principal muestra la siguiente figura.



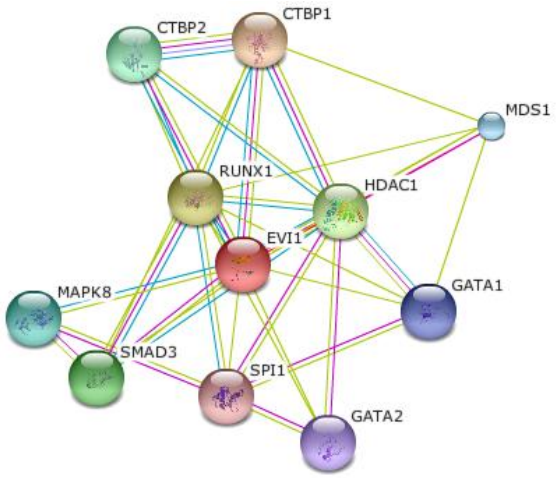
El primer cuadro es el cuadro por el que se hacen consultas, tiene varias opciones de consultas en varias pestañas, nos centraremos en las pestañas search by nome y multiples names para consultas multiples, los restantes cuadros muestran una breve información de la herramienta, cómo son fuentes de información, colaboradores, publicaciones acerca de STRING, etc.



Para hacer la búsqueda en protein name se digita la proteína de interés, en nuestro caso es evi1, seleccionamos el tipo de organismo (homo Sapiens), en interactors wanted se verifica que este activo el botón de proteins, como ejemplos se puede dar click en #1, #2 o #3 y se harán consultas de trpA para Escherichia coli K12, CDC15 para Saccharomyces cerevisiae, y wnt7A para ningún organismo especifico.

Cuando damos click en Go! Nos genera una red de las interacciones principales y el resultado lo muestra en tres partes: La gráfica, el resumen de predicción y los parámetros.

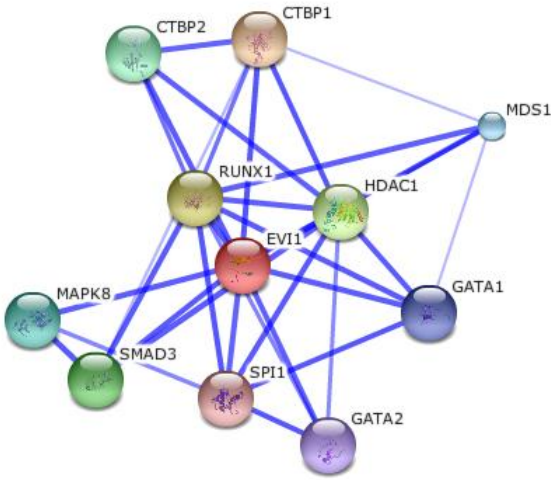
Gráfica de la red



This is the **evidence view**. Different line colors represent the types of evidence for the association.

(requires Flash player 10 or better)

Por defecto muestra las principales interacciones en modo de evidencia, combinado las principales fuentes, como son los experimentos, el Text mining y la información extraída de las bases de datos científicas, todo ello con algunos parámetros determinados que se explican más adelante. Los resultados en forma gráfica son mostrados en cuatro modos: el de confidencia, el de evidencia, el interactivo y el de acciones.

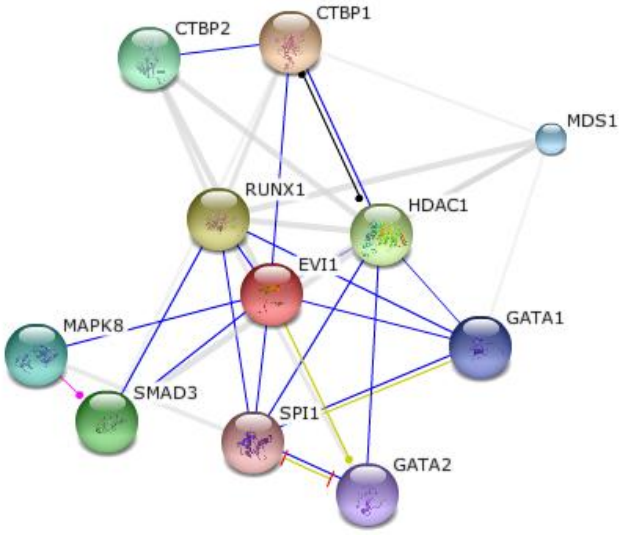


This is the **confidence view**. Stronger associations are represented by thicker lines.

Modo confidencia

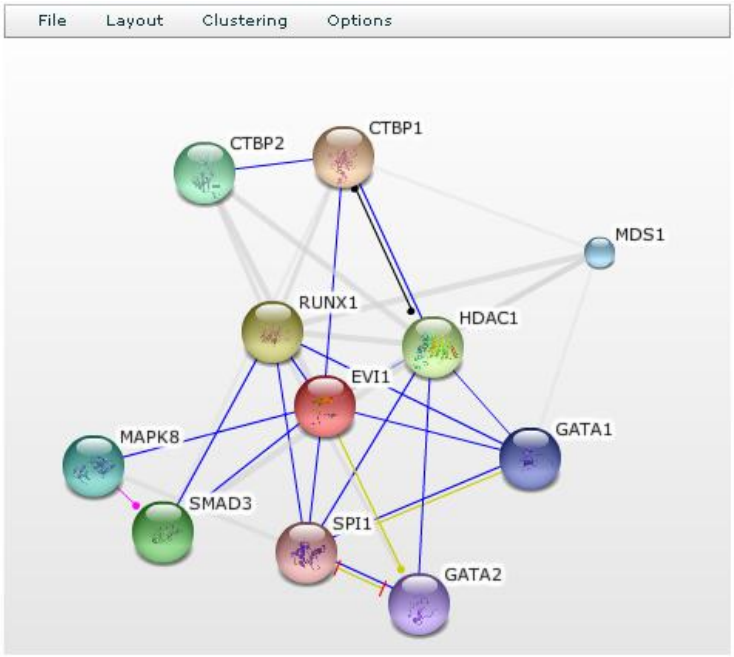
En este modo se puede diferenciar que tan fuertes son las asociaciones de las interacciones dependiendo del grueso de la línea.

Modo Action

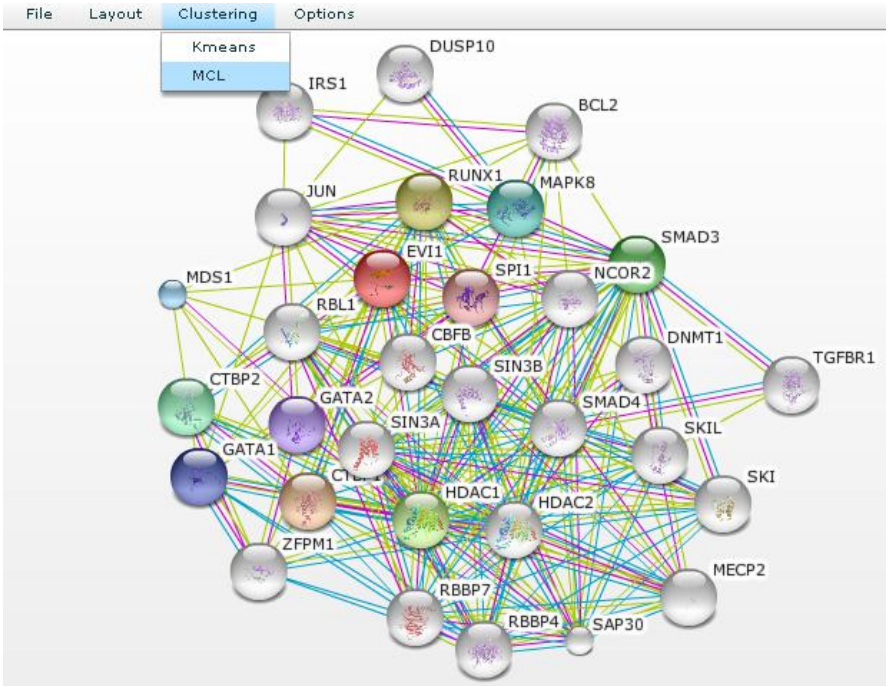


This is the **actions view**. Modes of action are shown in different colors.

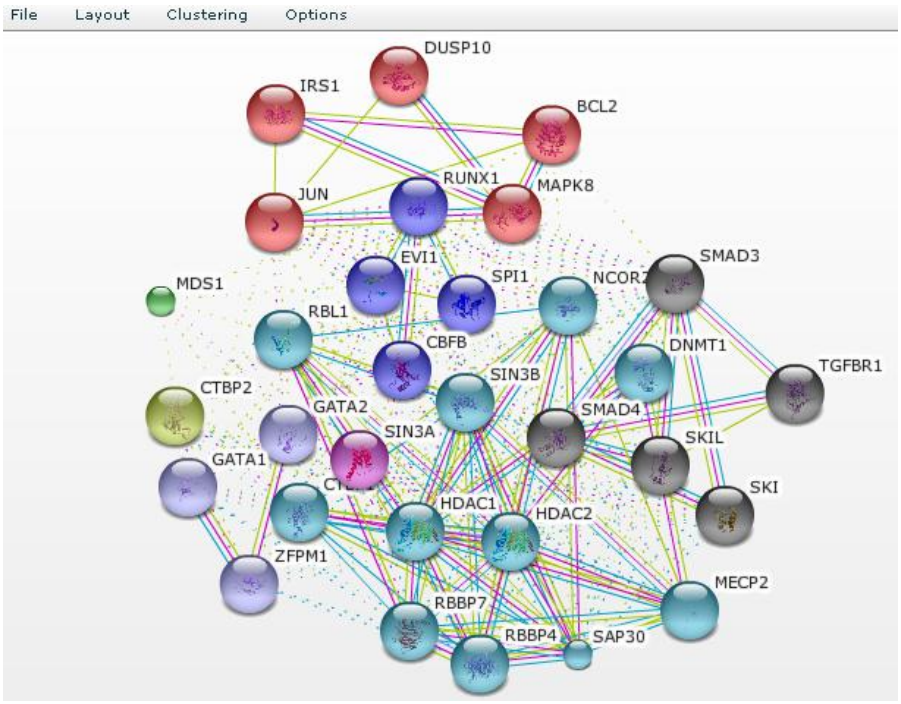
Modo interactive



Esta opción activa el botón avanced el cual posibilita características de presentación y otras opciones como la de clustering que permite aprendizaje no supervisado, que combinado con la estadística puede facilitar la inferencia porque agrupa nodos de características similares.



En el ejemplo se ha seleccionado el algoritmo MCL Markov cluster algorith, para agrupar en 6 clases las interacciones reportadas por la herramienta STRING.

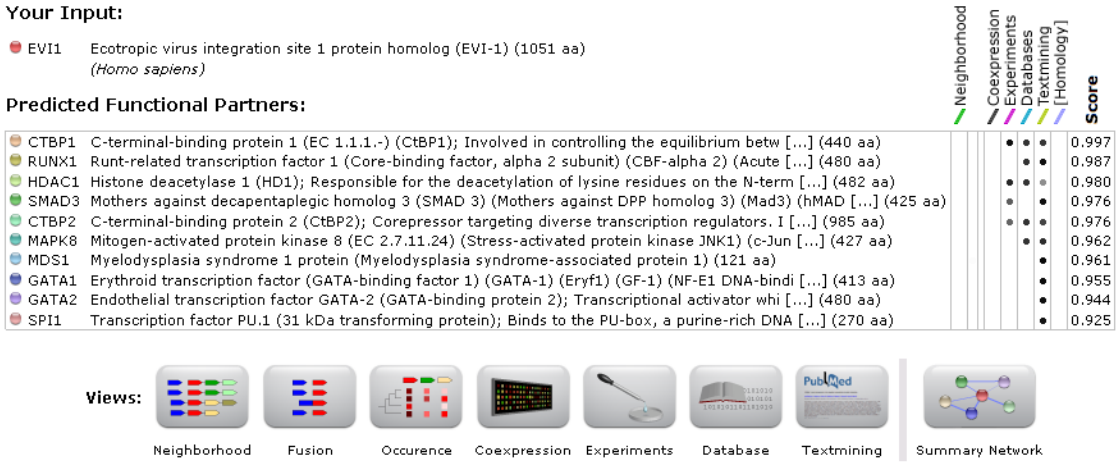


Debajo de las gráficas de la red aparece el navegador que permite escoger el modo de visualización que se desea, el botón more incluirá diez interacciones más, el criterio para mostrarlas serán las que tienen los valores de score más altos, y el botón less remueve interacciones con más bajo score.

Esta barra permite guardar los resultados en varios formatos, las imágenes de los diferentes modos, descripción de las moléculas involucradas en los resultados y una tabla con las relaciones y otros datos. Esta tabla es la que se usa para construir la estructura de la red.

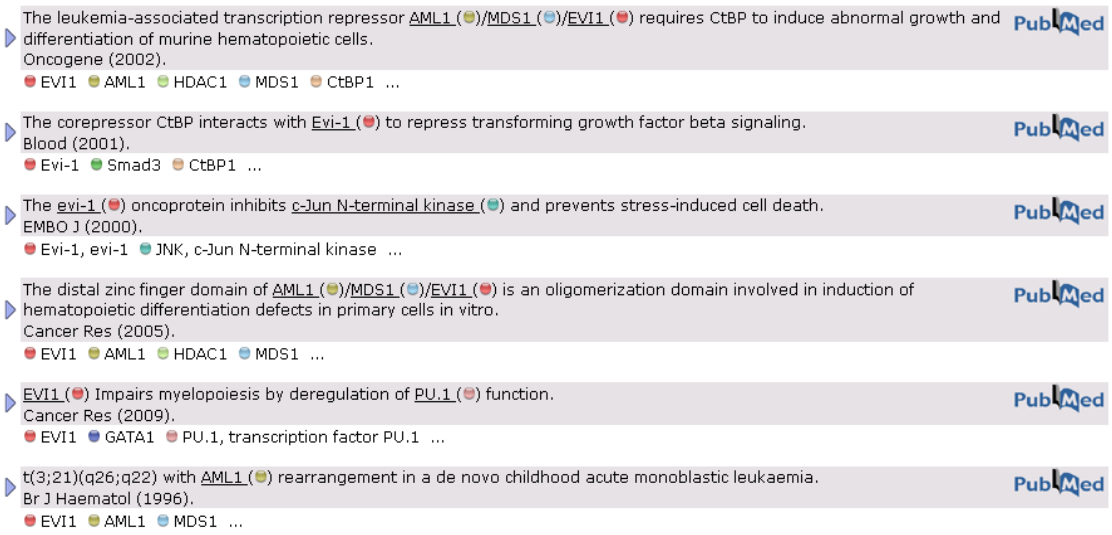
Se hicieron varias búsquedas variando parámetros.

Resumen de la predicción



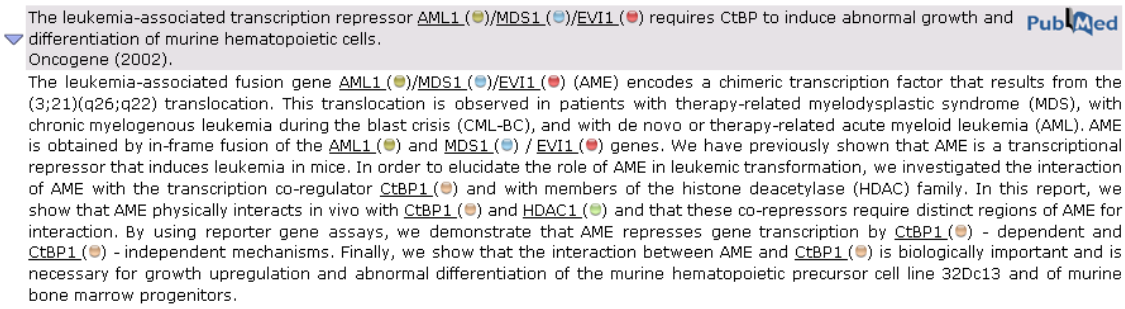
Esta segunda parte muestra una lista de las proteínas que resultan de la búsqueda, en el caso que sea búsqueda normal mostrará una sola entrada, y cuando la consulta sea para multiples proteínas las mostrará, el resto de proteínas podrán verse según el orden descendiente del Score combinado de cada asociación predicha, que es un valor que se calcula derivado de la probabilidad conjunta de las probabilidades de todos los canales de diferentes pruebas, bien sean estas puntuaciones que se han dado a vecindades, Text mining, co-ocurrencia, etc.

En la parte inferior muestra otro navegador con diferentes vistas, son estas las asociaciones obtenidas por cada fuente, se explicará la pertinente en nuestro estudio como es la textmining.



STRING lee automáticamente los abstracts de pubmed y otras fuentes, y también información que curada que se ha introducido en las bases de datos, donde curadores expertos han computarizado textos biológicos.

STRING identifica las proteínas y palabras claves que guarda en un diccionario, así en el texto va identificando relaciones, con un algoritmo hecho en pearl que funciona de la manera que se discutió la última vez.


The leukemia-associated transcription repressor AML1 / MDS1 / EVI1 requires CtBP to induce abnormal growth and differentiation of murine hematopoietic cells.
Oncogene (2002).

The leukemia-associated fusion gene AML1 / MDS1 / EVI1 (AME) encodes a chimeric transcription factor that results from the (3;21)(q26;q22) translocation. This translocation is observed in patients with therapy-related myelodysplastic syndrome (MDS), with chronic myelogenous leukemia during the blast crisis (CML-BC), and with de novo or therapy-related acute myeloid leukemia (AML). AME is obtained by in-frame fusion of the AML1 and MDS1 / EVI1 genes. We have previously shown that AME is a transcriptional repressor that induces leukemia in mice. In order to elucidate the role of AME in leukemic transformation, we investigated the interaction of AME with the transcription co-regulator CtBP1 and with members of the histone deacetylase (HDAC) family. In this report, we show that AME physically interacts in vivo with CtBP1 and HDAC1 and that these co-repressors require distinct regions of AME for interaction. By using reporter gene assays, we demonstrate that AME represses gene transcription by CtBP1 - dependent and CtBP1 - independent mechanisms. Finally, we show that the interaction between AME and CtBP1 is biologically important and is necessary for growth upregulation and abnormal differentiation of the murine hematopoietic precursor cell line 32Dc13 and of murine bone marrow progenitors.

STRING contiene información derivada de la literatura de dos tipos: información derivada de manera automática, y que se ha introducido manualmente. Este último es importado de diferentes bases de datos externas, donde curadores expertos que han leído y computarizado los textos de biología. Estos datos son accesibles en STRING en la vistas de 'base de datos' y 'experimentos que están debajo del resumen de predicción en el mismo navegador donde está la opción texmining.

Sin embargo, debido a la enorme cantidad de conocimientos biológicos publicados (incluidos los textos más antiguos), procesamiento de textmining accede a una gran cantidad de conocimientos, pero también es muy difícil para el software de supervisión de STRING no cometer errores en este proceso. Hay dos principales errores que pueden arrojar los resultados en la consulta, se debe idear alguna manera de procesar estos resultados para minimizar errores, estos errores principalmente se generan por falsos positivos y negativos falsos.

Resultados Falsos Positivos

La mayoría de falsos positivos se derivan de las dificultades para identificar proteínas mencionadas en un texto. Nombrar entidades biológicas ha sido notoriamente inconsistente, se incurre en ambigüedades porque a menudo existen varios nombres y abreviaturas para la misma proteína y nombres idénticos se han prestado a las familias distintas de proteínas (en diferentes organismos). STRING contiene una larga lista de sinónimos de una variedad de fuentes, y trata de

abstenerse de asignar una proteína cuando la situación es ambigua, pero todavía se cometen errores. Una segunda fuente de falsos positivos es el nivel de comprensión que tiene STRING para un texto que lee. No es un humano y no puede discernir con propiedad ciertos elementos en un párrafo, Aparte de los temas seleccionados actualmente en desarrollo, STRING no trata de "comprender" lo que realmente se dijo en un texto. La comprensión semántica de los textos científicos es extremadamente difícil, así que STRING simplemente busca proteínas que se mencionan juntos en los textos, con más frecuencia que lo que se esperaría por azar en función de su incidencia global. Este enfoque ha sido punto de referencia a nivel mundial y funciona bastante bien (las proteínas que se mencionan a menudo en conjunto son de hecho a menudo vinculadas funcionalmente). Pero, por supuesto, es un supuesto simplificador de nuevo de vez en cuando introducen errores.

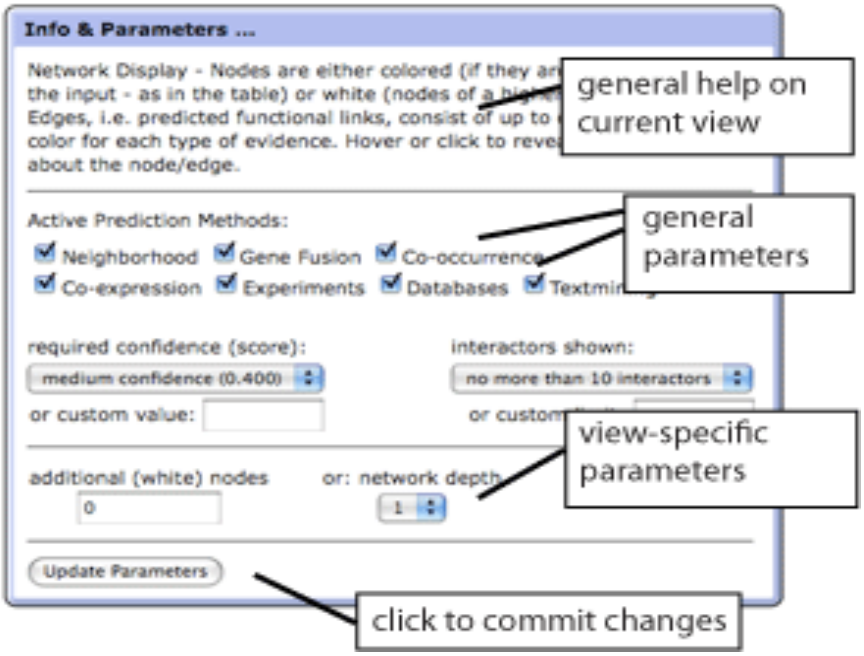
Resultados negativos falsos

Muchas asociaciones proteína-proteína conocida son ajenas al módulo de text mining. En primer lugar, STRING necesita tener acceso a los textos existentes: desde PubMed, puede leer en la actualidad sólo los resúmenes no - el texto completo de los artículos (sin embargo, otras fuentes de información tales como DGS o OMIM también las puede leer, y se espera añadir más fuentes en el futuro). En segundo lugar, STRING necesita saber de que organismo se está hablando en un texto- esto es a menudo difícil de "adivinar" a partir de unas pocas líneas de texto. En tercer lugar, muchas proteínas se denominan de una manera que es casi imposible para recoger a partir del texto Inglés - esto es a menudo el caso de organismos con una larga tradición de la genética (los genes son entonces a menudo nombrado por fenotipos, que a su vez, son palabras normales en Inglés). STRING mantiene una lista de 'Stop-words" palabras que son demasiado frecuentes como para ser útil como identificadores de la proteína.

Pero para cualquier el modulo text mining de STRING permite al usuario acceder a los textos pertinentes y descubrir información de manera manual que STRING no pueda discernir, para mejores resultados se pueden activar por el usuarios otros ítems como son la información extraida de las bases de datos, de experimentos, etc. Y text mining se puede desactivar también por el usuario. De esta manera, la cobertura y fiabilidad de la información puede ser adaptado a la cuestión que nos concierna.

Parámetros de consulta.

Esta caja de parámetros permite personalizar la consulta, son estos los métodos de predicción, número de interacciones a mostrar, el score, la profundidad, las relaciones funcionales o interacciones, consisten en hasta ocho líneas: un color para cada tipo de evidencia.

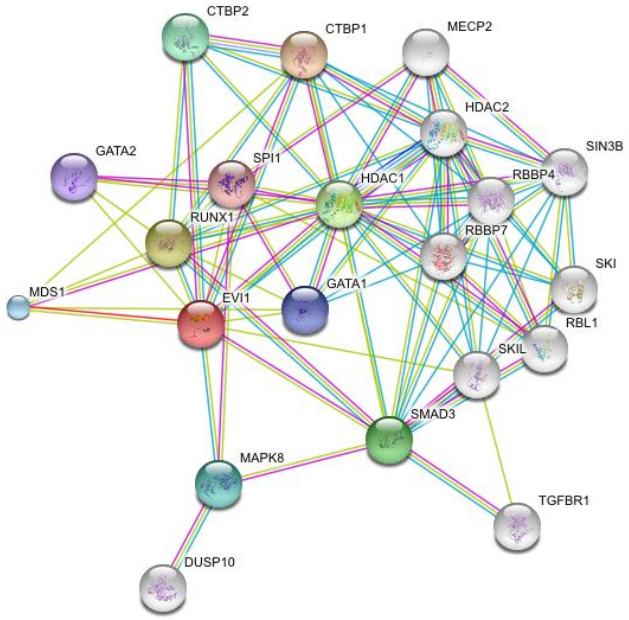


En el parámetro “profundidad de la red”. Un valor mayor que 1 significa que la búsqueda para las interacciones es iterativa - después de que un primer nodo de entrada, entrará otra vez para una siguiente ronda de búsquedas con los nodos asociados al nodo de entrada. Los nodos de una iteración más alta serán coloreados blancos. La salida se espera sea mayor y esto puede dar lugar a las imágenes bastante grandes que pueden tardar un rato para computar y para transferir. Esta característica permite “caminar” a través de la red de interacciones. Al dar click en cualquier nodo, permite un acoplamiento para utilizar ese nodo como la entrada poniéndola en el centro de la imagen. El uso repetido de este mecanismo permite explorar las regiones grandes de la red.

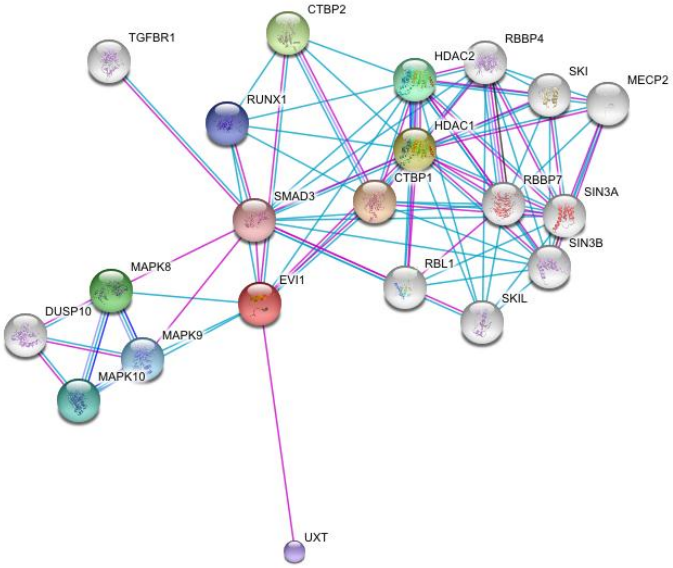
Se muestran unos ejemplos en las siguientes figuras, la primera se aumenta en 10 la primera red de evi1, la siguiente se hace consulta sólo por text mining y las

últimas son ejemplos de búsqueda múltiples, cabe anotar que se puede parametrizar de diferentes formas, variando el score, las interacciones a mostrar, las fuentes, etc.

Ejemplos



Esta red se obtuvo aumentando en 10 el numero de relaciones, y consultando por text mining, base de datos, experimentos y fusión.

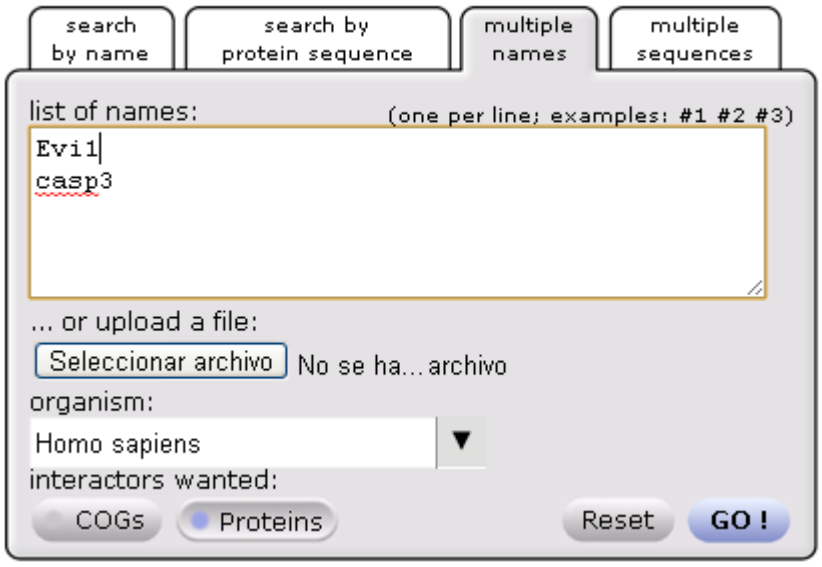


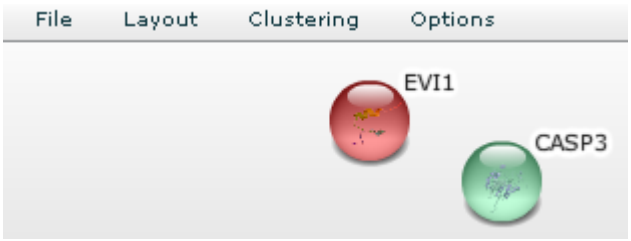
En esta consulta se eliminó la búsqueda por Text mining, y se dejo de reportar proteínas como GATA 1 y ÇGATa 2 entre otras.

Ejemplo para múltiples proteínas

En este aparte se ilustrará la búsqueda por dos o más proteínas para que se examine si se puede hacer este tipo de búsqueda para obtener una estructura de red más robusta.

Al hacer consulta con la proteína Evi1 nos dimos cuenta que no estaban muchas moléculas que había Mafe separado para la tabla que nos proporcionó, como estás moléculas ya han sido probadas hay unas que STRING no las reporta, o al menos no con la combinación de parámetros adecuados que pueden involucrar muchas moléculas que no son de interés cuando se escoge un umbral bajo. O se hace una consulta sin muchas restricciones, entonces se probó por ejemplo con Caspasa 3, que no aparecía en búsqueda inicial y se





En este caso de la consulta no hay interacciones reportadas entre estas dos moléculas o al menos con las características de la búsqueda.