Universidad de Granada

Departamento de Ciencias de la Computación e Inteligencia Artificial

Aproximación a la Medicina Personalizada mediante el desarrollo de nuevas metodologías en Inteligencia Artificial

Memoria que presenta

Carmen Navarro Luzón

Para optar al Grado de Doctor en Informática

Programa Oficial de Doctorado en Tecnologías
de la Información y la Comunicación

DIRECTORES:

Armando Blanco Morón

Carlos Cano Gutiérrez

Editor: Universidad de Granada. Tesis Doctorales
Autora: Carmen Navarro Luzón
ISBN: 978-84-9163-618-2
URI: http://hdl.handle.net/10481/48604

La memoria **"Aproximación a la Medicina Personalizada mediante el desarrollo de nuevas metodologías en Inteligencia Artificial"**, que presenta Dª Carmen Navarro Luzón para optar al grado de Doctor en Informática, ha sido realizada en el Dpto. de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada bajo la supervisión de los doctores D. Armando Blanco Morón y D. Carlos Cano Gutiérrez, ambos miembros del mismo Departamento.

La doctoranda Dª Carmen Navarro Luzón y los directores de la tesis D. Armando Blanco Morón y D. Carlos Cano Gutiérrez garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de los directores de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Guarantee, by signing this doctoral thesis, that the work has been done by the doctoral candidate under the direction of the thesis supervisors and, as far as our knowledge reaches, in the performance of the work, the rights of other authors to be cited (when their results or publications have been used) have been respected.

Lugar y fecha / *Place and date*: Granada, diciembre de 2016.

La doctoranda / *Doctoral candidate*

Carmen Navarro Luzón

Los directores / *Thesis supervisors*

Armando Blanco Morón

Carlos Cano Gutiérrez

Agradecimientos

Hay muchas personas sin las cuales no habría sido capaz de llevar a cabo este proyecto. Me gustaría aprovechar este momento para agradecerles su apoyo en este largo camino.

En primer lugar quiero dar las gracias a mis directores de tesis, Armando Blanco y Carlos Cano, por ayudarme a avanzar en este complicado proceso con vuestro conocimiento y experiencia, pero, sobre todo, por vuestra amistad y compañía a lo largo de todos estos años. Sois un apoyo que va mucho más allá de lo académico.

He tenido la suerte de poder desarrollar parte del trabajo de esta tesis en dos centros de investigación donde me he sentido como en casa. Quiero por ello agradecer a Boris Adryan, mi tutor en el Cambridge Systems Biology Centre, Reino Unido, su tutela y su guía a lo largo de mi estancia, y a su grupo de investigación al completo por todo lo aprendido y su buena recepción. En particular, mis agradecimientos a Daphne, Xiaoyan, Jamie y Shlomit, por su compañía, apoyo y amistad.

También quiero agradecer a Roderic Guigó su acogida en su grupo de investigación en el Centro de Regulación Genómica de Barcelona. Me gustaría agradecer a todo su grupo el apoyo y el cálido recibimiento del que disfruté. En particular, mi más sincero agradecimiento a Rory Johnson por su tutela y apoyo, que se han extendido fuera del periodo de mi estancia. A Juna, Andrés, Tommaso, y Reza por todos esos cafés en la terraza del “dry lab”.

También me gustaría dar las gracias a Paul Lizardi, tanto por sus inestimables enseñanzas como por transmitirme su gran pasión por la ciencia.

Muy especialmente quiero agradecer a mis amigos, los que siempre habeis estado ahí para tenderme una mano justo cuando lo he necesitado, independientemente de la distancia a la que nos encontremos. Mi alegría es poder celebrar esto con vosotros.

De verdad pienso que nunca habría llegado hasta aquí, ni a ninguna parte en general, sin el apoyo de mi familia. De corazón doy las gracias a mis padres, Silvia y Luis. En vosotros me veo a diario y espero hacerlo siempre. Por vuestro cariño incondicional, gracias por apoyarme y alentarme en todos mis proyectos. A Celia y Luis, por ser más que hermanos, amigos, y más que amigos, hermanos.

A vosotros os dedico esta tesis.

Resumen

Los últimos años han visto florecer y consolidarse a un extenso catálogo de metodologías experimentales, las cuales han dado lugar a vastas cantidades de información biológica. La extracción de conocimiento de esta enorme cantidad de datos representa a día de hoy un reto. La Bioinformática constituye, por tanto, una herramienta imprescindible para entender los complejos procesos que guían los entresijos de la maquinaria celular.

En esta memoria se presentan metodologías de Inteligencia Artificial para adquirir conocimiento sobre procesos de regulación génica a partir de datos biológicos heterogéneos, con el fin de contribuir al avance de la Medicina Personalizada. Las metodologías presentadas abarcan tres ámbitos: extracción de conocimiento de redes biológicas heterogéneas, búsqueda de elementos y módulos reguladores en el genoma y análisis de entidades biológicas de renovado interés: los ARN largos no codificantes.

En primer lugar, se propone una metodología general para la priorización en redes heterogéneas que puede ser aplicada a un número arbitrario de dominios biológicos gracias a un modelo de abstracción. Esta metodología se aplica a la búsqueda de nuevas aplicaciones para medicamentos comercializados.

Posteriormente, se acomete la búsqueda de elementos reguladores de la expresión de los genes. Se proponen una medida difusa de influencia de mutaciones en lugares de interés regulador como los lugares de unión de factores de transcripción, y una metodología difusa de minería de *itemsets* frecuentes para la búsqueda de módulos reguladores.

Por último, se buscan elementos reguladores en los ARN largos no codificantes mediante análisis computacionales que integran características propias del ARN: su estructura secundaria y su dirección de transcripción.

Abstract

The recent years have seen the bloom and establishment of an extensive catalog of experimental methodologies that have generated an immense amount of biological information. Extracting knowledge from these massive amounts of data represents a challenge that can only be overcomed by computational approaches. Bioinformatics is, in this sense, essential to understand the complex processes that guide the sophisticated function of the cell machinery.

This dissertation provides Artificial Intelligence methodologies to acquire knowledge about gene regulatory processes based on heterogeneous biological data, in the path towards Personalized Medicine. The proposed methodologies encompass three areas: knowledge extraction from heterogeneous biological networks, genomic regulatory elements and modules search, and the study of biological entities of increasing interest: long non-coding RNAs.

First, a general methodology for heterogeneous network prioritization, is proposed. This approach uses an abstraction model of heterogeneous networks in order to be applied to an arbitrary number of biological domains. The proposed methodology is then applied to a problem of interest: finding new applications for commercialized drugs, a process known as drug repositioning.

Secondly, we propose a methodology for a fuzzy assessment of mutations' regulatory potential according to their placement in putative transcription factor binding sites. Moreover, we propose a fuzzy itemset mining methodology for *cis*-regulatory module search.

Finally, we search for regulatory elements in long non-coding RNAs. We propose two computational methodologies that exploit two RNA-specific features, respectively: RNA secondary structure and transcription strand.

Índice general

Introducción

Desde la publicación del primer genoma humano de referencia [1], las tecnologías de secuenciación han sido partícipes de un vertiginoso avance. Atrás quedó el titánico esfuerzo del Proyecto Genoma Humano durante cerca de catorce años con un coste de alrededor de 2.700 millones de dólares que supuso el primer proceso [2]. Actualmente, la secuenciación de un genoma humano completo puede ser realizada en cuestión de días y a cambio de tan sólo unos pocos miles de dólares [2].

El estudio y la comprensión del organismo humano se ha beneficiado de este tremendo abaratamiento y mejora de las tecnologías de secuenciación en la aparición de las que se conocen actualmente como tecnologías de secuenciación NGS y de tercera generación. Dichas tecnologías han puesto al alcance de la comunidad científica cantidades descomunales de datos que no pueden ser entendidos sin la aplicación de metodologías computacionales y de Inteligencia Artificial. Del análisis de todos estos datos generados por laboratorios surgen constantemente nuevos resultados que no sólo aportan conocimiento sobre la naturaleza de los procesos celulares, sino que a su vez generan nuevas fuentes de datos que estudiar, aumentando exponencialmente la cantidad masiva de información a analizar.

En este ecosistema en continua ebullición se desarrolla la Bioinformática, un área de investigación interdisciplinar que trata de arrojar luz sobre complejos procesos biológicos mediante el uso de metodologías computacionales. En particular, el campo de la Medicina Personalizada ha cobrado en los últimos años gran protagonismo, al tratarse de una aplicación directa del conocimiento biológico que redunda en beneficio de la salud pública. La Medicina Personalizada

tiene como meta la explotación de las características individuales de cada paciente para mejorar su diagnóstico y tratamiento. Dicho de otro modo: la Medicina Personalizada integra el conocimiento general ya existente sobre el funcionamiento del organismo humano con las particularidades que hacen único al paciente, ayudando al personal médico a emitir diagnósticos más exactos y a elaborar tratamientos específicos para cada individuo. Uno de los principales retos de este planteamiento recae sobre las Ciencias de la Computación: la adaptación de la medicina moderna a estas prácticas en los sistemas de salud pública sólo será posible a través de herramientas eficientes y fiables que sean capaces de analizar estas enormes cantidades de información.

La Medicina Personalizada es hoy día por tanto una realidad casi palpable. El análisis del genoma de un paciente se ha convertido en algo muy asequible gracias a los últimos avances tecnológicos, y el conocimiento de los mecanismos celulares es ahora mucho más sofisticado. Sin embargo, existen aún numerosas incógnitas sobre el funcionamiento del organismo que necesitan ser respondidas antes de que la Medicina Personalizada pueda convertirse en una realidad clínica. El presente trabajo pretende contribuir a avanzar en el esclarecimiento de estas incógnitas mediante el desarrollo de metodologías de Inteligencia Artificial que permitan arrojar luz sobre las particularidades que hacen únicos a los individuos y poder así adecuar a ellas sus tratamientos. A su vez, el conocimiento obtenido en el proceso pretende contribuir a los modelos existentes, cada vez más sofisticados, sobre el funcionamiento del organismo humano.

Objetivos

La hipótesis sobre la que se construye esta tesis es que los procesos biológicos asociados a una célula, principalmente dirigidos por su información genética, influyen de manera crucial en el estado de salud o enfermedad de un individuo. Por lo tanto, un mejor conocimiento de dicha información genética así como de los procesos a los que afecta y cómo éstos se relacionan, permitirá en un futuro mejorar la evaluación, diagnóstico y tratamiento de los pacientes, contribuyendo así a la Medicina Personalizada.

En este sentido, la finalidad que persiguen las aportaciones de este trabajo es la mejora de la comprensión de las posibles relaciones entre los elementos heterogéneos que componen la maquinaria biológica celular, tomando como protagonistas entidades clave en la biomedicina, como son los genes, las enfermedades y las variaciones genómicas que diferencian a unos individuos de otros, así como entidades de relativamente reciente y renovado interés científico como son los ARNs largos no codificantes, en el marco de la Medicina Personalizada.

Por tanto, el objetivo general del presente trabajo es el desarrollo de nuevas metodologías y herramientas computacionales para extraer conocimiento de datos biológicos heterogéneos, facilitando así el avance de la Medicina Personalizada. Con este fin, se han desarrollado metodologías para el estudio de entidades y procesos biológicos mediante el análisis de dos tipos de datos principales: redes heterogéneas de entidades biológicas y secuencias de ADN y ARN.

Estructura de la memoria

A continuación se presenta la estructura de la memoria enumerando los capítulos que la componen junto con una breve descripción de los temas a tratar en cada uno de ellos.

- Parte I: *Preliminares*. En esta parte se realiza una introducción a los términos y conceptos previos necesarios para comprender el trabajo realizado y expuesto en esta memoria.
- Parte II: *Contribuciones*. En esta parte se describen las contribuciones realizadas durante el desarrollo de la tesis.

  En el capítulo 2, *Priorización en redes biológicas heterogéneas*, se detallan las metodologías desarrolladas con la finalidad de inferir conocimiento a partir de redes biológicas heterogéneas construidas a partir de múltiples fuentes de información. Se propone ProphTools, una metodología general y flexible de propagación en redes heterogéneas que consta de un meta-modelo de dichas redes, lo que permite aplicar la metodología propuesta a cualquier dominio de utilización. Como muestra de la utilidad de esta metodología, dos herramientas son descritas: i) DiGSNP, una herramienta que prioriza mutaciones potencialmente reguladoras respecto a enfermedades; y ii) DrugNet, una metodología de propagación en redes heterogéneas aplicada a un ámbito de gran interés para la Medicina Personalizada: la búsqueda de nuevas aplicaciones para medicamentos ya comercializados, área conocida como reposicionamiento de medicamentos.

  A continuación, en el capítulo 3, *Detección de elementos reguladores en secuencias de ADN*, se describen las aportaciones basadas en tecnología difusa para la búsqueda de elementos reguladores en el ADN. Se detallan i) IntuitSNP, una herramienta para búsqueda de mutaciones funcionales en regiones reguladoras basada en el cálculo de la influencia de una mutación en un potencial lugar unión de factor de transcripción, y ii) CisMiner, una metodología para la búsqueda de módulos reguladores en *cis*.

Por último, en el capítulo 4: *ARNs largos no codificantes*, se proponen enfoques para la investigación en una entidad biológica cuyo interés se ha renovado en los últimos años por su gran interés regulador: los ARNs largos no codificantes. Se aborda la búsqueda computacional de motivos específicos de ARNs largos no codificantes en dos aproximaciones. La primera, RNAIntuit, teniendo en cuenta la estructura secundaria de plegado del ARN, con un caso de estudio relacionado con la carcinogénesis. La segunda, MoSI, un pipeline de análisis de secuencias que utiliza la dirección de transcripción para discernir motivos específicos de ARN.

- Parte III: *Conclusiones y trabajo futuro*. En esta parte se describen las conclusiones extraídas de los resultados presentados en esta memoria, así como los caminos de trabajo futuro que se abren a raíz de los mismos.
- Parte IV: *Publicaciones*. Enumeración de las publicaciones realizadas durante el proceso de desarrollo de la tesis.
- Bibliografía: Relación de las fuentes citadas en este trabajo.

Introduction

Since the publication of the first human reference genome [1], sequencing technologies have quickly evolved. We are now far from that huge first effort performed by the Human Genome Project, that lasted approximately fourteen years and costed around 2.7 billion dollars [2]. Nowadays, the sequencing of a whole human genome can be performed in days, for only a few thousand dollars [2].

The study and understanding of the human organism has benefited from this technology improvement and price reduction on the advent of next generation sequencing technologies (NGS) and third generation sequencing technologies. These technologies have made massive amounts of biological data available to the scientific community. Such data cannot be understood without the application of computational and Artificial Intelligence methodologies. From the analysis of all this information, new results are constantly being released by laboratories. These results do not only provide knowledge about the nature of cell processes, but also generate themselves new data sources to study, increasing exponentially the massive amount of information to be analysed.

In this ecosystem, Bioinformatics arises as an interdisciplinary field of research that designs computational methodologies in order to shed light on the complex biological processes. Furthermore, a field known as Personalized Medicine has been drawing attention in the past recent years, since it is a direct application of the acquired biological knowledge to the improvement of public healthcare. One of the main goals of Personalized Medicine is exploiting the individual features of each patient to improve his or her diagnosis and treatment. In other words, Personalized Medicine integrates the general existing knowledge about how organisms function

and the special features that make each patient unique, helping doctors to produce better diagnoses and to ellaborate personalized treatment for each individual. Computer Science can help overcome one of the main challenges of this approach: the adaptation of modern medicine to these practices in public healthcare systems is only possible through efficient and reliable tools that are able to analyze these massive amounts of data.

Today, Personalized medicine is close to become a reality. The fast analysis of a patient's genome has become feasible due to the last technological improvements, and knowledge of cell mechanisms has increased significantly. However, there are many questions about an organism's function that still need to be answered.

This work intends to contribute to clarify these questions, developing Artificial Intelligence methodologies that can help to shed some light on the individual features that make each patient unique, in order to better adequate treatments and diagnoses. In this sense, the knowledge obtained in the process also intends to contribute to existing models about the functioning of the human organism.

Objectives

The main hypothesis of this work is that biological processes associated to a cell, mainly directed by its genetic information, are key influencers on the health or disease state of an individual. Therefore, a better knowledge about such genetic information, along with a better understanding of the processes it affects and how these relate to each other, will allow us to improve diagnosis and treatment of patients, contributing to Personalized Medicine.

In this sense, the aim of the contributions presented in this work is the improvement of the understanding about the relationships between the heterogeneous elements that conform the complex cell machinery, taking as main actors key entities in biomedicine such as genes, diseases and genomic variations that differentiate individuals, along with entities of recent and renovated scientific interest, such as long non-coding RNAs, in the area of Personalized Medicine.

More specifically, the goal of this work is the development of new methodologies and computational tools to extract knowledge from heterogeneous biological data, helping in the progress towards Personalized Medicine. To this end, several methodologies have been developed to study biological entities and processes by means of the analysis of two main types of data: heterogeneous biological networks and DNA and RNA sequences.

Structure of this work

The structure of this work is detailed below, enumerating its chapters together with a description of the topics described in each one.

- Part I: *Introduction*. In this part we provide an introduction to the concepts and terminology necessary to understand the developed work and described in this dissertation.
- Part II: *Contributions*. In this part, we describe the contributions developed during the development of this work.

  In chapter 2, *Prioritization on heterogeneous biological networks*, we describe the methodologies developed to infer knowledge from heterogeneous biological networks, built from multiple data sources. First, ProphTools, a general and flexible prioritization methodology that can be applied to any domain of interest, is proposed. As a proof of its versatility and usefulness, two methodologies are described: i) DigSNP, a prioritization tool to obtain potentially regulatory variations related to diseases; and ii) DrugNet, a heterogeneous biological network propagation tool applied to an area of great interest for Personalized Medicine: drug repositioning.

  Next, in chapter 3, *Detection of regulatory elements and modules in DNA sequences*, we describe the contributions based on fuzzy technology for the search of regulatory elements and modules in DNA sequence. More specifically, we describe two approaches: i) IntuitSNP, a tool for the search of functional mutations in regulatory regions, based on the influence of a mutation in a putative transcription factor binding site; and ii) CisMiner, a methodology for the search of *cis*-regulatory modules.

  Chapter 4: *Long non-coding RNAs* describes research on a biological entity that has been drawing interest recently: long non-coding RNAs. We address the computational search of long non-coding RNA-specific motifs in two different approaches: i) RNAIntuit, a tool that exploits RNA-secondary structure, and it is applied to a case study related to carcinogenesis; and

ii) MoSI, a sequence analysis pipeline that use transcription strand bias as a signal to obtain long non-coding RNA-specific motifs.

- Part III: *Conclusions and further work*. In this part we describe the conclusions drawn from the results presented in this dissertation, together with the open lines of future work.
- Part IV: *Publications*. List of all publications that were written during the development of this work.
- Bibliography: List of sources cited in this work.

Parte I

Preliminares

Capítulo

1

Conceptos previos

La interdisciplinaridad del presente trabajo requiere introducir el área de conocimiento de cada una de las disciplinas involucradas para poder comprenderlo en su conjunto. En este capítulo se introducen los aspectos básicos, tanto biológicos como computacionales, relacionados con el desarrollo de esta tesis. En la primera sección se describen los conceptos relativos al funcionamiento de la célula, haciendo especial hincapié en los mecanismos de regulación de los genes, de especial relevancia en el desarrollo de las metodologías propuestas. A continuación se introducen unas nociones básicas sobre teoría de grafos, redes complejas y el análisis de los mismos en el marco de la biología computacional. Por último, se detallan los términos necesarios sobre la lógica difusa y su aplicación a las técnicas de análisis de secuencia de ADN.

1.1. Introducción a la biología molecular

En esta sección se presentan nociones básicas sobre biología molecular relevantes al desarrollo de las líneas de trabajo descritas en la segunda parte. Para una descripción más completa y exhaustiva de estos conceptos, recomendamos la consulta de las referencias [3, 4]. En primer lugar, se describe el funcionamiento y estructuración de la información genética. Esta información es la base sobre la que actúan procesos dinámicos como la regulación de los genes, detallada posteriormente.

Por otro lado, el continuo proceso de estudio del ADN y el aumento de la precisión de las tecnologías que nos permiten obtener información sobre el mismo mejoran incrementalmente el modelo con el que lo representamos, incorporando nuevos procesos y entidades clave que nos permiten comprender mejor su funcionamiento. En este sentido, también son descritos en detalle los ARNs largos no codificantes, unas entidades biológicas clave en numerosos procesos celulares que han adquirido un interés científico renovado debido a su gran potencial regulador.

El análisis de todos estos procesos biológicos durante los últimos años han dado pie a la consolidación del ámbito conocido como Medicina Personalizada o de Precisión, concepto cuyas bases y situación actual se describen en el último apartado.

1.1.1. Estructura de la información genética

Todos los seres vivos estamos formados por **células**. De hecho, la célula es considerada la unidad elemental de la vida y, en este sentido, los seres vivos se clasifican en unicelulares o pluricelulares según estén formados por una sola de estas unidades o por un conjunto de las mismas. En los organismos pluricelulares pueden existir multitud de tipos celulares, cuya funcionalidad y características son principalmente determinadas por las proteínas que se sintetizan en ellas.

Una **proteína** es un compuesto bioquímico que realiza una función biológica concreta, cuya estructura es determinada por el **ADN**[1] presente en el núcleo celular en el caso de los organismos eucariotas[2], frente a los organismos procariotas tales como las bacterias, cuyo material genético se encuentra disperso en el citoplasma. El ADN contiene las instrucciones necesarias para fabricar la extensa variedad de proteínas presente en el organismo.

Esta información también incluye la maquinaria necesaria para coordinar todos los procesos de fabricación (**regulación génica**) e incluso para generar una nueva célula a su imagen (**replicación**), que contendrá una nueva copia de su material genético. En este sentido, la célula no es sólo la unidad fundamental del organismo vivo, sino también el vehículo a través del cual se transmite la información genética que define cada especie.

Del mismo modo que el código fuente del núcleo de un sistema operativo contiene la información necesaria para gestionar el hardware y la ejecución de los procesos en un computador, podría decirse que el ADN orquesta el funcionamiento de la célula, más allá de contener la información estructural para la construcción de todas las proteínas que lo conforman.

La información necesaria para el correcto funcionamiento, desarrollo y reproducción de todo organismo vivo está contenida en su **ADN**. El ADN es una molécula en forma de doble hélice, formada por dos largas cadenas de nucleótidos. Un **nucleótido** es una molécula formada por un azúcar (la desoxirribosa) unida a un grupo fosfato y una base, que puede ser adenina (A), guanina (G), citosina (C) o timina (T). Sustentándose sobre la teoría postulada por Chargaff en 1950 en la que éste enunció que en todo organismo vivo el número de bases A-T era similar al de G-C [5] y los trabajos de rayos X de Rosalind Franklin [6], Watson y Crick propusieron la ya famosa estructura de doble hélice del ADN que muestra la figura 1.1 en 1953 [7]. En dicha estructura, cada base se une a su complementaria (A-T, C-G).

Ahora bien, a pesar de su reducido tamaño, cada célula guarda una copia completa del ADN del organismo al que pertenece compactada en su núcleo

[1] Ácido desoxirribonucleico

[2] En este texto siempre nos referiremos a organismos eucariotas.

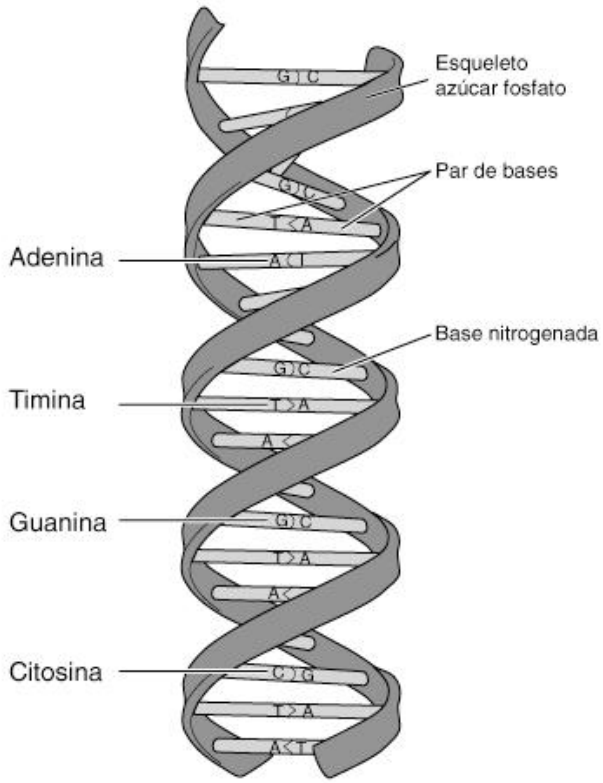


Figura 1.1: ***Estructura de doble hélice del ADN****. Sustentándose la primera regla de Chargaff y los trabajos de rayos X de Rosalind Franklin, Watson y Crick descubrieron la estructura de doble hélice del ADN, en la que cada base o nucleótido se une a su complementario (A-T, C-G).*

celular en forma de **cromosomas**. De otra forma sería imposible almacenar esta información, ya que cada célula humana contiene aproximadamente dos metros de ADN [8]. Esta compactación es realizada por un tipo especial de proteínas asociadas al ADN, las cuales se encargan de su plegado y estructuración en niveles de organización superiores. Un cromosoma es una sección lineal de ADN compuesta por una sola molécula, la cual adquiere su característica forma de X durante el ciclo de división celular. Al conjunto que forman el ADN y estas proteínas encargadas de su estructuración se le denomina **cromatina**.

Dentro de estas largas secuencias de ADN se encuentran los **genes**. Por lo general, un gen se define como una porción consecutiva de ADN que codifica la formación de una proteína[3] o grupo de proteínas relacionadas entre sí (isoformas o variantes). Los genes contienen la secuencia de **codones** necesaria para generar las secuencias de aminoácidos que darán lugar a una serie de proteínas. Una particularidad de los genes en los organismos eucariotas radica en que los genes contienen a su vez secciones codificantes y no codificantes. Las secciones codificantes reciben el nombre de **exones** y son las que dan lugar a la secuencia de aminoácidos final de la proteína. Las secciones no codificantes o **intrones** son secuencias mucho más largas que no aparecerán en el ARNm, al ser omitidas en una etapa del proceso de transcripción denominada *splicing*.

Pese a la gran importancia de las proteínas en el funcionamiento de un organismo, los genes conforman tan sólo un 2% de la totalidad de la secuencia de ADN del genoma humano [10]. Son las zonas **intergénicas** (como su nombre indica, las áreas que se encuentran entre los genes), las que abarcan la mayoría del genoma. Dicho de otra forma, si el ADN humano fuera un océano, los genes serían tan sólo pequeñas islas alejadas entre sí. Sin embargo, actualmente se sabe que más de tres cuartas partes del genoma pueden presentar actividad de transcripción [11], una proporción muy amplia frente a la cantidad de genes que constituyen el genoma humano. Poco a poco se van descubriendo funcionalidades presentes

[3] Por simplicidad en la introducción se ha mantenido esta definición algo estricta, ya que hay genes que dan como resultado otros productos biológicos e incluso a las secuencias que transcriben los lncRNAs, entidades que describiremos más adelante, se las ha llegado a llamar *un nuevo tipo de genes* [9].

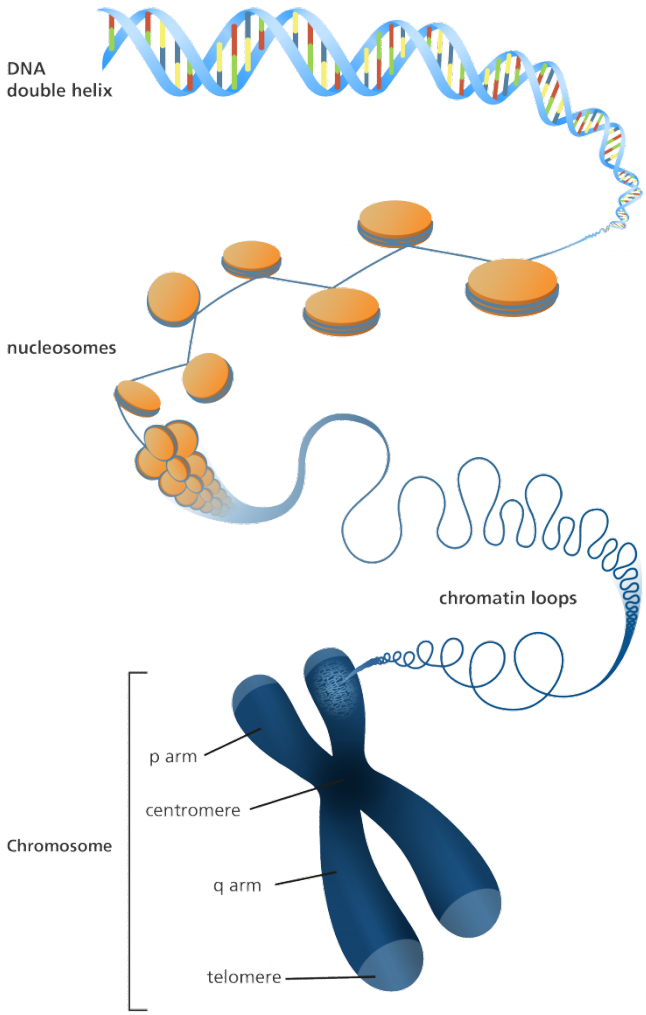


Figura 1.2: ***Niveles de estructuración del ADN****. En la figura se muestra la estructura del ADN en un cromosoma. El ADN se empaqueta alrededor de proteínas llamadas histonas, formando nucleosomas. Estos nucleosomas se estructuran formando bucles de cromatina, los cuales se compactan a su vez en forma de cromosoma (imagen de Genome Research Limited).*

en estas zonas, antes conocidas como *junk DNA*[4], entre las cuales aparece la ya mencionada regulación de los genes, de especial interés en este trabajo.

1.1.2. Dogma central de la biología molecular

Hasta el momento se ha explicado de forma sencilla cómo se estructura la información que necesitan las células para fabricar proteínas, reproducirse y, en términos generales, funcionar correctamente. A continuación se pretende profundizar en cómo dicha información es utilizada por la célula con este fin, es decir, cuál es la relación entre el ADN y las proteínas, las unidades de trabajo de las células, y qué otros procesos y entidades biológicas intervienen en este complejo sistema.

La figura 1.3 muestra un esquema de las relaciones que existen entre ADN y proteínas, las cuales presentan cierta circularidad: el ADN contiene instrucciones para la síntesis de todas las proteínas que la célula necesita; el ARN es una molécula que actúa como intermediaria entre las instrucciones del ADN y la formación de proteínas; y proteínas específicas a su vez participan en la síntesis y el metabolismo del ADN y el ARN. El ADN, además, se **replica**, durante un proceso de reproducción de la célula conocido como ciclo celular, y que consta principalmente de dos etapas: la síntesis, donde se realiza una copia del ADN, y la mitosis, donde se reparte el material genético y el citoplasma. A este flujo de información se le denomina el **dogma central de la biología molecular**[5].

El proceso de **transcripción** es aquel por el cual un trozo de ADN (gen) sirve como molde para la producción de una secuencia de ARN mensajero o ARNm, que es enviada fuera del núcleo para ser procesada posteriormente por un ribosoma. El ARNm contiene una secuencia complementaria a la secuencia de ADN que la genera, con la salvedad de que la timina (T) es reemplazada por uracilo (U). El uracilo, al igual que la timina, se empareja con la adenina. Este proceso es realizado

[4] En castellano traducido al poco afortunado término *ADN basura*, refiriéndose a la ausencia de utilidad conocida en el momento de su descubrimiento.

[5] Hoy en día se sabe que esta relación entre ADN, ARN y proteínas es más sofisticada, ya que, por ejemplo, no todo el ARN transcrito termina convirtiéndose en una proteína.

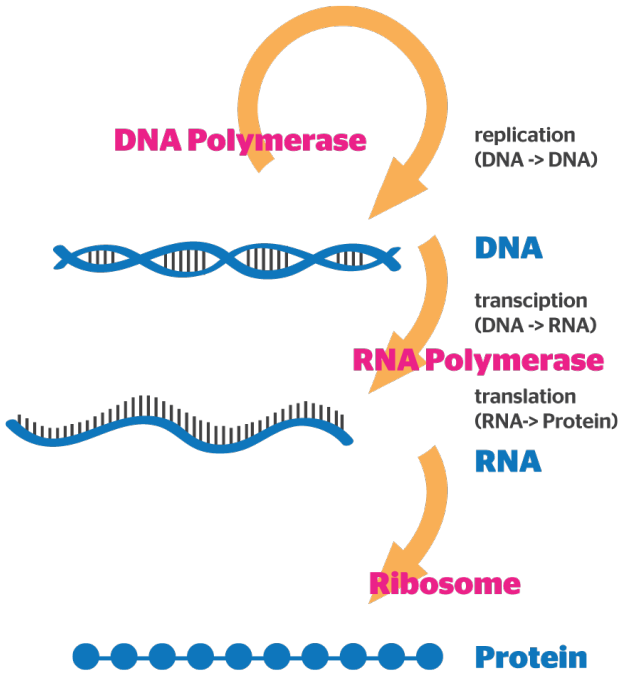


Figura 1.3: ***Dogma central de la biología molecular****. Flujo de información entre ADN, ARN y proteínas. El ADN se reproduce medidante procesos de* ***replicación****, se crea una copia de una de sus cadenas simples en forma de ARN durante la* ***transcripción*** *y dicha cadena es* ***traducida*** *por un ribosoma para formar la secuencia de aminoácidos que darán lugar a la proteína.*

por unas enzimas llamadas ARN polimerasas, las cuales recorren de forma lineal la secuencia de ADN a copiar, desenrollándola para así poder exponer una sección del ADN al emparejamiento de bases. Para que estas enzimas puedan iniciar el proceso de transcripción, necesitan la actuación de unas proteínas concretas, conocidas como factores generales de transcripción.

Una vez el ARNm es generado y enviado fuera del núcleo, se produce en los ribosomas el proceso de **traducción**. Los ribosomas son los complejos macromoleculares encargados de sintetizar proteínas. Durante la traducción, el ARNm es procesado secuencialmente, de forma que cada grupo de tres bases

Tabla 1.1: *Código genético. A cada aminoácido (AA) le corresponden uno o más codones. INI y TER indican señales de inicio y terminación de transcripción, respectivamente.*

AA	**Codones**	**AA**	**Codones**
Ala	GCT, GCC, GCA, GCG	Leu	TTA, TTG, CTT, CTC, CTA, CTG
Arg	CGT, CGC, CGA, CGG, AGA, AGG	Lys	AAA, AAG
Asn	AAT, AAC	Met	ATG
Asp	AAT, AAC	Phe	TTT, TTC
Cys	TGT, TGC	Pro	CCT, CCC, CCA, CCG
Gln	CAA, CAG	Ser	TCT, TCC, TCA, TCG, AGT, AGC
Glu	GAA, GAG	Thr	ACT, ACC, ACA, ACG
Gly	GGT, GGC, GGA, GGG	Trp	TGG
His	CAC	Tyr	TAT, TAC
Ile	ATT, ATC, ATA	Val	GTT, GTC, GTA, GTG
INI	ATG	**TER**	TAA, TGA, TAG

codifica por un aminoácido. Cada una de estas agrupaciones de tres bases recibe el nombre de **codón**. Cada codón se asocia a un aminoácido a través de una molécula llamada transfer ARN o tARN. Estas moléculas poseen una estructura tridimensional que por un lado se une a una secuencia de tres aminoácidos (extremo conocido como anticodón) y en el otro extremo está unida mediante enlace covalente al aminoácido correspondiente. En la figura 1.4 se puede observar el proceso completo en una célula eucariota. Cada tipo de tARN sólo está unido a un único tipo de aminoácido. Ahora bien, sólo existen 20 tipos de aminoácidos y $4^3 = 64$ posibles combinaciones de tres nucleótidos, así que varios codones (y, por tanto, distintos tipos de tARN) pueden corresponderse con el mismo aminoácido. Esta correspondencia entre codones y aminoácidos que se conoce como **código genético**. El código genético completo se muestra en la tabla 1.1. Cabe destacar entre ellos los llamados codones de terminación, que marcan el final del encadenamiento de aminoácidos que da lugar a la proteína.

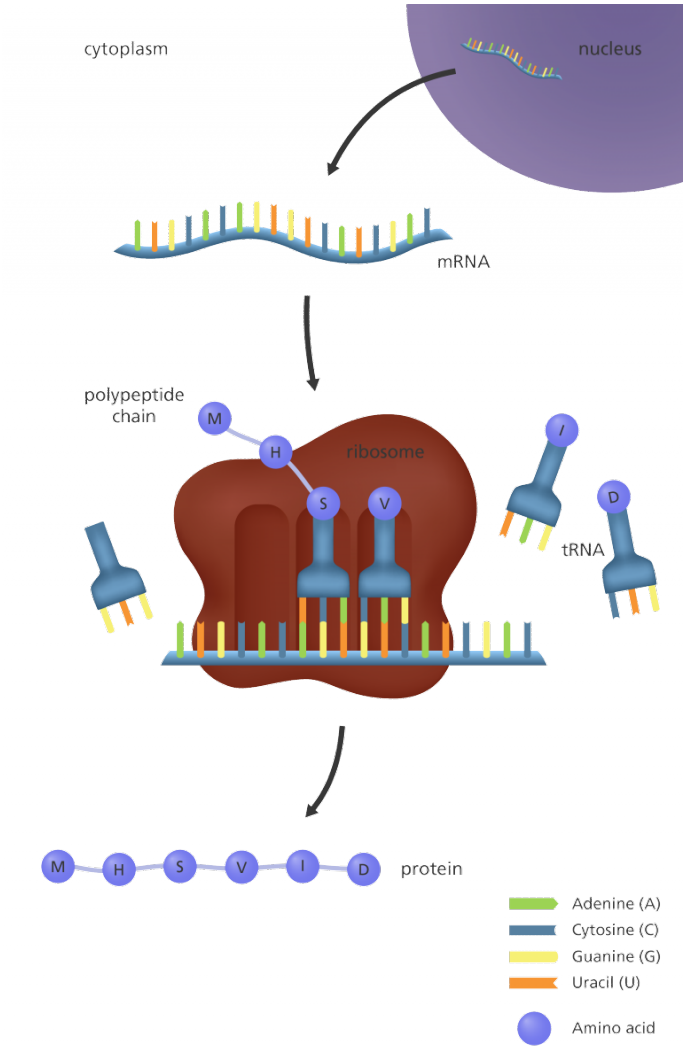


Figura 1.4: ***Traducción del ARN****. El ARNm transcrito es enviado fuera del núcleo, donde es procesado por ribosomas. En ellos, los transferARN se unen a sus codones complementarios, generando la cadena de aminoácidos que dará lugar a la proteína (imagen de Genome Research Limited).*

1.1.3. La regulación y expresión de los genes

A pesar de que todas las células contienen el mismo ADN, no todas desempeñan la misma función ni poseen las mismas propiedades. Una célula perteneciente a tejido óseo es diferente a un glóbulo rojo, y sin embargo el código genético presente en sus núcleos es el mismo. Esto se debe a que las características de una célula vienen determinadas por las proteínas que sintetiza, ya que estas moléculas se agrupan en bloques que conforman estructuras celulares más complejas, como tejidos u órganos. Dicho de otra manera, si entendiéramos a la célula como una fábrica de proteínas, el ADN en su núcleo contiene la información necesaria para fabricar **todas** las piezas que podría necesitar dicha célula alguna vez para funcionar, pero esto no significa que todas las células fabriquen todas las piezas al mismo tiempo, es decir: no todos los genes se **expresan** al mismo tiempo, ni lo hacen en todas las células. Esta idea es muy relevante para la comprensión del presente trabajo, ya que hay que hacer hincapié en que el ADN modela tanto las estructuras fijas de la biología como sus procesos dinámicos. Es necesario, pues, entender a la célula como un una entidad dinámica, y a su secuencia de ADN como al manual de instrucciones cuyo estudio nos permitirá entender la realidad de sus procesos. En este sentido, la expresión de los genes es coordinada por una serie de entidades biológicas que a su vez están orquestadas desde el contenido mismo del ADN, y a toda esta infraestructura que define qué genes se expresan, cuándo y en qué medida se la conoce como **regulación génica**.

Hasta ahora se han descrito los procesos por los cuales el ADN se transforma en proteínas. Sin embargo, no sólo se trata de saber qué aminoácidos forman una determinada proteína para poder fabricarla. Son necesarios mecanismos para **regular** la expresión de los genes. Los genes precisan por tanto recibir señales de activación o desactivación para matizar la cantidad de proteínas que están produciendo. Estas señales se producen, entre otras zonas, en las regiones promotoras de los genes, ubicadas en la zona inmediatamente anterior al comienzo del gen. En esta tarea de activación/desactivación juegan un papel clave los factores de transcripción o TFs[6], un tipo concreto de proteínas capaces de

[6]Transcription Factors. Es importante notar que estos factores de transcripción son un conjunto

interactuar con las zonas promotoras de determinados genes y en determinados segmentos del ADN, provocando una reacción que facilita o entorpece el ensamblaje de la ARN polimerasa, influyendo finalmente en la producción de proteínas. Los segmentos de ADN a los que se unen los TFs reciben el nombre de lugares de unión de factores de transcripción o TFBSs[7]. Los TFBSs se corresponden con secuencias de ADN breves (de unos 5-20 nucleótidos) que responden a un patrón o motivo determinado[8]. Se dice que un TF tiene una alta afinidad de unión con una secuencia de un TFBS. Si las zonas promotoras se pueden considerar interruptores que activan y desactivan la transcripción de los genes, los TFs son actuadores que juegan un papel importante en su encendido y apagado. Además, las proteínas reguladoras no suelen actuar de forma individual, sino que se coordinan en módulos de regulación, es decir, un gen podría necesitar varios TFs distintos coordinados en los TFBSs de su zona promotora, para poder transcribir ARN.

1.1.4. Variaciones en el ADN

El ADN de nuestras células nos define como especie. Como individuos *Homo sapiens*, los humanos compartimos el 99.9 % de nuestro genoma [12]. Las diferencias existentes entre personas se deben a ese 0.1 % de variabilidad que aparece en el ADN en forma principalmente de variaciones estructurales y un gran número de variaciones de una sola base, conocidas como SNVs[9]. Entre ellas, aquellas en las que un nucleótido o base es sustituido por otro, ampliamente cononcidas como SNPs[10], han sido extensamente estudiadas y guardan relación con numerosas características o fenotipos, desde diferencias en la apariencia física hasta la aparición de enfermedades [13].

más amplio que el de los factores generales de transcripción, los cuales forman parte de la maquinaria general de transcripción.

[7] Transcription Factor Binding Sites.

[8] En este texto se utilizan los términos TFBS y motivo de forma análoga, ya que existe una correspondencia entre ambos conceptos

[9] Single Nucleotide Variation. Variación de una única base.

[10] Single Nucleotide Polymorphisms, polimorfismos de nucleótido único o simple

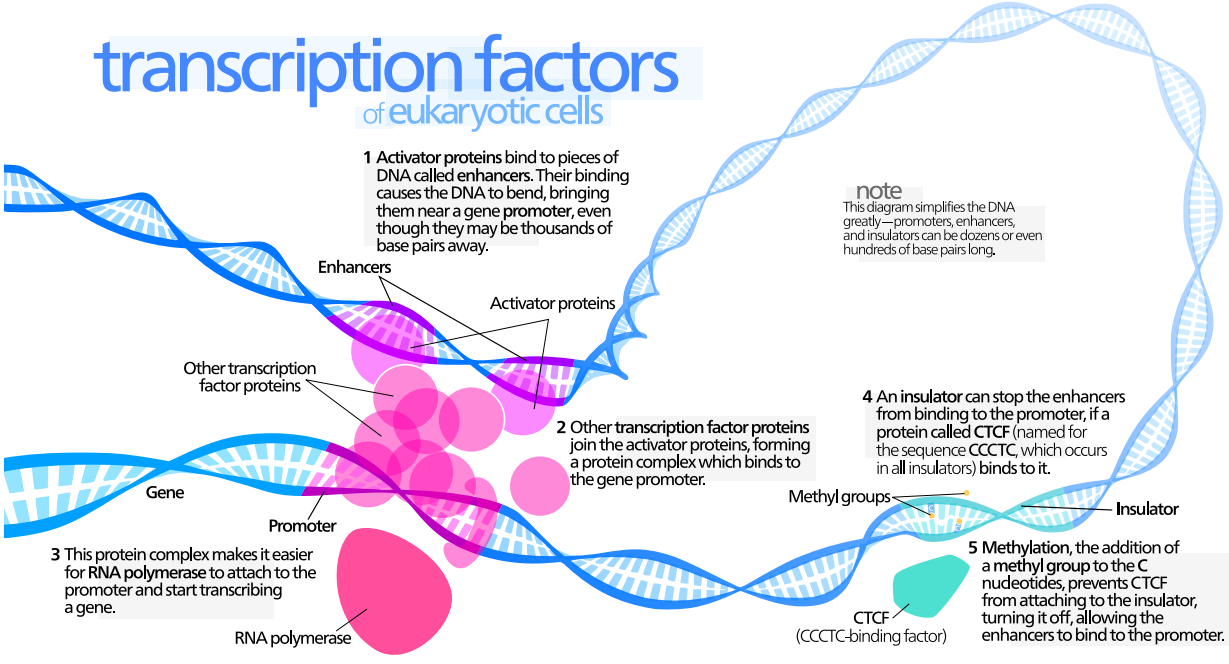


Figura 1.5: ***Factores de transcripción, zonas promotoras y regulación***[a]. *Los Factores de Transcripción se unen a unas secuencias concretas llamadas Lugares de Unión de factores de transcripción, o TFBS, en unas áreas más amplias llamadas regiones promotoras de los genes, iniciando un complejo proceso biológico que desemboca en la transcripción del gen.*

[a]Imagen de Kelvinsong (https://commons.wikimedia.org/wiki/File:Transcription_Factors.svg) bajo licencia Creative Commons CC BY 3.0 (https://creativecommons.org/licenses/by/3.0/).

Una de las formas de estudio de dichas variaciones es según su ubicación. Si un SNP está ubicado en una zona codificante de un gen, según el tipo de cambio que genere puede resultar en una mutación **sinónima**, es decir, la que no modifica la proteína resultante, ya que algunos aminoácidos son codificados por más de un codón, o una mutación **no sinónima**: la que sustituye un aminoácido por otro. Un caso muy extremo de este tipo de mutación son los SNPs que causan una aparición prematura de un codón de finalización. Este efecto tan drástico en la estructura de la proteína está asociado a veces a enfermedades conocidas, como la fibrosis quística o numerosos tipos de cáncer [14]. Sin embargo, este tipo de variaciones no son las únicas que pueden afectar a la compleja maquinaria de la célula. De forma más indirecta, los SNPs también pueden afectar a la regulación de los genes, por ejemplo, si están ubicados en un TFBS y afectan a la afinidad de unión del factor de transcripción correspondiente a dicho TFBS, activando o desactivando así genes que de otra forma estarían en el estado contrario de activación. A este tipo de SNPs se los conoce como reguladores, y, contrario a lo que se pudiera pensar por el tipo de influencia indirecta que tienen en la regulación de los genes, alrededor de un tercio de los SNPs relacionados con enfermedades pertenecen a regiones no codificantes, tal y como corroboran recientes estudios GWAS[11] [15].

1.1.5. ARNs largos no codificantes

Como se ha mencionado anteriormente, el porcentaje de secuencia de ADN humano que se traduce en proteínas es muy reducido en comparación con el tamaño del genoma completo. Además, el tamaño del genoma humano es muy diferente al de otros organismos menos complejos, mientras que la diferencia en el número de genes codificantes con respecto a esas mismas especies no es tan elevada [16]. Esta diferencia parece indicar que el ADN intergénico es responsable en parte de la complejidad de nuestro organismo [17]. Por otro lado, el descubrimiento paulatino de nuevas entidades que participan en los procesos de regulación ha extendido el dogma central de la biología molecular expuesto

[11] Genome-Wide Association Studies

anteriormente. Entre estas entidades se encuentran *loci* capaces de transcribir ARN que no está destinado a finalmente convertirse en una proteína, por lo que se denomina ARN no codificante o ncRNA[12].

Estos transcritos de ARN no codificantes se dividen en dos categorías: los ncRNAs *pequeños*, es decir, aquellos cuya longitud es menor de 200 nucleótidos, y los ncRNAs *largos*, o lncRNAs[13], los cuales tienen una longitud mayor. En general, los primeros han conocido mayor estudio y subsecuente clasificación en los últimos años [18], siendo los lncRNAs de más reciente interés científico. Seguramente por esta causa, los ARN no codificantes de pequeño tamaño se subdividen en numerosas categorías, como los micro ARNs, de aproximadamente $\sim 22nt$ de longitud, el ARN ribosomal, el transfer ARN ya descrito en la sección anterior o los ARNs pequeños de interferencia. La funcionalidad de estos ARNs es diversa y va desde funciones generales como las desempeñadas por el transfer ARN en la traducción de ARNm a proteínas o los pequeños ARN nucleares, relacionados con el fenómeno de *splicing*, hasta funcionalidades como la destrucción del ARNm o la edición del ARN [19].

Por otra parte, el interés científico en los lncRNAs es bastante más reciente, pese a que han resultado casi triplicar al número de genes codificantes [20]. Estos transcritos forman parte de importantes procesos biológicos tales como la impronta genética, la regulación alostérica de la actividad enzimática o la conformación cromosómica [21]. Sin embargo, hasta años recientes han sido difíciles de monitorizar, debido entre otras cosas a que su volumen de transcripción es mucho más reducido [22, 23].

Recientemente han sido descubiertos miles de lncRNAs en el transcriptoma de los mamíferos [24]. Su papel regulador atrae desde entonces mucho interés debido no sólo al hecho ya mencionado de que un tercio de los SNPs relacionados con enfermedades conocidas a través de estudios GWAS están ubicadas en regiones no codificantes [15], sino en general a la idea cada vez más asentada en la comunidad científica de que el *ADN basura* es ADN de funcionalidad aún desconocida. Visto

[12] Non-coding RNA.
[13] Long non-coding RNAs.

desde la óptica del análisis de ARN, se acepta cada vez más la idea de que todo transcrito juega un papel en el funcionamiento de la célula, incluso si éste no sintetiza una proteína [25]. El estudio de estas nuevas entidades biológicas por tanto no hace sino reforzar esta nueva concepción del análisis del ADN. Se ha demostrado que los lncRNAs juegan un papel en los múltiples niveles de los caminos de expresión de los genes [26, 21]. Por ejemplo, actúan como silenciadores o activadores en la transcripción de los genes y además presentan una elevada cantidad de TFBSs [27]. Hasta ahora, sólo un número limitado de lncRNAs ha sido estudiado, ya que la investigación se ha centrado en los pocos lncRNAs con una asociación conocida a enfermedades.

De la misma manera que hace unos años se hiciera con los genes codificantes de proteínas, han surgido iniciativas con el fin de aunar el conocimiento existente sobre lncRNAs, para poner a la disposición de la comunidad científica bases de datos sobre este tipo de entidades, las cuales engloban ya más de 15000 genes de lncRNA:

- **GENCODE**[14] [28]. Es una iniciativa del consorcio ENCODE[15], cuya meta es identificar todos los elementos funcionales en la secuencia del genoma humano. GENCODE abarca no sólo lncRNAs, sino también genes codificantes, otros ncRNAs, pseudogenes y otras entidades funcionales. La última versión de GENCODE, GENCODEv25, cuenta con 15767 genes lncRNA, los cuales producen un total de 27692 transcritos.
- **lncRNAdb** [29]. Una base de datos de anotación de funcionalidad de lncRNAs cuyas entradas son obtenidas manualmente de la literatura biomédica, donde para cada lncRNA se proporciona información de secuencia, contexto genómico, datos de expresión, información estructural y conservación, entre otras.
- **lncRNAdisease** [30]. Una base de datos que contiene 480 relaciones entre lncRNAs y enfermedades validadas experimentalmente, las cuales incluyen

[14] http://www.gencodegenes.org
[15] Encyclopedia Of DNA Elements. Enciclopedia de elementos de ADN

166 enfermedades y unas 1500 predicciones computacionales noveles de asociaciones lncRNA-enfermedad.

- **NONCODE v4.0** [31]. NONCODE es una base de datos de anotación de ARNs no codificantes, con especial hincapié en los lncRNAs. En el caso de esta base de datos no sólo se anotan ncRNAs humanos, sino también de otras especies. NONCODE acaba de publicar una actualización [32].
- **LNCipedia** [33]. Una base de datos de anotación de transcritos de lncRNA humanos basados en su potencial codificador y en estudios de conservación de secuencia entre especies.
- **Lnc2Cancer** [34]. Una base de datos de asociaciones manualmente curadas de relaciones experimentales validadas entre lncRNAs y varios tipos de cáncer.

Ante la aparición de todas estas bases de datos de anotaciones de lncRNAs se abre un amplio campo de investigación, ya que gran cantidad de estudios realizados en genes codificantes son aplicables o adaptables al estudio de lncRNAs y todavía no han sido realizados. En este sentido, los datos correspondientes a lncRNAs disponibles en la actualidad se circunscriben casi exclusivamente a su ubicación en el genoma junto con características tales como potencial codificador o conservación entre especies. Otros estudios o anotaciones, tales como aquellos que incluyan elementos reguladores en los *loci* que transcriben lncRNAs, están aún por realizar.

1.1.6. Secuenciación de ADN

Para conocer la secuencia exacta de nucleótidos que conforman la información genética de un individuo es necesario realizar un proceso bioquímico denominado **secuenciación** del ADN. Dicho procedimiento tiene su origen en el año 1977, cuando Fred Sanger y Alan R. Coulson publicaron una metodología para la secuenciación de cadenas de ADN [35] que supuso un hito en la investigación en biomedicina, con la consecución del código genético completo de una bacteria. En 1990 surgió el Proyecto Genoma Humano, que permitió la secuenciación del

genoma completo de un humano [1], tras unos 14 años de trabajo y un coste aproximado de unos 2.700 millones de dólares.

Cabe destacar que esto fue un proceso de secuenciación y ensamblado conocido como ensamblado *de novo*: los *reads*, o fragmentos de ADN obtenidos mediante el procedimiento, habían de ser ensamblados como piezas de un puzzle sin conocimiento sobre la fotografía final. Una vez ensamblado, este genoma actúa como *genoma de referencia*, permitiendo sucesivos procesos de secuenciáción de la misma especie ser más rápidos, al poder alinear los *reads* contra dicho genoma de referencia. La figura 1.6 muestra ambos procesos. No obstante, un genoma de referencia no es algo inamovible sino susceptible de mejora mediante procesos sucesivos de *resecuenciación*, en los que secuencias obtenidas por tecnologías más avanzadas permiten mejorar la calidad del genoma de referencia.

Por otro lado, a finales de los años 80, los científicos Stephen Fodor, Michael Pirrung, Leighton Read y Lubert Stryer, desarrollaron una tecnología innovadora para la determinación y cuantificación del ADN de una muestra, tecnología que daría lugar a la primera plataforma de microarrays de expresión, la cual permite cuantificar los niveles de expresión de una serie de genes en un momento determinado.

Estas tecnologías sentaron las bases para los avances que hemos presenciado en la actualidad. En las siguientes secciones se detalla el estado del arte en las tecnologías de secuenciación de ADN así como tecnologías afines para la secuenciación de ADN que interactúa con proteínas concretas o la secuenciación del transcriptoma, es decir, el ARN que está siendo transcrito en un organismo en un momento determinado.

1.1.6.1. Tecnologías de secuenciación NGS y de tercera generación

En los últimos diez años la comunidad científica ha presenciado un avance enorme en las tecnologías de secuenciación. Desde la metodología Sanger, considerada como la tecnología de primera generación [36], pasando por las tecnologías de secuenciación de siguiente generación o NGS [37], hasta las tecnologías más recientes denominadas de tercera generación. Estas técnicas han

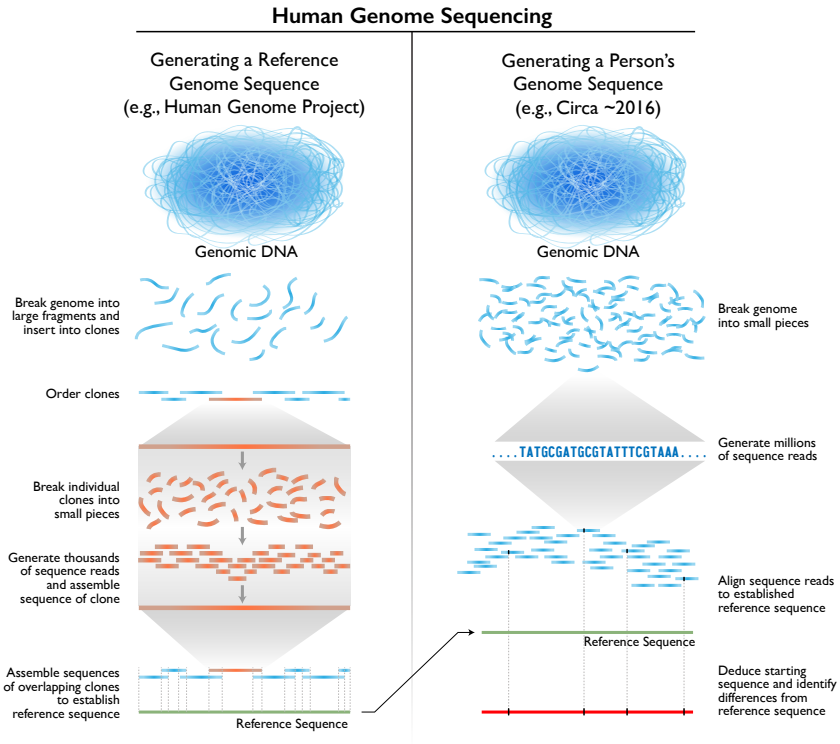


Figura 1.6: ***Secuenciación y ensamblado de novo y contra genoma de referencia.*** *La referencia obtenida en el proceso de novo se utiliza para acelerar el proceso en subsecuentes alineamientos contra genoma de referencia. Imagen disponible en:* `https://www.genome.gov/sequencingcosts/`. *Acceso el 20-12-2016.*

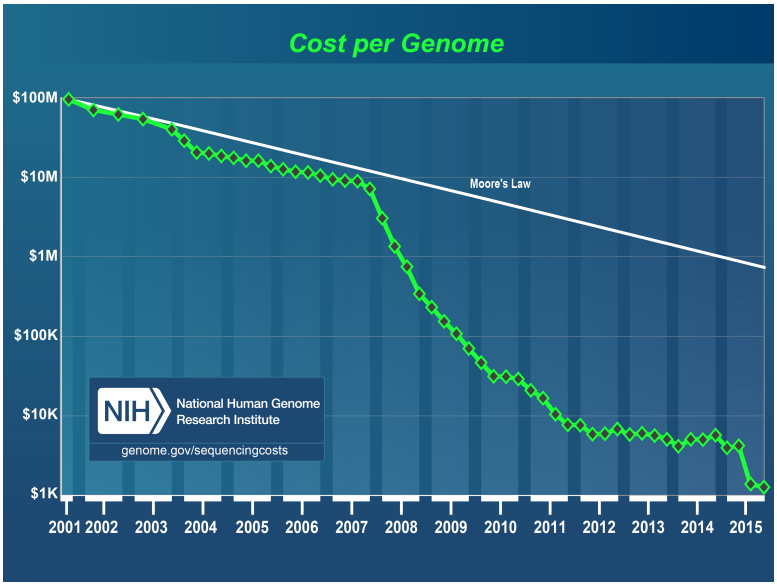


Figura 1.7: ***Descenso del coste de secuenciación de un genoma en los últimos años.*** *Disponible en:* `http://www.genome.gov/sequencingcostsdata`. *Acceso el 20-12-2016.*

permitido abaratar costes y mejorar la velocidad y calidad de secuenciación frente a los métodos basados en la metodología Sanger [38]. En la actualidad, el coste de secuenciación se ha reducido drásticamente a cerca de mil de dólares, como se puede observar en la figura 1.7.

El primer hito en la reciente carrera en las tecnologías de secuenciación lo supusieron las tecnologías NGS o de siguiente generación. Dichas tecnologías se basan en la generación de una cantidad muy grande de secuencias cortas o *reads* a un coste bajo. Estas son posteriormente alineadas contra un genoma de referencia mediante técnicas computacionales. En la actualidad existen multitud de plataformas para la secuenciación de siguiente generación. A continuación se enumeran las plataformas más extendidas para secuenciación agrupadas por compañía [39].

Illumina. La compañía Illumina es uno de los gigantes del mercado de la

secuenciación. Desde que en 2006 lanzó el Genome Analyzer II, han sucedido a esta plataforma multitud de mejoras. Actualmente cuenta con una serie de secuenciadores muy extendidos en todo el mundo: MiSeq, NextSeq y la serie HiSeq, que actualmente se encuentra en los modelos HiSeq 3000/4000. El HiSeq 2500, por ejemplo, es capaz de producir datos de secuenciación de genoma humano completo en 27 horas.

Life Technologies – ThermoFisher – Ion Torrent. Life Technologies lanzó en 2010 la tecnología Ion Torrent con la plataforma de secuenciación Ion PGM. Posteriormente, en 2012, lanzaron Ion Proton, la cual incrementa en un orden de magnitud la producción de Ion PGM, aunque proporciona reads más cortos.

Pacific Biosciences. Pacific Biosciences comercializó su tecnología SMRT[16], aplicada en la plataforma RS II en 2010, la cual continúa siendo la plataforma vigente en la actualidad. La tecnología SMRT permite obtener reads mucho más largos que otras tecnologías, de hasta 20000 nucleótidos de longitud, con una media de 3000 y tasas de error que oscilan entre el 11 % y el 14 % [40].

Oxford Nanopore Technologies. Oxford Nanopore ha prelanzado al mercado MinION [41], una herramienta que secuencia moléculas individuales en tiempo real [36] mediante una tecnología basada en nanoporos, siendo además de tamaño portable, como se puede observar en la figura 1.8. MinION proporciona reads de entre 5000 y 50000bp con tasas de error entre el 5 % y el 40 % [42].

Las propuestas de Pacific Biosciences y Oxford Nanopore se consideran actualmente como tecnologías de tercera generación. Entre sus características destacan la capacidad para secuenciar moléculas individuales en tiempo real [36]. La diferencia entre tasas de error y longitud de reads de estas tecnologías frente a las mencionadas anteriormente ha propiciado la aparición de multitud de metodologías nuevas de alineamiento y ensamblado capaces de sacar partido a sus características y manejar las tasas de error, más elevadas que en otras metodologías de reads cortos. Entre las ventajas de estos reads largos se incluyen la capacidad de ubicar reads en zonas de *repeats* [40]. Esta capacidad hace a este tipo de tecnologías muy útiles tanto para resecuenciar genomas, rellenando los huecos

[16] Single-Molecule Real-Time.

Figura 1.8: *Secuenciador de tercera generación MinION, de Oxford Nanopore.*

generados por los repeats por las tecnologías de reads más cortos, como para el ensamblado de genomas *de novo*. Además, en el caso de tecnologías como la de Oxford Nanopore, su portabilidad ha abierto el área de la secuenciación *in situ* para estudiar organismos en su lugar de aparición, sin necesitar el traslado del material genético a un laboratorio.

1.1.6.2. RNA-seq: Secuenciando el transcriptoma.

Al secuenciar el ARN que está siendo transcrito en lugar del ADN es posible observar qué secuencias de ADN están siendo transcritas en un momento dado y en qué cantidad, es decir, *cuantificar* la expresión del genoma en un momento dado. Tradicionalmente, la expresión de ARNm ha sido evaluada mediante enfoques basados en qPCR[17] o microarrays de expresión. La primera es más eficiente y económica para estudios a escala genómica [37]. Propuestas como SAGE[18] [43] se han centrado en el análisis computacional de secuenciado del ADN. Este tipo de enfoques ha estado limitado históricamente por el elevado coste de secuenciación

[17] PCR cuantitativa (quantitative PCR).
[18] Serial Analysis of Gene Expression.

de ADN. Sin embargo, el abaratamiento de costes asociado a las nuevas tecnologías de secuenciación ha permitido el progreso de este tipo de tecnologías, dando lugar a nuevas propuestas.

En este sentido, la tecnología RNA-Seq [44] es desde hace unos años un estándar en el estudio de expresión génica. RNA-seq es proceso de secuenciación del transcriptoma de una muestra de un individuo, es decir, las cadenas de ARNm de esta muestra celular del individuo en un momento concreto. En la figura 1.9 se puede observar el proceso. Para poder obtener el ARNm es preciso obtener previamente ADN complementario o ADNc. Este tipo de ADN se obtiene del ARNm transcrito en un organismo, por tanto contiene únicamente los genes expresados en él. El ADNc es después fragmentado y analizado para posteriormente secuenciarlo con alguna técnica de secuenciación.

1.1.6.3. ChIP-seq: Interacción ADN-proteína.

La interacción entre ADN y proteínas juega un papel clave en la regulación y expresión de los genes. Estas interacciones, sin embargo, no son reflejadas en las secuencias de ADN obtenidas mediante las técnicas explicadas en apartados anteriores. No obstante, existen técnicas experimentales que permiten estudiar estas interacciones. Tal es el caso de la inmunoprecipitación de cromatina o ChIP[19] [46]. ChIP conlleva una serie de pasos [37]:

1. El ADN y las proteínas asociadas se enlazan químicamente.
2. Se aíslan los núcleos, se someten a un proceso de lisis y el ADN es fragmentado.
3. Se utiliza un anticuerpo específico para la proteína de interés que se une al ADN (por ejemplo, un factor de transcripción) para inmunoprecipitar de forma selectiva los complejos proteína-ADN asociados.
4. Los enlaces químicos entre ADN y proteína son revertidos y el ADN analizado. En las aproximaciones iniciales, este análisis del ADN involucrado en la interacción proteína-ADN incluía el análisis del gen involucrado mediante

[19] Chromatine Immuno-Precipitation

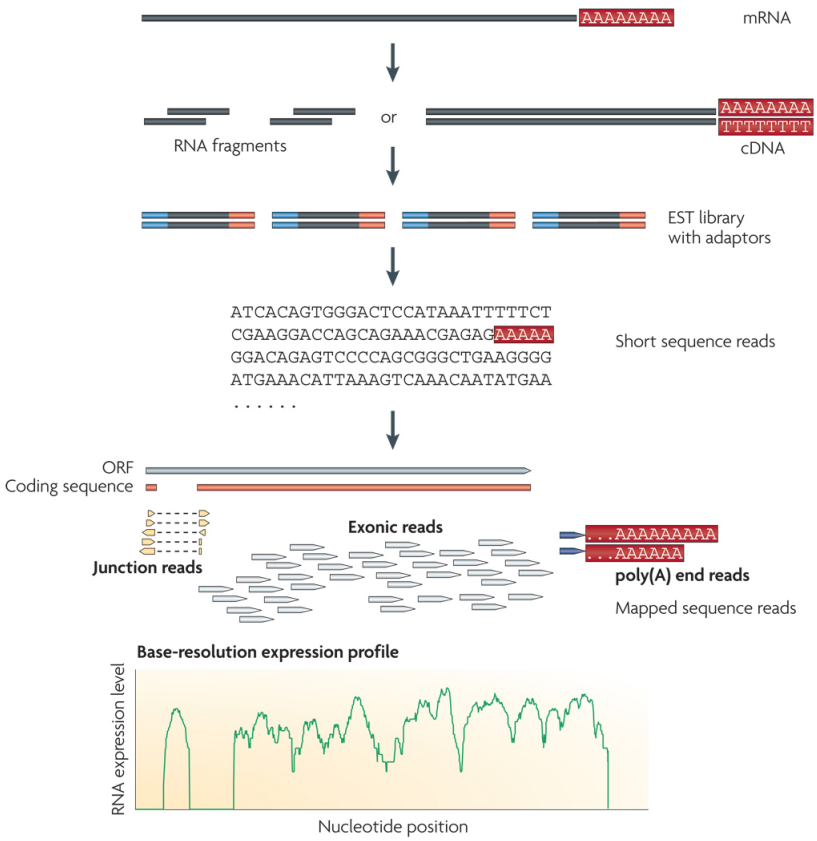


Figura 1.9: ***Experimento de RNA-seq****. Los ARNs se convierten en una librería de fragmentos de ADNc y una secuencia corta se obtiene de cada ADNc utilizando una tecnología de secuenciación de alto rendimiento. Los reads resultantes se alinean al genoma de referencia. La imagen ha sido tomada de [45].*

qPCR [47]. Más recientemente ha sido posible realizar estudios mediante microarrays. Este enfoque es conocido como ChIP-chip o ChIP-on-chip [48]. La aplicación de técnicas de secuenciación NGS o posteriores al ADN obtenido mediante ChIP da lugar a la técnica que hoy en día se conoce como ChIP-seq, cuyo origen se encuentra en Johnson et al. [49].

Las metodologías ChIP-seq permiten obtener grandes cantidades de secuencias que han interactuado con una proteína de interés, por lo que son de utilidad para descubrir TFBSs *de novo* a partir de un conjunto de secuencias que se saben que los contienen. Sin embargo, estas metodologías son específicas de proteína.

1.1.7. La Medicina Personalizada

La Medicina Personalizada o de Precisión promueve un uso del conocimiento sobre las características específicas de cada individuo para una mejora de su diagnóstico y tratamiento [50], la cual ha ido adquiriendo gran protagonismo en los últimos años [51].

La idea de utilizar características del paciente para adaptar su tratamiento no es en sí misma nueva, tomemos por ejemplo los tipos sanguíneos, los cuales llevan ya un siglo en uso para guiar las transfusiones de sangre [50]. Por otro lado, los médicos también llevan años personalizando los tratamientos de una forma *ad hoc*, por ejemplo, si prescriben medicamentos a pacientes basándose en ciertas características y las monitorizan para estimar el funcionamiento de los mismos sobre el paciente [52]. Sin embargo, la posibilidad de aplicar este concepto sistemáticamente no ha existido hasta los avances tecnológicos de los últimos años, con el advenimiento de las tecnologías de secuenciación de alto rendimiento y el consiguiente abaratamiento de los costes económicos y temporales que supone la obtención de datos del paciente. La caracterización del individuo a nivel molecular permite adaptar el tratamiento a las particularidades de cada paciente, a la par que se mejora la comprensión de la enfermedad y sus mecanismos [2].

En la actualidad, los medicamentos aún son únicos y se aplican a multitud de pacientes de naturaleza diversa, obteniendo asimismo diversos resultados, donde algunos colectivos, probablemente los más numerosos, son beneficiados

en detrimento de las individualidades no representadas. La meta de la Medicina Personalizada, por contra, es el tratamiento de cada paciente como un caso individual, en el que sus características genómicas, epigenéticas, de entorno, estilo de vida e historia clínica sean tenidas en cuenta [51].

Sin embargo, pese al gran avance del que ha sido sujeto la Medicina Personalizada en los últimos años, aún no es una realidad clínica, especialmente en las enfermedades de baja incidencia. Se hace pues necesario el descubrimiento de biomarcadores efectivos y asequibles para la expansión del área de aplicación de la Medicina Personalizada. Una muestra de este creciente interés son sus apariciones como área de interés en la iniciativa europea Horizonte 2020 [51] y en la iniciativa estadounidense PMI[20] presentada por Obama en enero de 2015 [50].

En la iniciativa Horizonte 2020 se ha definido la Medicina Personalizada como *"un modelo médico que utiliza la caracterización del fenotipo y el genotipo de los individuos para diseñar la estrategia terapéutica adecuada para la persona adecuada en el momento adecuada, y/o determinar la predisposición a enfermedad y/o proporcionar prevención temprana y dirigida"*[21].

Con respecto a la iniciativa PMI presentada por Obama, su meta principal es *"dar paso a una nueva era en la medicina a través de investigación, tecnología y políticas que capaciten a los pacientes, investigadores y personal médico para trabajar juntos hacia el tratamiento personalizado"*[22].

Para ello es necesario poder trasladar los tratamientos a las características únicas de cada paciente, desde su genoma individual, su microbioma (el conjunto de microorganismos que habitan en un individuo) hasta su historial clínico y estilo de vida. Con esta finalidad se proponen una serie de líneas principales de trabajo[23]:

[20] Precision Medicine Initiative

[21] http://ec.europa.eu/research/health/index.cfm?pg=policy&policyname=personalised – acceso el 08-12-2016.

[22] Información obtenida de https://www.whitehouse.gov/precision-medicine – acceso el 08-12-2016.

[23] https://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative – acceso el 08-12-2016.

1. **Mejorar los tratamientos para el cáncer**. Acelerar el diseño y prueba de métodos *a medida* efectivos mediante la expansión de ensayos clínicos en cáncer basados en genética.
2. **Creación de grupo nacional de voluntarios para investigación**. Se espera reunir a un millón o más estadounidenses para participar en la investigación de Medicina Personalizada con sus datos individuales.
3. **Compromiso para la protección de la privacidad**. Se pretende asegurar la protección de la privacidad de los pacientes de forma rigurosa. Además, identificar posibles problemas éticos y de privacidad en la aplicación de la Medicina Personalizada.
4. **Modernización reguladora**. La iniciativa incluye revisitar las instituciones reguladoras para observar si es necesario algún cambio para el correcto funcionamiento del modelo.
5. **Iniciativas público-privadas**. Con el fin de promover el desarrollo de la infraestructura necesaria, se pretende promover iniciativas de investigación público-privadas en el marco de la Medicina Personalizada.

La aparición de la Medicina Personalizada como un punto clave en los planes globales de actuación sobre la salud pública de la Unión Europea y Estados Unidos es, por tanto, señal inequívoca de la importancia que está tomando como área de interés científico y clínico, objetivos con los que se alinean las líneas de trabajo descritas en esta tesis.

1.2. Integración de fuentes de información heterogénea

Tal vez uno de los hechos más patentes en la biología computacional de los últimos años sea el continuo crecimiento de una inmensa cantidad de datos disponibles para el análisis y la extracción de conocimiento de los mismos. La mejora de la calidad de los resultados y el abaratamiento de los costes de las tecnologías experimentales han resultado en una cantidad de datos heterogéneos en continuo crecimiento. La integración de estas fuentes de datos constituye un reto cuya solución resultará en un conocimiento más completo de los procesos biológicos y sus actores.

Por otro lado, análogamente a la heterogeneidad y diversidad de las fuentes de datos disponibles, existen multitud de agentes en el gran abanico de procesos biológicos que tienen lugar en la compleja maquinaria celular. Esta multitud de entidades diferentes interactúan entre sí de una forma parecida a la de los nodos en una red. Así, el modelo de grafo o red se postula como una herramienta muy potente para la inferencia de hipótesis sobre el funcionamiento de los mencionados procesos [53]. Las redes o grafos modelados a partir de los sistemas biológicos no son aleatorias, sino que siguen unos principios de organización en sus estructuras que las diferencian de redes aleatorias. Es por ello que la aplicación de metodologías de minería de datos en grafos permite la extracción de conocimiento de dichas redes [54].

En este apartado se detallan los conceptos necesarios sobre nomenclatura, teoría y propiedades de los grafos necesarios para comprender los enfoques de análisis de redes en el ámbito de la biología computacional. A continuación se enumeran una serie de tipos de redes complejas de reciente estudio en relación con las propiedades esperadas en las redes biológicas. Por último, se describen los enfoques y retos existentes en cuanto al análisis de redes heterogéneas.

1.2.1. Definiciones básicas sobre grafos

A continuación se enumeran algunas definiciones básicas relativas a los grafos y a su representación como matrices de adyacencia en los términos en que serán

utilizados a lo largo del presente capítulo.

Definición 1. ***Grafo no dirigido***. *Se define un grafo no dirigido como* $G = (V, E)$ *donde* $V = \{v_1, ..., v_n\}$ *es un conjunto no vacío y finito de vértices y* E *es el conjunto de arcos o aristas, tal que* $E = \{(v_i, v_j) : i \neq j \forall i, j \in 1, ..., n\}$[24].

Definición 2. ***Grafo completo***. *Se dice que un grafo* $G = (V, E)$ *es completo si existen aristas conectando todos los vértices:* $\forall v_i, v_j \in V$, $\exists (v_i, v_j) \in E$. *El grafo completo de* N *vértices se denomina* $\mathcal{K}_N$.

Definición 3. ***Grafo vacío***. *Un grafo* G *se dice vacío si no tiene aristas, es decir* $E = \emptyset$.

Además, existe un tipo de grafo en el que sus vértices están *particionados* en dos subconjuntos disjuntos. Se define a continuación por su relevancia en las secciones posteriores.

Definición 4. ***Grafo bipartito***. *Un grafo bipartito es un grafo* $G = (V, E)$ *cuyos vértices* V *se pueden particionar en dos conjuntos* V_1 *y* V_2 *tales que no existen aristas que conecten dos nodos de* V_1 *ni dos nodos de* V_2*, es decir:* $V_1 \cup V_2 = V$, $V_1 \cap V_2 = \emptyset$, *y* $(v, w) \in E \implies v \in V_1 \wedge w \in V_2$.

La figura 1.10 muestra un grafo bipartito que relaciona enfermedades con genes. Las únicas relaciones presentes en dicho grafo son las que relacionan una enfermedad con un gen, no habiendo relaciones ni enfermedad-enfermedad ni gen-gen.

Las aristas E de un grafo no dirigido G inducen una relación binaria simétrica:

Definición 5. ***Relación de adyacencia***. *Dado un grafo no dirigido* $G = (E, V)$, *las aristas de* E *inducen en* V *una relación binaria simétrica denominada relación de adyacencia. Para cada arco* $e = (v_i, v_j)$ *en* E, *los nodos* v_i, v_j *de dicen adyacentes, es decir* $v_i \sim v_j$. *Además, se dice que* v_i *y* v_j *son incidentes a* e.

De la relación de adyacencia se introduce el concepto de grado de un nodo.

[24] Nótese que esta definición de grafo no permite *bucles*, es decir, aristas que empiezan y terminan en el mismo nodo.

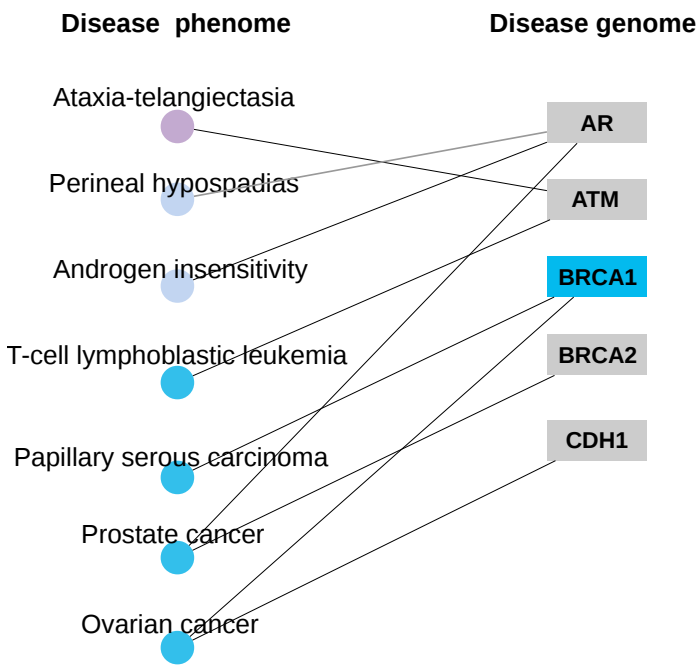


Figura 1.10: *Ejemplo de red biológica bipartita: red bipartita enfermedades-genes (fragmento de [55]).*

Definición 6. ***Grado de un nodo.*** *Se define también el grado de un vértice $v \in V$ $d_G(v)$ como el número de aristas a las que v es incidente, o el número de nodos a los que v es adyacente:*

$$d_G(v) = |\{e = (v, w)\}|, \forall w \in V, \forall e \in E$$

Definición 7. ***Vecindario de un vértice.*** *Dado un grafo no dirigido $G = (V, E)$, se define el vecindario de un nodo $v \in V$, $\mathcal{N}(v)$ como el conjunto de vértices adyacentes a v:*

$$\mathcal{N}(v) = \{u : (v, u) \in E\}$$

Nótese que $|\mathcal{N}(v)| = d_G(v)$.

En esta sección se omiten las consideraciones sobre grafos dirigidos, ya que en las herramientas que se describirán en subsecuentes secciones el concepto de dirección no se utiliza. En general, los vínculos entre nodos vendrán dados por relaciones de similitud, las cuales cumplen propiedades de simetría. Hasta ahora se han descrito los grafos donde las aristas o arcos son una propiedad binaria. A continuación se define el concepto de **grafo ponderado**.

Definición 8. ***Grafo ponderado***. *Se dice que un grafo es ponderado*[25] *si existe un* ***peso*** *asociado a cada arista, es decir, G se puede definir como* $G = (V, E, w)$ *donde* w *es una aplicación* $w : E \rightarrow \mathcal{R}^+$ *que asigna a cada arista* $e = (v, w) \in E$ *un valor* $w(e) > 0$.

Dados dos nodos u y v en los vértices de un grafo, una de las preguntas que se suelen formular es si dichos vértices están o no conectados y a qué distancia. En este sentido, se define el concepto de camino:

Definición 9. ***Camino***. *Dado un grafo no dirigido* $G = (V, E)$ *y dos vértices* $u, v \in V$, *un camino es una secuencia de vértices en* V $P_{u,v} = \{p_1, ..., p_n\}$ *donde* $p_i \in V$ *y* $p_1 = u, p_n = v$ *y existe una arista conectando cada par de nodos consecutivos:* $\forall p_i, p_{i+1} \in P_{u,v}$, $\exists (p_i, p_{i+1}) \in E$. *Se dice que un camino es un* ***camino simple*** *si no existe repetición en los vértices en* $P_{u,v}$ *salvo la posibilidad de que* $u = v$.

Definición 10. ***Ciclo***. *Un ciclo es un camino cerrado.*

Definición 11. ***Longitud de un camino***. *Dado un grafo ponderado no dirigido* $G = (V, E, w)$, *un camino en* G $P_{u,v}$ *entre los vértices* $u, v \in V$, *se define* $d(P_{u,v})$ *como:*

$$d(P) = \sum_{k=2}^{K} w((v_{k-1}, v_k))$$

[25] Por simplicidad, en este texto se utilizan los términos *red* o *grafo* de forma indistinta para el concepto de *grafo ponderado*, ya que en las aplicaciones biológicas que se describen todos los grafos son ponderados. Algunos autores denominan red a un grafo ponderado [56]

Definición 12. ***Camino más corto (camino geodésico)***. *Se define el camino más corto en $G = (V, E)$ entre los vértices $u, v \in V$ como*

$$d_{u,v} = min_{\mathcal{P}_{u\to v}} d(\mathcal{P}_{u\to v})$$

donde $\mathcal{P}_{u\to v}$ es el conjunto de todos los caminos posibles entre u y v en G.

Definición 13. ***Distancia entre dos vértices***. *Dado un grafo ponderado no dirigido $G = (V, E, w)$, la distancia entre dos vértices $u, v \in V$ $d_{u,v}$ será siempre la longitud del camino más corto entre ambos. Nótese que el camino más corto siempre es un camino simple.*

Representación de los grafos

Normalmente, un grafo, ponderado o no, se suele representar mediante una **matriz de adyacencia**. A continuación se define formalmente este tipo de representación.

Definición 14. ***Matriz de adyacencia***. *Sea $G = (V, E, w)$ un grafo ponderado no dirigido. Su matriz de adyacencia $\mathcal{A}$ se define como sigue:*

- *Las dimensiones de la matriz vienen dadas por el número de vértices: $\mathcal{A}_{|V|\times|V|}$*
- *El elemento en la fila i, columna j de $\mathcal{A}$, denominado $\mathcal{A}_{i,j}$ representa el valor de la aplicacion w para la arista (v_i, v_j), es decir: $\forall(v_i, v_j) \in E, \mathcal{A}_{i,j} \neq 0$ y $\forall(v_i, v_j) \notin E, \mathcal{A}_{i,j} = 0$*

Nótese que si G es un grafo no ponderado, entonces la matriz $\mathcal{A}$ es una matriz binaria, y si G es un grafo no dirigido, la matriz $\mathcal{A}$ es una matriz simétrica.

1.2.2. Tipos de grafos o redes complejas

Existen algunos tipos de grafos o redes, los cuales cumplen ciertas propiedades topológicas o estructurales de interés. Dentro de estos, se pueden enumerar los siguientes [57]: redes aleatorias, redes aleatorias agrupadas, redes de mundo pequeño, redes libres de escala y redes centro-periferia. A continuación se describen brevemente las características de cada una de ellas.

1.2.2.1. Redes aleatorias

Dados los vértices de V, se establecen las aristas de E de forma aleatoria donde para cada par de vértices existe una probabilidad $p > 0$ de unir dichos vértices. Esta generación de redes aleatorias se conoce como el modelo de Erdös y Réyni [58]. La característica principal de este tipo de redes es que la distribución de los grados de los nodos sigue una distribución Binomial de parámetros $|V| - 1, p$.

1.2.2.2. Redes aleatorias agrupadas

Muchas redes reales, tales como las redes sociales, presentan unos módulos denominados comunidades [59]. La característica principal de estos módulos es análoga a la definición común de comunidad: los nodos de una comunidad presentan un alto nivel de conectividad con los nodos dentro de su comunidad, mientras que la conectividad con nodos externos a la comunidad es mucho más reducida. La forma de modelar este tipo de red consiste básicamente en dividir la probabilidad p de conectar dos vértices en dos probabilidades: p_{in} si se trata de vértices pertenecientes a la misma comunidad, p_{out} si se trata de vértices de diferentes comunidades. En [59] se presenta un modelo aglomerativo para construir este tipo de redes, partiendo de un conjunto de V vértices que se aglutina en M comunidades.

1.2.2.3. Redes de mundo pequeño o redes sociales

Se podría pensar que existe una correlación directa entre el tamaño de una red y la distancia máxima posible entre sus nodos. Sin embargo, muchas redes reales, como por ejemplo las redes sociales, presentan la característica de que la mayoría de sus vértices se pueden conectar entre sí con caminos de longitud corta [60].

Las redes sociales presentan dos características de interés [61]:

1. **Ley potencial de densificación**. En el pasado se creía que conforme evoluciona una red, el número de los grados de sus nodos crecía linealmente con respecto al número de nodos, lo que se conocía como la *asunción del grado medio constante*. Numerosos experimentos han demostrado que

las redes se densifican con el tiempo, es decir, el número de aristas crece exponencialmente con respecto al número de nodos:

$$e(t) \propto n(t)^a$$

Donde $e(t)$ y $n(t)$ representan el número de aristas y de nodos en el instante t, y el exponente a generalmente se encuentra en algún punto entre 1 y 2.

2. **Diámetro decreciente**. Se ha demostrado que el diámetro (la longitud de la distancia más larga existente entre sus nodos) de una red tiende a *disminuir* conforme aumenta el número de nodos de la red.

1.2.2.4. Redes libres de escala

Barabási y Albert [62] descubrieron en un estudio que algunas redes tienen un número reducido de nodos con un grado muy elevado mientras que la mayoría de los nodos tienen un grado muy bajo. En su definición de redes libres de escala, la distribución del grado de los nodos de la red sigue una ley potencial:

$$P(k) \sim k^{-\gamma}$$

donde γ es un exponente de escala. Esta propiedad hace que cuando el grado k crece, el número de vértices con grado k disminuya y viceversa, para k pequeño existirá una gran cantidad de vértices con grado k (para un valor fijo de γ).

La topología de estas redes la hace robusta al fallo si este fallo es aleatorio, es decir, si se elimina un nodo aleatorio de la red, es poco probable que sea uno de los nodos de alta conectividad. Incluso en caso de serlo, la red no pierde su conectividad, dado que seguramente existan otros nodos de alta conectividad. Sin embargo, si se atacan a la vez varios nodos de alta conectividad, la red se convierte en una multitud de nodos aislados. Es por ello que se suele decir que las redes libres de escala son robustas a fallos fortuitos, pero frágiles ante ataques intencionados.

1.2.2.5. Redes núcleo-periferia

La teoría de redes, por lo general, intenta descubrir estructuras a escalas locales, globales e intermedias en las redes. Mientras que se han hecho grandes progresos en lo que a detección de comunidades se refiere [63], existen aún estructuras que no están tan estudiadas. Las redes núcleo periferia se modelan como redes en las que existe un núcleo de alta conectividad y un conjunto de nodos que se consideran periferia. Los vértices del núcleo no sólo están bien relacionados entre sí, sino que también pueden presentar alta conectividad con nodos de la periferia. Sin embargo, los nodos de la periferia no están bien conectados ni entre sí ni con el núcleo. Por tanto, un núcleo en una red no es sólo una comunidad bien interconectada, sino que también posee cierta *centralidad*.

1.2.3. Propiedades de las redes biológicas

Las redes biológicas forman grafos con ciertas propiedades que se derivan de las características de los datos desde los que fueron construidas. En muchos casos las propiedades de estas redes derivan de su pertenencia parcial o total a las categorías mencionadas en la sección anterior. Se puede decir, pues, que las redes biológicas pueden presentar cierto nivel de agrupamiento en comunidades, características de red social o de redes libres de escala. Más específicamente, existen ciertas propiedades que se consideran ciertas para este tipo de redes o grafos [54]:

1. **Módulos**. Las redes biológicas muestran un cierto grado de *clustering*, es decir, grupos de nodos o comunidades que conservan una alta interconectividad en comparación con el nivel de conectividad exterior.
2. **Motivos.** En las redes biológicas existen subredes que se repiten más de lo esperado. A estas subredes se las suele llamar motivos y suelen estar relacionadas con alguna función biológica.
3. **Distribución de grados y *hubs*.** En una red aleatoria la mayoría de los nodos tienen por lo general un grado similar, y la distribución del grado de sus nodos sigue la distribución de Poisson. Sin embargo, en muchas redes biológicas hay un conjunto reducido de nodos de alto grado conocidos como

hubs que mantienen la red conectada. La figura 1.11 muestra una red de enfermedades, donde se puede observar este fenómeno.

4. **Propiedades de red de mundo pequeño.** En la mayoría de las redes existe este fenómeno ya comentado en las redes sociales: entre dos nodos cualesquiera existe un camino relativamente corto que los conecta. Esto quiere decir que cualquier nodo puede afectar no sólo a sus vecinos inmediatos sino a la totalidad de la red.
5. **Redundancia de caminos.** En los procesos biológicos se espera una cierta robustez. Una propiedad que refleja esto es la redundancia de caminos en las redes biológicas que los representa. Ésta característica de robustez se relaciona también con la existencia de esos *hubs* que le dan conectividad a la red. Además de existir caminos cortos entre muchos de los nodos, también suelen existir varios caminos que los conectan [64].

1.2.4. Medidas en redes complejas

Los grafos son una estructura de gran complejidad, y en particular las redes complejas que se han mencionado, poseen una gran cantidad de características estructurales, locales e intermedias. Las características que se han mencionado anteriormente de una forma más bien cualitativa tienen su contrapartida en medidas cuantitativas. En esta sección se hace un resumen de algunas de las medidas cuantitativas más frecuentes en el estudio de las propiedades de los grafos, divididas en varias categorías: las relacionadas con el grado de los nodos, las que se refieren a distancias entre nodos y conectividad, y las que se refieren a la estructura global de la red [57].

1.2.4.1. Medidas relativas al grado de los nodos

Densidad

La densidad D de una red mide cómo de fuertes son las conexiones entre los nodos de dicha red. Es la proporción de las conexiones existentes entre las conexiones posibles.

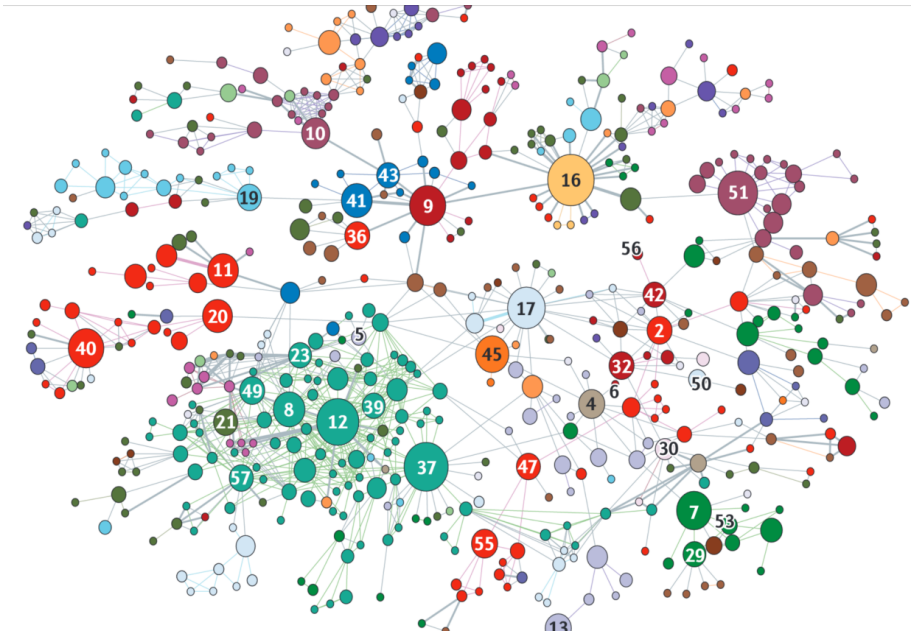


Figura 1.11: ***Ejemplo de red biológica*** *[54]. El tamaño de los nodos es proporcional a su grado. Se observa como unos pocos nodos (hubs) garantizan la conectividad de todo el conjunto. Los números se corresponden con enfermedades, siendo los hubs de mayor grado el 16 (sordera), 51 (retinitis pigmentosa), 37 (leucemia), 12 (cáncer de colon), 9 (cardiomiopatía).*

Definición 15. ***Densidad****. Sea $G = (V, E)$ un grafo simple no dirigido. Se define su densidad como:*

$$D = \frac{E}{\binom{V}{2}} = \frac{2E}{2V(V-1)}$$

D toma valores en $[0, 1]$. Cuando $D = 0$, se trata de un grafo vacío, mientras que si $D = 1$ se trata de un grafo completo.

En general, se habla de redes *dispersas* cuando D toma valores cercanos a 0, y *densas* en caso de que D se aproxime a 1. La densidad de las redes tiene implicaciones en gran parte de los algoritmos de aprendizaje automático aplicados a grafos.

Agrupabilidad

La agrupabilidad[26] de la red captura la preferencia de sus nodos de vincularse a otros nodos que tienen un grado similar o diferente [65]. A veces se entiende la agrupabilidad como la correlación de grados entre los vértices. El coeficiente de agrupabilidad r es esencialmente el coeficiente de correlación de Pearson del grado entre pares de vértices unidos. Por tanto, valores positivos de r indican relaciones entre vértices de similar grado, mientras que valores negativos de r indican relaciones entre vértices de grado diferente [66].

Se han hecho muchos estudios sobre la agrupabilidad de redes reales. En general, las redes sociales parecen ser agrupables, mientras que las redes tecnológicas, biológicas o financieras tienden a ser desagrupables [66].

Coeficiente "rich-club"

El coeficiente "rich-club" aparece por primera vez en [67]. Más recientemente, se ha parametrizado en base a un umbral de corte de grado k. El coeficiente "rich-club" mide la propiedad estructural de las redes complejas denominado fenómeno "rich-club", la cual se refiere a la tendencia de vértices con grado elevado a estar fuertemente conectados entre sí, formando estructuras cercanas a cliques (subgrafos completos). En general, los nodos con un elevado número de aristas, denominados *hubs* o nodos ricos, tienen más tendencia a formar estructuras altamente interconectadas (clubes) que los vértices de grado bajo.

Definición 16. ***Coeficiente "rich-club" no normalizado.*** *Sea $G = (V, E)$ un grafo simple no dirigido. Considerando $E_{>k}$ es el número de aristas entre los $N_{>k}$ vértices que tienen un grado mayor que un umbral $k \geq 0$, la versión escalada del coeficiente "rich-club" se define como [68]:*

$$\phi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k} - 1)}$$

[26] En inglés, *network assortativity*.

Una crítica a este coeficiente es que efectivamente puede ser alto incluso en redes aleatorias, ya que los nodos con grado alto tienen más probabilidad de hecho a estar conectados a cualquier nodo, y por ende a nodos con alto grado.

Definición 17. ***Coeficiente "rich-club" normalizado***. *Se define el coeficiente "rich-club" normalizado como [68]:*

$$\phi_{norm}(k) = \frac{\phi(k)}{\phi_{rand}(k)}$$

donde $\phi_{rand}(k)$ es el coeficiente "rich-club" no normalizado evaluado en una red randomizada con la misma distribución de grados P(k) de la red bajo estudio. Para esta métrica, si para ciertos valores de k tenemos $\phi_{norm}(k) > 1$, esto indica la existencia de un efecto "rich-club".

Nótese que las redes en las que el coeficiente de agrupabilidad se acerca a -1 (es decir, redes *desagrupables*) y que tienen regiones "rich-club" de grado alto apuntan a la existencia de una estructura núcleo-periferia.

1.2.4.2. Medidas relativas a la distancia entre los nodos

Definición 18. ***Diámetro***. *Se define el diámetro de una red como la longitud de la distancia más larga existente entre sus nodos:*

$$T = max_{u,v \in V} d_{uv}$$

Para un grafo no dirigido (y no ponderado) los valores de T están en [0, V-1].

Definición 19. ***Excentricidad de vértices***. *La excentricidad de un nodo $v \in V$ se define como la distancia máxima de v a cualquier otro nodo $u \in V \setminus \{v\}$:*

$$e_v = max_{u \in V \setminus \{v\}} d_{uv}$$

Definición 20. ***Radio***. *El radio ζ de una red es la excentricidad mínima:*

$$\zeta = min_{u \in V} e_u$$

Definición 21. ***Índice Wiener***. *El índice Wiener λ se define como la suma de las distancias geodésicas entre cada par de nodos del grafo:*

$$\lambda = \frac{1}{2} \sum_{u,v \in V, u \neq v} d_{uv}$$

Definición 22. ***Eficiencia global***. *La eficiencia global considera que la capacidad para enviar información entre dos nodos u y v es inversamente proporcional a la distancia geodésica:*

$$GE = \frac{1}{V(V-1)} \sum_{u,v \in V, u \neq v} \frac{1}{d_{uv}}$$

Definición 23. ***Armonía global***. *La armonía global h se define como la medida recíproca a la eficiencia:*

$$h = \frac{1}{GE}$$

1.2.4.3. Medidas estructurales

Definición 24. ***Coeficiente de clustering de un nodo*** *[69]. Es una medida que define cómo de cerca está el conjunto de vecinos de un nodo de formar un clique (un subgrafo completo). Se define como:*

$$CC_i = \frac{2|e_i|}{k_i(k_i-1)}$$

donde e_i es el número de aristas compartidas por los vecinos directos del vertice i y k_i es el grado del vértice i.

El coeficiente de clustering de una red se puede calcular como la media del coeficiente de clustering de sus nodos.

Definición 25. ***Coeficiente cíclico de un nodo*** *[70]. Se define el coeficiente cíclico θ_i del vértice i como la media del tamaño inverso del ciclo más pequeño que conecta ese vértice con dos de sus vecinos:*

$$\theta_i = \frac{2}{k_i(k_i-1)} \sum_{j,k \in \mathcal{N}(i)} \frac{1}{S^i_{jk}}$$

Donde S^i_{jk} es el tamaño del camino cerrado más pequeño que pasa por i conectándolo con los vecinos j y k. La suma itera sobre todos los pares de vecinos de i. $S^i_{jk} = \infty$ si no existe ningún camino entre j y k que no pase por i.

El coeficiente cíclico global se puede calcular como la media del coeficiente cíclico de sus nodos.

Cabe destacar otras medidas estructurales como la **modularidad** [71, 72]. La modularidad trata de encapsular la organización en comunidades de una red. Es un valor entre 0 y 1. Cuando la modularidad está cercana a 0 quiere decir que la red no presenta estructura en comunidades. Conforme crece la modularidad van apareciendo comunidades más definidas. La idea principal de la modularidad es calcular la proporción de aristas que están contenidas en comunidades en proporción a las esperadas si las aristas fueran dispuestas aleatoriamente.

Por otro lado existe la medida de **solapamiento topológico**, que pretende medir hasta qué punto dos vértices están conectados aproximadamente al mismo conjunto de nodos. Dicho de otro modo, esta medida compara cómo de parecidos son los vecindarios directos de ambos nodos.

1.2.4.4. Medidas de centralidad

Las medidas de centralidad calculan cómo de centrales o importantes son un nodo o una arista en una red. En este sentido, la medida más sencilla de centralidad que se puede imaginar es el grado de un nodo. Es lógico pensar que un nodo de grado elevado es relevante en una red. Además de esta medida, se han propuesto numerosas medidas de centralidad en la literatura.

Por un lado podemos encontrar las medidas **basadas en distancia**, las cuales a su vez se clasifican según el criterio usado para calcular la distancia de centralidad [73]:

- **Criterio Minimax**. Este criterio, análogamente al criterio que se podría desear seguir para ubicar un hospital, busca minimizar la máxima distancia de un nodo con respecto a cualquier otro nodo de la red.
- **Criterio Minisum**. Este criterio es análogo al que se usaría para ubicar un centro comercial. La idea es minimizar la distancia total en lugar de la máxima.

Definición 26. ***Criterio Minimax***. *Sea $u \in V$ un nodo cualquiera de la red, se denota la distancia máxima de u a otro nodo $v \in V$ cualquiera como la excentricidad e_u de u. Por tanto, la centralidad de u basada en su excentricidad:*

$$c_E(u) = \frac{1}{e_u} = \frac{1}{max_{v \in V} d_{uv}}$$

Definición 27. ***Criterio Minisum***. *Sea $u \in V$ un nodo cualquiera de la red, se denota la centralidad de u basada en la suma total de las distancias al resto de nodos de la red como:*

$$c_C(u) = \frac{1}{\sum_{v \in V} d_{uv}}$$

En el análisis de redes sociales, este concepto se llama ***cercanía***.

Existen otras medidas de centralidad que no consideran las distancias de nodo a nodo, sino que consideran el **flujo** que atraviesa un vértice, es decir, un nodo será relevante desde este punto de vista si muchos de los caminos más cortos deben pasar por dicho nodo.

Intermediación

La intermediación[27] mide en qué grado un nodo se encuentra en los caminos más cortos entre todos los nodos de la red [74]. Suponiendo que entre todos los nodos de la red se intercambian mensajes con igual probabilidad, al paso de un tiempo t determinado, el número de mensajes que hayan pasado por cada nodo

[27] En inglés *betweenness*.

será directamente proporcional al número de caminos geodésicos en los que se encuentra, suponiendo que los mensajes siempre se envían a través de los caminos mínimos [74]. Este número de caminos mínimos es lo que se conoce como índice de intermediación.

Los vértices con índice de intermediación elevado son los que causan más problemas en las comunicaciones si son eliminados.

Comunicabilidad

En las redes reales no siempre los trayectos que se siguen para comunicar dos nodos son los caminos geodésicos. La comunicabilidad [75] se define para pares de vértices, y refleja en términos generales cómo de fácil es la comunicación entre ambos vértices, mediante una combinación de caminos geodésicos y *paseos aleatorios* de distintas longitudes.

Existen multitud de medidas adicionales basadas en otros conceptos como la vitalidad [73] o el *feedback general*, entre las que se encuentran la centralidad de valores propios [76], el índice de Katz [77] o la conocida centralidad web PageRank [76]. Todas estas medidas buscan modelar distintas características que le dan relevancia a un nodo dentro de una red.

1.2.5. Procesos dinámicos en redes complejas

Una de las preguntas más frecuentes que se suele formular en el análisis de datos en forma de red es el nivel de asociación entre dos nodos, es decir, dadas dos entidades, cómo de relacionadas están considerando la estructura de la red.

En el caso de las redes biológicas, el interés de la pregunta suele estar relacionada con la red completa, puesto que se estudia en muchas ocasiones un sistema completo y su funcionamiento. La pregunta podría reformularse como: dados dos nodos i y j, cómo le afecta a i el estado de j y viceversa.

Se podría responder a esta pregunta mediante medidas como la distancia, la comunicabilidad u otras medidas complejas basadas en la estructura de la red. Sin embargo, existen también medidas que se basan en el lanzamiento de procesos

dinámicos en la red y el estudio de su comportamiento. Uno de los algoritmos más conocidos en este enfoque es el Random Walk with Restart (RWR) [78].

El algoritmo Random Walk with Restart se define de la siguiente forma. Se considera a una partícula aleatoria que se encuentra inicialmente en el nodo i. Iterativamente, la partícula pasa a uno de los nodos adyacentes al nodo en el que se encuentra, donde la probabilidad de pasar a dichos nodos es proporcional a los pesos de los arcos que unen el nodo actual con los nodos del siguiente paso. Además, existe una probabilidad c de que la partícula vuelva al nodo inicial i. El *score* de relevancia del nodo i con respecto al nodo j es igual a la probabilidad final $r_{i,j}$ de que la partícula permanezca finalmente en el nodo j:

$$\vec{r}_i = c\tilde{W}\vec{r}_i + (1-c)\vec{e}_i$$

La ecuación anterior define un sistema lineal donde $\vec{r}_i$ está determinado por:

$$\vec{r}_i = (1-c)(1-c\tilde{W})^{-1}\vec{e}_i$$

Donde $\tilde{W}$ es el grafo ponderado normalizado asociado a W, $\vec{e}_i$ es el vector de entrada inicial, donde el elemento i-ésimo vale 1 y el resto vale 0, y el vector $\vec{r}_i$ es el vector de ranking, donde cada elemento $r_{i,j}$ es la relevancia del nodo i con respecto al nodo j.

Adicionalmente, una vez respondida esta pregunta es posible responder una pregunta más compleja: dado un nodo v, obtener una lista ordenada del resto de nodos en orden de similitud o distancia a v. Este proceso se denomina **priorización** de los nodos de la red respecto a v, y puede ser costoso dependiendo del volumen de la red.

Existen además otros tipos de procesos dinámicos similares al RWR, como puede ser el *self-avoiding* walk, en el cual no se permite visitar el mismo nodo más de una vez. Este tipo de camino dinámico, al contrario que el RWR, necesita ser consciente de los nodos visitados, y fue introducido por primera vez en la teoría de polimerización [79]. Otro tipo de proceso de similares características son los *tourist walks*. En este tipo de camino, el agente o partícula se asemeja

a un turista que busca visitar ciertos sitios en un mapa P-dimensional. En cada paso, el turista intenta visitar el el sitio más cercano que no ha sido visitado en los últimos μ pasos. Se puede entender que el turista realiza algo parecido a un *self-avoiding* walk parcial de longitud μ, evitando únicamente los sitios que ha visitado recientemente.

1.2.6. Análisis de grafos en biología computacional

La gran cantidad de datos disponible en la actualidad, así como su diversidad en naturaleza y en el amplio espectro de tipos de entidades biológicas que representan, ha propiciado la aparición de multitud de enfoques computacionales para la extracción de conocimiento de datos biológicos basadas en grafos [53].

1.2.6.1. Ámbitos de aplicación en la biología

En el área de la biología computacional se aplica el análisis de redes biológicas a una gran diversidad de áreas de interés, entre las que podemos encontrar, entre otras, redes de interacción proteína-proteína (PPI), redes metabólicas, redes de enfermedades, redes de medicamentos o redes de interacción de ARN. La integración de estos tipos de datos es por ello de gran interés para la extracción de conocimiento biológico.

Redes de interacción proteína-proteína

Las proteínas son uno de los productos principales en la producción de la maquinaria celular, siendo las redes de interacción proteína-proteína (PPI) una fuente de información utilizada extensivamente para extraer conocimiento sobre la funcionalidad de las proteínas. Existen numerosas bases de datos que recopilan este tipo de información, entre las que destacan:

- **BioGRID**[28] [80]. Pretende aglutinar toda la información existente sobre interacción entre entidades biológicas de todas las especies extrayéndola de

[28] Biological General Repository for Interaction Datasets. https://thebiogrid.org

la literatura biomédica. Es actualizada frecuentemente y en la actualidad cuenta con un total de 515032 interacciones (no redundantes) repartidas entre 30 organismos modelo.

- **STRING**[29] [81]. Base de datos de interacción proteína-proteína que incluye relaciones tanto directas (físicas) como indirectas (funcionales).
- **UniProt**[30] [82]. Base de datos de secuencias de proteínas y sus anotaciones.
- **HPRD**[31] [83]. Una base de datos proteómicos en la especie humana manualmente curados. En la actualidad cuenta con 30047 proteínas y 41327 interacciones proteína-proteína.

Redes metabólicas

Las redes metabólicas son el conjunto de procesos físicos y metabólicos que determinan las propiedades fisiológicas y bioquímicas de una célula. Esto comprende las reacciones químicas del metabolismo, las rutas metabólicas y las interacciones reguladoras que comprenden estas actividades. Una ruta metabólica es una cadena de reacciones químicas que ocurren en una célula [84]. La principal base de datos que contiene este tipo de información es KEGG[32] PATHWAY [85]. Otros recursos para el estudio de redes metabólicas disponibles son Pathway Ontology [86], una ontología para la anotación estandarizada de genes de varias especies en términos metabólicos, y la Pathway Interaction Database [87], una base de datos de anotaciones manualmente curadas realizada como una colaboración entre el US Cancer National Institute y el Nature Publishing Group.

Redes de enfermedades

Existen multitud de bases de datos que relacionan enfermedades entre sí, ya que este tipo de redes son de interés para gran cantidad de áreas de aplicación,

[29] Search Tool for Recurring Instances of Neighbouring Genes. http://string-db.org
[30] http://www.uniprot.org
[31] Human Protein Reference Database http://www.hprd.org
[32] Kyoto Encyclopedia of Genes and Genome. http://www.genome.jp/kegg/

desde priorización gen-enfermedad [88], priorización mutación-enfermedad [89] o reposicionamiento de medicamentos [90]. Entre estas fuentes de información existen ontologías, bases de datos manualmente construidas o bases de datos construidas a partir de sofisticadas metodologías de text-mining aplicadas a la literatura biomédica:

- **OMIM**[33] [91]. Base de datos que incluye información manualmente curada sobre fenotipos, genes y la relación entre ellos. Actualmente incluye 15427 entradas para genes y 4888 entradas de fenotipo.
- **Disease Ontology (DO)**[34] [92]. Una ontología con información sobre enfermedades que engloba actualmente 8043 fenotipos diferentes.
- **DISEASES**[35] [93]. Una base de datos que contiene relaciones gen-enfermedad obtenidas de la literatura biomédica mediante aplicación de metodologías de *text-mining* y curada manualmente, de estudios GWAS y datos de mutación en cáncer.

Redes de medicamentos, compuestos químicos, efectos secundarios

La información sobre compuestos químicos, sus características, su relación con dianas terapéuticas así como los posibles efectos secundarios de su administración en pacientes, son fuentes de información que también abundan entre las bases de datos públicas. En particular, estas bases de datos suelen tener gran volumen.

- **DrugBank 4.0**[36] [94]. Base de datos que incluye información bioquímica sobre los principios activos de los fármacos y sus dianas terapéuticas. Incluye 8246 compuestos incluyendo 2012 pequeñas moléculas aprobadas por la FDA.
- **PubChem**[37] [95]. Es un repositorio público de datos procedentes de experimentos de alto rendimiento proporcionados por más de 50

[33] Online Mendelian Inheritance in Man. http://www.omim.org
[34] http://disease-ontology.org
[35] http://diseases.jensenlab.org
[36] https://www.drugbank.ca
[37] http://pubchem.ncbi.nlm.nih.gov

organizaciones, incluyendo empresas farmacéuticas, instituciones públicas y laboratorios de investigación, entre otros.

- **ChEMBL** [96]. Base de datos de actividad biológica de compuestos químicos, dianas terapéuticas, indicaciones y literatura biomédica. Su versión más reciente, 22.1, ha sido lanzada en noviembre de 2016 con más de 11.000 dianas terapéuticas, y más de millón y medio de compuestos químicos
- **SIDER**[38] [97]. Una base de datos pública de efectos secundarios de medicamentos compuesta por conexiones entre 888 compuestos y 1450 términos de efectos secundarios.

Redes de interacción de ARN

Las bases de datos de interacción de ARN incluyen redes de interacción ARN-ARN yo redes de interacción ARN-ADN. Existen bases de datos como microRNA[39], miRbase [98] o miRDB [99] para interacciones miRNA-gen, StarBase [100] para interacciones lncRNA-proteína, miRNA-ncRNA.

1.2.6.2. Metodologías existentes para el análisis de redes biológicas

El análisis de datos biológicos conlleva en muchos casos la aplicación de metodologías de minería de datos en grafos. En el caso de la aplicación en la biología computacional, estas metodologías se dividen en dos categorías generales: aquellas que realizan sus análisis sobre una sola *red homogénea*, y aquellas que integran diversas fuentes de información y tipos de entidades biológicas en *redes heterogéneas*.

Entre las metodologías existentes para el análisis de *redes homogéneas*, donde sólo un tipo de entidad es representada (*e.g.* redes de interaccion proteína-proteína, redes de genes), se pueden dividir los posibles análisis en dos categorías.

El primero consiste en un análisis de tipo *estructural*, que busca ciertas propiedades dentro de la red, ya sea con alguna información de entrada o sin

[38] Side-Effect Resource http://sideeffects.embl.de/
[39] http://www.microrna.org

conocimiento a priori. Entre este tipo de análisis encontramos varios problemas diferentes [101]:

1. Búsqueda de estructuras similares a una subred de entrada dentro de la red principal (*network querying*).
2. *Alineamiento de redes*. Consiste en buscar las subredes en común entre dos redes de entrada.
3. *Búsqueda de motivos en redes*. Búsqueda de patrones o subredes que se repiten dentro de la red principal, las cuales se espera que tengan alguna funcionalidad.

Por otro lado, encontramos el problema de la *priorización*, que consiste en cuantificar el nivel de relación de los nodos de una red con respecto a un nodo de entrada. Se puede definir el problema de priorización de la siguiente forma:

Definición 28. ***Priorización***. *Sea $G = (V, E)$ un grafo simple no dirigido, $u \in V$ un nodo que denominaremos nodo de consulta. Se define la priorización de u con respecto a G como una lista ordenada de los elementos de $V \setminus \{u\}$ en base a una función $\varphi : V \times V \to \mathcal{R}$ que define la fuerza de la relación entre dos elementos de V.*

Nótese que la definición de la función φ es lo que define a los diversos enfoques priorizadores existentes, y, en general, es lo que se entenderá como *score* de un nodo en un problema de priorización. Entre los enfoques priorizadores existentes se encuentran metodologías computacionales más o menos generales que permiten el análisis de una red de entrada introducida como una matriz de adyacencia. Un ejemplo de este tipo de metodología es RANKS [102], una herramienta de priorización basada en funciones kernel. RANKS calcula la relación entre todos los nodos de la red mediante una transformación kernel de la matriz de adyacencia. Así, en la nueva matriz transformada, $M(i, j)$ representará la distancia del nodo i al nodo j, que ha sido calculada teniendo en cuenta la estructura global de la red. Luego, para un nodo de consulta y un conjunto de nodos considerados *positivos* con respecto a la propiedad que se pretende analizar, se calcula la distancia entre ambos grupos mediante una función *score* que integra el conocimiento local del entorno

del nodo de consulta en el cálculo. Esta agregación se puede realizar de varias maneras posibles (máximo, media, *k*-vecinos). Las funciones de transformación que calculan la matriz kernel también se pueden elegir de entre un conjunto de funciones diferentes, cada una con sus parámetros propios: lineal, gaussiana, cauchy, polinómica, laplaciana, entre otras.

El resto de metodologías existentes para el análisis de redes homogéneas en el ámbito de la biología computacional están fuertemente acopladas al área de aplicación. En este sentido, las herramientas tratan de resolver un problema concreto, y en ellas la construcción de la red de análisis a partir de una o varias fuentes de información forma parte de la metodología ofrecida. Nótese que, aunque en estos casos pueda existir una integración de varias fuentes de información, el tipo de elementos que representan los nodos pertenece a un único ámbito, y es por ello que los llamamos métodos de análisis en redes homogéneas. Entre estos enfoques encontramos metodologías de análisis de redes PPI, como [103], o de genes, como SVD-phy [104], que hace uso de información filogenética para conectar los genes entre sí. DRaWR [105], por otro lado, trata de extender el concepto de red homogénea añadiendo la posibilidad de varios tipos de aristas, lo que convierte el grafo en un multigrafo. Por último, metodologías como FunRich [106] tratan de solucionar el problema de la generalidad proporcionando al usuario numerosas fuentes de información entre las que elegir para construir la red de análisis e incluso la personalización de la misma.

Existen multitud de metodologías que extienden el número de tipos de elementos incluidos en el análisis a un número mayor de uno, por lo que las consideraremos metodologías de análisis en *redes heterogéneas*. Sin embargo, las metodologías de este tipo existentes en la literatura están circunscritas a ámbitos concretos de aplicación, como por ejemplo, los métodos que realizan priorización entre genes o proteínas y enfermedades [107, 108].

Otro ámbito que está recibiendo gran interés entre las metodologías de análisis de redes heterogéneas desde hace unos años es la búsqueda de nuevas aplicaciones para fármacos ya comercializados, ámbito que se conoce como reposicionamiento de medicamentos. El éxito en la propuesta de una nueva aplicación para un

medicamento ya existente en el mercado puede ahorrar enormes costes tanto temporales como económicos en el proceso de desarrollar un tratamiento para una enfermedad. Entre las metodologías que integran fuentes de información para el reposicionamiento de medicamentos encontramos metodologías que construyen redes enfermedad-medicamento [109] y otras más específicas, las cuales buscan la relación no con enfermedades sino con dianas terapéuticas [110, 111]. De nuevo, estas metodologías están vinculadas a las fuentes de información que utilizan para construir las redes, no permitiendo al usuario definir sus propias redes de análisis.

1.3. Análisis de secuencia

Los diversos niveles de estructuración de la información genética permiten su estudio en una amplísima variedad de enfoques, entre los cuales el análisis de secuencia se puede considerar uno de gran actividad, y en el cual se centra una de las partes principales de este trabajo. Por ello, en esta sección se incluye una introducción a las técnicas de análisis de secuencia más relevantes en relación a las contribuciones propuestas en esta tesis. En primer lugar, se realiza una introducción a la teoría de conjuntos difusos e intuicionistas aplicada al análisis de secuencias. A continuación, se realiza un análisis de los problemas relacionados con la detección y anotación de lugares de unión de factores de transcripción (TFBSs), la representación computacional de los mismos y las herramientas actuales disponibles para este fin. Posteriormente, se realiza una descripción de las bases de datos disponibles en la actualidad y relacionadas con la propuesta que se plantea en el presente trabajo. Por último, se describen las propuestas ya existentes en cuanto a detección de mutaciones en lugares de unión de factores de transcripción.

1.3.1. Medidas difusas de comparación secuencia-motivo

Trabajar con información imperfecta, generalmente más fiel al tipo de información que manejamos para desenvolvernos en el mundo real, plantea una serie de problemas a la hora de interactuar con un sistema de información. Esto ocurre porque trabajamos con vaguedad e incertidumbre, mientras que los enfoques computacionales, en muchos casos, evitan esta incertidumbre. Los seres humanos somos muy buenos trabajando con este tipo de información, ya que no tenemos una percepción tan exacta como para diferenciar, por ejemplo, si un vehículo circula a 60 o a 65 kilómetros por hora, pero sin embargo sí sabremos decir si iba *deprisa* cuando lo vimos acercarse, y puede que tal vez esa información imprecisa incluida dentro de la etiqueta *deprisa* nos sea suficiente para tomar una decisión, como puede ser la de cruzar la calle.

Sin embargo, para poder utilizar este tipo de información en metodologías de Inteligencia Artificial es necesario un modelo que permita operar en este tipo de términos. Dicho de otro modo, cuando la información que el usuario tiene disponible no es suficientemente buena como para hacer consultas al sistema o introducirla en éste, o la que obtiene no es suficiente, se hace necesario desarrollar mecanismos que permitan manipular la vaguedad e incertidumbre presente en los datos. En el mundo real se trabaja mucho más con conceptos e información difusos, más afines al razonamiento humano que la información perfecta, y muy especialmente, en el caso de la información biológica, nos vemos obligados a tratar con imprecisión e incertidumbre.

Con el fin de resolver este tipo de problemas aparece en 1965 el concepto de conjunto difuso, introducido por Zadeh [112]. Zadeh extiende el concepto de conjunto en la matemática clásica introduciendo el concepto de *grado de pertenencia* de un elemento a un conjunto, pudiendo éste ser cualquier valor entre 0 y 1, y por tanto creando un conjunto de bordes difusos, definido por una propiedad imprecisa.

Definición 29. *Sea X un conjunto cualquiera (finito o infinito). Se define un* ***conjunto difuso*** *A sobre X cuando existe función de pertenencia μ_A:*

$$\mu_A : X \to [0,1]$$

donde $\mu_A(x \in X)$ es el grado de pertenecia de x a A:

$$\mu_A(x) = \begin{cases} 1 & \textit{si es claro que } \ x \in A \\ 0 & \textit{si es claro que } \ x \notin A \\ \alpha \in (0,1) & \textit{en otro caso} \end{cases}$$

Siendo μ_A la función de pertenencia de cierto elemento x a un conjunto X, aparece la siguiente definición:

Definición 30. *α-corte. Sea A un conjunto difuso sobre X, y μ_A su función de pertenencia. Se define el α-corte como el conjunto:*

$$A_\alpha = x \in X : \mu_A(x) \geq \alpha$$

Donde:

$$A = \bigcup_{\alpha>0}^{1} A_\alpha$$

A partir de esta definición aparecen la *moda* o α-corte de nivel 1, y el *soporte*, conjunto de todos los elementos tales que $\mu_A > 0$. Una de las principales utilidades de los conjuntos difusos es representar información imprecisa, frecuente en el lenguaje humano. Así, un conjunto difuso puede ser definido por una propiedad imprecisa, como por ejemplo ser joven (etiqueta lingüística). En este caso, $\mu_{joven}(x)$ sería el grado de cumplimiento de esta propiedad para un individuo x.

1.3.1.1. Operaciones sobre conjuntos difusos

Definición 31. *Inclusión difusa. Dados dos conjuntos difusos A y B sobre X, se dice que A está incluido en B:*

$$A \subseteq B \iff \forall x \in X \mu_A(x) \leq \mu_B(x)$$

Definición 32. *Intersección difusa. Dados dos conjuntos difusos A y B sobre X, su intersección C se define como:*

$$\forall x \in X; \mu_C(x) = \mu_A(x) \odot \mu_B(x) = i(\mu_A(x), \mu_B(x))$$

Normalmente, la operación $\odot$ suele definirse como el mínimo entre $\mu_A(x)$ y $\mu_B(x)$, aunque puede ser cualquier operación que cumpla las propiedades de t-norma:

Definición 33. *t-norma. Se define una t-norma como una aplicación*

$$i : [0,1] \times [0,1] \to [0,1]$$

que cumple las siguientes propiedades:

- *Asociatividad.* $\forall x, y, z \in [0,1];\ i(i(x,y),z) = i(x,i(y,z))$
- *Conmutatividad.* $\forall x, y \in [0,1];\ i(x,y) = i(y,x)$
- *Monotonía.* $\forall x, y, z \in [0,1];\ y \leq z \implies i(x,y) \leq i(y,z)$
- *Acotación.* $\forall x \in [0,1];\ i(1,x) = x$

Definición 34. *Unión difusa. Dados dos conjuntos difusos A y B sobre X, su unión C se define como:*

$$\forall x \in X; \mu_C(x) = \mu_A(x) \oplus \mu_B(x) = u(\mu_A(x), \mu_B(x))$$

Normalmente, la operación $\oplus$ suele definirse como el máximo entre $\mu_A(x)$ y $\mu_B(x)$, aunque puede ser cualquier operación que cumpla las propiedades de t-conorma:

Definición 35. *t-conorma. Se define una t-conorma como una aplicación*

$$u : [0,1] \times [0,1] \to [0,1]$$

que cumple las siguientes propiedades:

- *Asociatividad.* $\forall x, y, z \in [0,1];\ u(u(x,y),z) = u(x,u(y,z))$
- *Conmutatividad.* $\forall x, y \in [0,1];\ u(x,y) = u(y,x)$
- *Monotonía.* $\forall x, y, z \in [0,1];\ y \leq z \implies u(x,y) \leq u(y,z)$
- *Acotación.* $\forall x \in [0,1];\ u(0,x) = x$

Definición 36. *Complemento. Dado un conjunto difuso A sobre X, su complemento C se define como:*

$$\forall x \in X;\ \mu_C(x) = n(\mu_A(x))$$

Inicialmente esta operación suele definirse como $n(x) = 1 - x$, aunque puede definirse cualquier operación que cumpla las propiedades de negación.

Definición 37. *Negación. Se dice que* $n : [0,1] \to [0,1]$ *es una negación si verifica:*

- $n(0) = 1$ *y* $n(1) = 0$.
- n *es continua.*
- *Monotonía.* $\forall x, y \in [0,1];\ x \leq y \implies n(x) \geq n(y)$.
- n *es involutiva.* $\forall x \in [0,1], n(n(x)) = x$.

1.3.1.2. Relaciones difusas

El concepto de relación difusa es una extensión natural del concepto de relación *crisp*, en donde las interacciones entre elementos no son binarias sino más o menos fuertes. Se ha trabajado mucho en las relaciones binarias difusas y en la generalización de conceptos crisp como equivalencia y orden, dando lugar a la similaridad y orden difusos. Sin embargo, se ha visto que muchas de estas herramientas no son totalmente nuevas, ya que la similaridad guarda conexión con el concepto de distancia, y los órdenes difusos contienen elementos no dominantes o no dominados.

1.3.1.3. Conjuntos intuicionistas

La teoría de conjuntos difusos se centra en el concepto de *grado de pertenencia* de un elemento a un conjunto. Sin embargo, no considera información relativa a la *no pertenencia* de un elemento a un conjunto. En determinadas circunstancias podría suceder que hubiera cierta información al respecto. Con esta idea en mente, Attanassov [113] propuso la teoría de conjuntos difusos intuicionsitas añadiendo el concepto de *grado de no pertenencia*, lo que significa mayor flexibilidad a la hora de representar nuestro grado de comprensión de la realidad. Un conjunto difuso intuicionista se define de la siguiente forma:

Definición 38. *Conjunto difuso intuicionista. Sea* X *el universo de discurso. Un conjunto intuicionista* A *en* X *es un objeto de la forma:*

$$A = \{(x, \mu_A(x), \nu_A(x)) : x \in X\}$$

donde $\mu_A, \nu_A \rightarrow [0,1]$ *representan las funciones de pertenencia y no pertenencia en A, respectivamente, satisfaciendo además que* $0 \leq \mu_A + \nu_A \leq 1$ *para todo* $x \in X$. *Por tanto, el grado de incerteza de x con respecto a A es* $\pi_A(x) = 1 - \mu_A(x) - \nu_A(x)$.

Es interesante destacar que no sólo se introducen los conceptos pertenencia y no pertenencia, sino que además su suma está acotada entre 0 y 1 y define un tercer concepto, que es el grado de incerteza. La teoría de conjuntos difusos intuicionistas ha sido aplicada en numerosos campos, tales como programación lógica [114], diagnóstico médico [115, 116], toma de decisiones [117] o reconocimiento de patrones [118].

1.3.2. Lugares de unión de factores de transcripción

Como se ha mencionado previamente, los lugares de unión de factores de transcripción o TFBSs son fragmentos cortos de ADN a los cuales se unen los factores de transcripción, proteínas que influyen en la activación o desactivación de la producción de proteínas en una determinada célula. Estas secuencias comparten una serie de características que hacen que los TFs se unan a ellas. Estas características son las que actualmente se estudian con el objetivo de arrojar luz sobre el descubrimiento de TFBSs. Por lo general, existen dos vertientes diferenciadas en la investigación de lugares de unión de factores de transcripción: el descubrimiento de TFBSs *de novo*, y la búsqueda de TFBSs conocidos en secuencias.

1.3.2.1. Búsqueda de motivos de novo

El descubrimiento de TFBSs de novo consiste en la búsqueda de estas regiones sin conocimiento previo sobre el motivo o las posibles secuencias que podrían formar parte de él. Existen técnicas experimentales para descubrir regiones de ADN que interactúan con proteínas, como es el caso de TFBSs, mediante las técnicas basadas en inmunoprecipitación de cromatina o ChIP. En este tipo de experimentos, se aplica un reactivo inmunológico específico para un factor de unión al ADN con el fin de enriquecer los lugares del ADN objetivo en los que el factor se unió

en células vivas. Estos lugares se miden y cuantifican a continuación. El análisis de la información producida por técnicas NGS aplicadas a experimentos ChIP (ChIP-Chip [119] o ChIP-seq [120]) ha resultado en una mayor disponibilidad de información sobre motivos obtenidos de forma experimental. La tecnología ChIP-Chip ha sido paulatinamente sustituida por ChIP-Seq. Esta última consiste en la inmunoprecipitación de complejos proteína-ADN seguidos de secuenciación masivamente paralela de *short ends* de ADN inmunoprecipitado [49]. Al término de uno de estos experimentos, se obtienen millones de etiquetas direccionales cortas ($\sim 35-50bp$) de ADN, las cuales se pueden alinear al genoma de referencia para el organismo objetivo. Cada etiqueta representa el final de una secuencia más larga ($\sim 200-400bp$) o *read*. Teniendo en cuenta estas secuencias se puede realizar una tarea de análisis de las posiciones donde existe abundancia de solapamiento de reads, los cuales se identifican como posibles lugares de la interacción de la proteína con el ADN.

Esta información debe ser posteriormente analizada mediante herramientas computacionales para obtener lo que se conoce como motivos, o representaciones de secuencias a las que se puede unir un factor de transcripción. Existen en la actualidad multitud de herramientas como FindPeaks [121] o MICSA [120], un sistema que incluye FindPeaks en su desarrollo y que posteriormente elimina regiones satélite, picos falsos y aplica RepeatMasker[40]. RepeatMasker se encarga de buscar repeats y secuencias de ADN de baja complejidad y enmascararlas. Otros enfoques incluyen Hybrid Motif Sampler [122], que propone un modelo bayesiano que incorpora información sobre secuenciación para mejorar la identificación de motivos, DBChIP [123], que propone obtener listas de posibles TFBSs de varias fuentes, mezclarlas y aplicar clustering sobre ellas, o peak-motifs [124], un pipeline que descubre TFBSs analizando resultados de ChIP-Seq para posteriormente compararlos con bases de datos existentes y visualizarlos en el Genome Browser de la UCSC.

Sin embargo, estos experimentos únicamente aportan información sobre los tipos específicos de tejidos y las condiciones utilizadas. Además, la mayoría de

[40] http://www.repeatmasker.org

los factores de transcripción no han sido perfilados a escala genómica, debido tanto al coste de dichos experimentos como a la necesidad de un anticuerpo específico disponible para cada tipo de factor de transcripción [125]. Otros métodos proponen obtener información de varias fuentes para determinar si existe o no probabilidad de que la secuencia analizada sea un TFBS. Tal es el caso de Ernst et al. [125], quienes proponen un score GBP[41] que aglutina información general sobre la localización del presunto TFBS (distancia al punto de comienzo de transcripción más cercano, conservación, etc.) junto con información específica de motivo para mejorar la predicción.

Otra metodología para el descubrimiento de TFBSs de novo parte de la base de que, entre especies relacionadas evolutivamente, regiones de ADN conservadas entre especies relacionadas (aquellas secuencias que guardan similitud en dichas especies) tienen mayor probabilidad de ser regiones relevantes en el ADN. De hecho, la tasa de cambio en los TFBSs es más baja que en otras secuencias no codificantes [126]. Esta técnica se conoce como *phylogenetic footprinting*, y se aplica mediante la comparación de secuencias ortólogas. Dos secuencias se dice que son ortólogas si presentan un alto índice de similaridad entre especies debido a que tienen un ancestro común. Existen algoritmos para comparar este tipo de secuencias y encontrar motivos repetidos [127]. Con respecto al análisis de secuencias provenientes de diversas especies, lo que se conoce computacionalmente como alineamiento múltiple de secuencias, Kawrykow et al. proponen Phylo [128], un juego online donde los usuarios buscan patrones manualmente en secuencias relacionadas por filogenia, procesándose después esa información para obtener patrones de ella.

Además, la identificación de TFBSs de novo no implica que el motivo que se encuentre no haya sido encontrado previamente. Por ello son necesarias medidas de similitud entre motivos que permitan averiguar si el perfil obtenido corresponde a otro motivo ya obtenido con anterioridad. Por ello hay herramientas que proponen buscar en bases de datos de motivos conocidas como JASPAR [129] o TRANSFAC [130] una vez se ha encontrado un motivo, y actualizarlo en caso

[41] General Binding Preference.

de que esto sea necesario [124, 125]. Por otra parte, para determinar si un motivo se encuentra o no en bases de datos conocidas, es necesario compararlo de alguna forma contra los motivos existentes en dichas bases de datos. Con este fin, existen numerosas aproximaciones computacionales, tales como FISim [131], una medida de similitud entre motivos basada en integral difusa, propuestas basadas en métodos estadísticos, MatCompare [132], que compara matrices de motivos basándose en distribuciones multinomiales mediante un test χ^2 de Pearson, propuestas como la de Choi et al. [133], que utilizan distancia euclídea entre columnas, o Tomtom [134] que permite cualquier medida de similitud entre columnas.

1.3.2.2. Representación de motivos

Un TFBS no está representado por una única secuencia, sino que existe cierta variación en las secuencias que pueden ser identificadas como TFBS de un determinado factor de transcripción. Como ejemplo, en la tabla 1.2 se puede observar como para el motivo USF1 de la figura 1.12 existen 30 secuencias, y no todas son iguales. Existe cierta variación en las posiciones, siendo algunas muy conservadas, como por ejemplo la primera posición, donde siempre aparece la citosina. Una forma de representar estos motivos se puede ver en la figura 1.12, bajo el título de Sequence Logo. Esta forma de visualizar los motivos, conocida como logo [135], permite visualizar qué posiciones de la secuencia están más conservadas y cuales ofrecen mayor variación. Una aplicación web de generación de logos, WebLogo, puede encontrarse en [136].

En la figura 1.12 se puede observar además una matriz de frecuencias, que indica, en términos absolutos, el número de veces que aparece cada base en cada posición. Las versiones normalizadas de estas matrices se conocen como PWMs (Position-Weight Matrices), y son la base para el cálculo de muchas medidas de similitud secuencia-motivo. Este tipo de información se conoce como motivo, y se obtiene mediante métodos de descubrimiento de TFBSs de novo como los descritos en el apartado anterior. Los resultados de éstos son almacenados en bases de datos de motivos contra las cuales, mediante diversos enfoques computacionales,

Tabla 1.2: *Secuencias proporcionadas para el motivo USF1 en JASPAR.*

Secuencias	
CACGTGG	CACGTGA
CACGTGG	CACGTGA
CACGTGG	CACGTGA
CACGTGG	CACGTGT
CACGTGG	CACGTGT
CACGTGG	CACGTGT
CACGTGG	CACGTGT
CACGTGG	CACGTGT
CACGTGG	CACGTGC
CACGTGG	CACGTGC
CACGTGG	CATGTGG
CACGTGG	CATGTGA
CACGTGA	CACATGA
CACGTGA	CACGCGG
CACGTGA	CACGGGA

es posible comparar una secuencia determinada para saber si dicha secuencia es candidata a ser uno de estos TFBSs conocidos. Las dos bases de datos más utilizadas para obtener este tipo de información son JASPAR y TRANSFAC. Ambas contienen una lista de TFBSs encontrados donde para cada uno de ellos se detalla la información disponible sobre ellos. Entre la información más relevante que se puede encontrar en estas bases de datos están las matrices PWM que indican las proporciones de bases en cada posición, y la lista de secuencias que generaron esa matriz. En ciertos casos, dependiendo de la técnica con la que se encontró el motivo, las secuencias concretas que lo generaron aparecerán o no detalladas.

Las características principales de JASPAR y TRANSFAC se detallan a continuación. **JASPAR** es una base de datos de acceso gratuito cuya información se puede usar sin restricciones (es open-source). Se divide en dos bloques principales, JASPAR CORE y JASPAR COLLECTIONS.

- **JASPAR CORE**. Es la más utilizada y contiene un conjunto de motivos no redundante. Uno de los principales objetivos de JASPAR CORE es

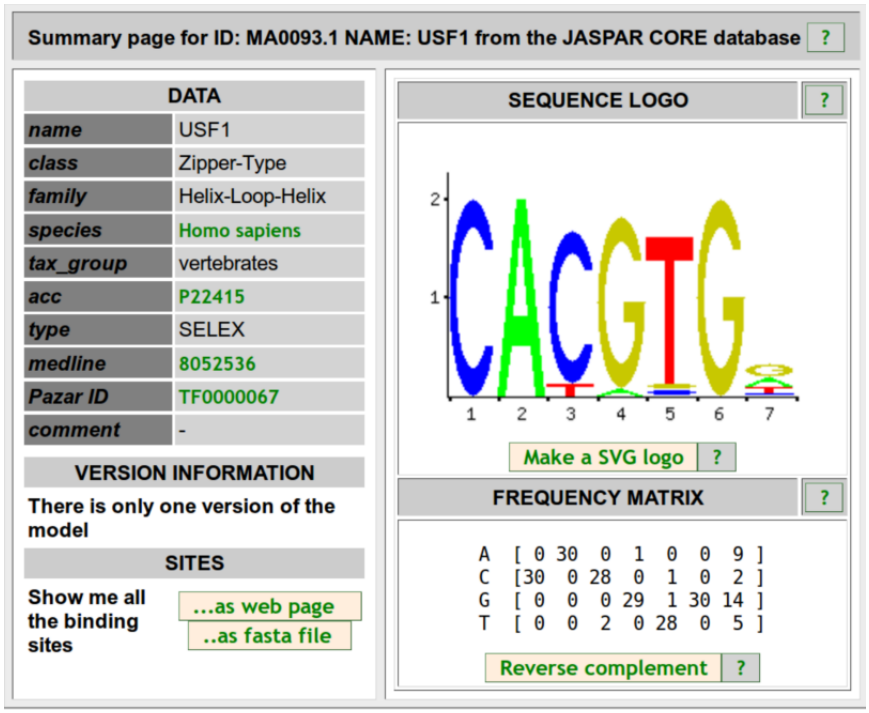


Figura 1.12: *Información proporcionada por JASPAR para el motivo MA0093.1, USF1, donde se observan datos básicos sobre el motivo, la matriz de frecuencias y logo que la representa. De forma visual observamos que el motivo generalmente consta de una secuencia CACGTG relativamente constante seguido de un nucleótido que varía bastante entre A, C, G y T. Es posible además consultar las secuencias que dieron lugar a la matriz de frecuencias en el apartado sites.*

proporcionar la mejor representación para cada TFBS. En este sentido, la base de datos es no redundante ya que se procura evitar que existan varias representaciones para el mismo TFBS (con algunas excepciones debido a la complejidad biológica en ciertos casos).

- **JASPAR COLLECTIONS**. Es un conjunto de bases de datos que no se pueden incluir en JASPAR CORE. Entre este tipo de información se incluyen variantes de splicing, patrones derivados computacionalmente o metamodelos. Las bases de datos incluidas en ellas on:
  - JASPAR FAM, la cual contiene 11 modelos que describen propiedades compartidas de unión de clases estructurales de factores de transcripción. Son lo que se puede entender como perfiles familiares, matrices consenso o metamodelos.
  - JASPAR PHYLOFACTS contiene 174 perfiles que fueron extraídos de elementos conservados filogenéticamente en zonas anteriores al gen [137]. La idea principal es que zonas altamente conservadas entre especies relacionadas evolutivamente representan zonas relevantes del genoma. En el caso de los TFBSs, áreas muy conservadas en la zona anterior al gen son candidatas a ser TFBSs. Alineando el genoma de especies relacionadas evolutivamente se pueden buscar estas zonas.
  - JASPAR POLII contiene 13 patrones enlazados con los promotores de núcleo de la RNA polimerasa II.
  - JASPAR CNE. Los elementos no codificantes muy conservados son una característica distintiva de los genomas animales. Muchos de ellos actúan como entradas reguladoras, y es por ello que se están empezando a investigar. JASPAR CNE contiene 233 perfiles derivados de Xie et al. [138] basado en clustering de motivos sobre-representados de elementos conservados no codificantes en humanos.
  - JASPAR SPLICE. Contiene perfiles humanos de lugares de splicing canónicos y no-canónicos. Actualmente sólo contiene 6 perfiles.
  - JASPAR PBM. Las colecciones PBM se construyen mediante el uso de nuevas técnicas *in vitro* basados en k-mer microarrays. Los modelos de

matrices PBM tienen su propia base de datos especializada: UniPROBE [139].

- JASPAR PBM HOMEO. Al igual que la anterior, proviene de la aplicación de técnicas *in vitro*. En este caso, los datos fueron obtenidos a partir de [140]. Incluye 176 perfiles de homeodominios de ratón.
- JASPAR PBM HLH. Al igual que las dos anteriores, proviene de la aplicación de técnicas *in vitro*. En este caso incluye 19 modelos de factores de transcripción de *C. elegans bHLH*.

TRANSFAC es una base de datos que contiene elementos reguladores *cis* en organismos eucariotas. Al contrario que JASPAR, TRANSFAC no es una base de datos de acceso gratuito. Ofrece una versión anterior pública, de 2005, pero para poder disponer de la información actualizada se requiere una subscripción. En general, la información presente en TRANSFAC proviene de la literatura biomédica. En algunos casos se recoge de otras fuentes, aunque en dichos casos estas fuentes están documentadas. TRANSFAC consta de las siguientes tablas:

- FACTOR. Describe las proteínas que regulan la transcripción, es decir, los factores de transcripción, y el micro ARN (miARN) que controla la estabilidad o traducción del ARN mensajero (ARNm).
- GENE. Describe el gen al que pertenecen los TFBSs o fragmentos ChIP y/o genes que codifican factores de transcripción o miARNs.
- SITE. Proporciona información sobre lugares de unión de factores de transcripción individuales y partes de secuencia de mARN que son objetivo para interacción con miARN.
- FRAGMENT. Contiene fragmentos de ADN para los que se ha probado unión in vivo de un factor de transcripción mediante ChIP-on-chip, ChIP- Seq o experimentos relacionados.
- MATRIX. Contiene matrices de distribución de nucleótidos para los TFBSs. Además de la matriz PWM, TRANSFAC ofrece la secuencia de consenso IUPAC correspondiente. Se trata de un código de correspondencias entre

Tabla 1.3: *Correspondencia de secuencia de consenso IUPAC.*

Símbolo	Descripción	Nucleótidos			
A	**a**denine	A			
C	**c**ytidine		C		
G	**g**uanosine			G	
T	**t**hymidine				T
U	**u**ridine				U
W	**w**eak	A			T
S	**s**trong		C	G	
M	a**m**ino	A	C		
K	**k**eto			G	T
R	pu**r**ine	A		G	
Y	p**y**rimidne		C		T
B	cualquiera menos A		C	G	T
D	cualquiera menos C	A		G	T
H	cualquiera menos G	A	C		T
V	cualquiera menos T	A	C	G	
N	cualquier base	A	C	G	T

posiciones y letras más amplia que las habituales A, C, G, T, cuando hay posiciones donde no hay una predominancia evidente de una base.

- CLASS. Contiene información de background sobre las familias de factores de transcripción.
- CELL. Proporciona información breve sobre la fuente celular de las proteínas que se ha mostrado que interactúan con estos lugares.
- REFERENCE. Contiene referencias extraídas para TRANSFAC junto con enlaces a estas referencias en PubMed.

1.3.2.3. Búsqueda de TFBSs conocidos

La información sobre los TFBSs descubiertos de novo son almacenados en bases de datos de motivos contra las cuales, mediante diversos enfoques

computacionales, es posible comparar una secuencia determinada para saber si dicha secuencia es candidata a ser uno de estos TFBSs conocidos. En general, esta aproximación depende principalmente del criterio propuesto para decidir si la secuencia objetivo se parece o no a un motivo dado. En un principio se asumió independencia entre las posiciones de los motivos conocidos, por ejemplo en Patser [141] y ConSite [142]. Sin embargo, esta asunción ha resultado ser falsa [143, 144], y por ello se propusieron posteriormente dos métodos que tenían en cuenta dependencias entre posiciones de un motivo. Tomovic y Oakeley [145] propusieron un método que incorpora una medida de interdependencia al score final. Zare-Mirakabad et al. [146] desarrollaron un método basado en contenido de información junto con información mutua. En este método la interdependencia entre posiciones viene reflejada tomando en consideración todas las posibles combinaciones de pares de posiciones en el motivo.

El hecho de que las secuencias de TFBSs sean muy cortas aumenta la probabilidad de que aparezcan por azar las mismas secuencias o algunas muy parecidas. Esto hace que la tasa de falsos positivos sea más susceptible de aumentar, y por tanto, las mejoras en este campo tienden a buscar propuestas que reduzcan el número de falsos positivos. Los métodos que asumen interdependencia entre posiciones tienden a ser un poco más efectivos. En esta línea encontramos la propuesta de García et al. [147], SC_{intuit} que propone calcular un score de similitud considerando interacción entre columnas, teniendo en cuenta las secuencias que dieron lugar a la matriz.

La propuesta SC_{intuit} tiene la ventaja de ser más sofisticada que otras propuestas en el sentido de que no sólo se basa en la matriz PWM sino también en las secuencias concretas que dieron lugar a ella, estableciendo así una dependencia *horizontal* entre posiciones. Dicho de otro modo, asume cierta sinergia entre posiciones de la matriz, cuestión que, como se ha mencionado con anterioridad, permite mejorar sensiblemente la tasa de falsos positivos. Por otro lado, al ser una medida intuicionista, permite manejar la incertidumbre e imprecisión inherentes a los datos biológicos.

1.3.3. Bases de datos sobre secuencias, SNPs y regulación

El estudio sobre la variación en el genoma es de especial relevancia en lo que concierne al diagnóstico y tratamiento de enfermedades. Así pues, encontrar relaciones entre mutaciones genómicas y posibles trastornos o desórdenes es un campo importante dentro de los avances en el conocimiento sobre la biología de nuestro organismo en general y en el marco de la Medicina Personalizada en particular. Entre las posibles mutaciones que pueden aparecer en el genoma de un individuo, las más estudiadas son los polimorfismos de nucleótido simple o SNPs, ya que son las más presentes en el genoma, con una frecuencia de aparición aproximada cada 1200 bps en el genoma humano [148]. En esta sección se explica en detalle el contenido de las bases de datos más relevantes disponibles en la actualidad cuyo contenido involucra secuencias, SNPs, información sobre regulación génica y literatura biomédica.

1. **EnsEMBL.** El proyecto EnsEMBL[42] [149] comenzó en 1999, unos años antes de que la primera secuencia completa del genoma humano se completara [1]. No obstante, incluso entonces se hizo evidente que anotar manualmente las más de tres mil millones de bases que conforman el genoma humano no sería factible. En este sentido, la meta de EnsEMBL es ofrecer información sobre las secuencias de los genomas de diversas especies, así como la automatización de las anotaciones en ellas. Además, se ofrece la integración de esta información con el resto de información biológica disponible en otras fuentes.
2. **UCSC Genome Browser.** Genome Browser Database[43] [150] de la Universidad de California en Santa Cruz es una base de datos actualizada de secuencias genómicas integradas con una gran cantidad de anotaciones relacionadas. De ella se pueden obtener secuencias genómicas por coordenadas, anotaciones, y todo tipo de información relevante relacionada.
3. **dbSNP [148].** La base de datos con mayor información sobre SNPs disponible es dbSNP. En colaboración con el National Human Genome

[42] http://www.ensembl.org
[43] http://genome.ucsc.edu

Research Institute, el NCBI ha establecido dbSNP como repositorio central para tanto SNPs como polimorfismos cortos de inserción y borrado. Una vez descubiertos, estos polimorfismos podrían ser utilizados por laboratorios adicionales, mediante el uso de información sobre la secuencia de contexto y las condiciones experimentales específicas presentes en dbSNP. En ella se encuentran indexados los SNPs encontrados en las secuencias del genoma de diferentes especies, independientemente de si están o no relacionados con enfermedades. Entre la información que se puede encontrar en una entrada para un SNP en dbSNP, se encuentran su identificador único, información sobre la secuencia de contexto en la que aparece, sobre las condiciones experimentales en las que fue descubierto, descripciones de la población que contiene el SNP, así como la información de la frecuencia de alelos por población o por genotipo individual.

4. **ORegAnno [151].** Es una base de datos abierta que permite a la comunidad científica realizar anotaciones sobre regiones reguladoras del ADN. Esta base de datos ha sido diseñada para gestionar de forma conjunta el envío, indexación y validación de anotaciones de usuarios en cualquier lugar del mundo. Además, todas las nuevas entradas en ORegAnno están enlazadas con EnsEMBL, dbSNP, Entrez Gene, la base de datos taxonómica del NCBI y PubMed.

1.3.4. Detección de SNPs en TFBSs

Actualmente hay pocas propuestas relativas al estudio de las variaciones en las áreas reguladoras del genoma, centrándose la mayoría de las propuestas en el estudio de regiones codificantes, es decir, regiones que se traducen directamente en mARN que posteriormente dará lugar a la secuencia de aminoácidos que conforman la proteína. Cuando un SNP aparece en una secuencia de ADN en la que hay un lugar de unión de un factor de transcripción, dicha mutación puede alterar la afinidad de esta cadena de ADN con el factor de transcripción en cuestión, alterando así el proceso de regulación. En particular, un SNP en un TFBS se puede encontrar en una de cuatro situaciones distintas:

- **Inhibición de un TFBS**. El SNP altera la afinidad del TFBS de forma que el factor de transcripción que se unía a esta secuencia ya no es capaz de unirse. Se dice entonces que el SNP *inhibe* el proceso de transcripción, silenciando el gen.
- **Creación de un nuevo TFBS**. El SNP produce una alteración en la cadena de ADN de forma que se genera una afinidad con un TF que antes no existía, haciendo que un gen que antes no se expresaba lo haga.
- **Intercambio de TFBSs**. Se produce una combinación de las dos situaciones anteriores, silenciando un proceso de transcripción e iniciando otro.
- **SNP sinónimo**. El SNP no altera la afinidad del TFBS, así que no altera el proceso de regulación.

Es interesante distinguir entre estas cuatro situaciones, ya que de esta manera se obtiene una mayor cantidad de información sobre la probabilidad de que cada SNP altere el estado de salud de un individuo.

Existen metodologías que tratan de averiguar de forma computacional si un SNP afecta a un TFBS putativo. Entre ellas se encuentra is-rSNP [152]. Is-rSNP utiliza PWMs de TRANSFAC y JASPAR para evaluar la afinidad de unión alrededor del SNP estudiado. A continuación, calcula si la afinidad de unión es *alta* en comparación con la distribución de scores posibles. Otra metodología computacional para abordar este problema es sTRAP[44] [153]. sTRAP extiende la herramienta TRAP [154] para calcular afinidad de unión de fatores de transcripción y seleccionar mutaciones con alto log-ratio entre secuencia mutada y no mutada, aunque también permiten seleccionar secuencias con alta afinidad de unión con un TFBS de forma independiente del efecto estimado de la mutación. Existen otros enfoques relacionados con la búsqueda de elementos con funcionalidad en secuencias de ADN, como por ejemplo BCRANK[45] [155] o PupaSNP [156]. Más recientemente han aparecido métodos que basan su análisis en otras características epigenéticas diferentes a la afinidad de unión [157, 158].

[44] http://trap.molgen.mpg.de/cgi-bin/trap_two_seq_form.cgi – Acceso el 23-12-2016

[45] http://www.bioconductor.org/packages/release/bioc/html/BCRANK.html--Accesoel23-12-2016

1.3.5. Búsqueda de módulos reguladores

Por otro lado, en muchos casos la regulación de los genes no se produce mediante elementos aislados, tales como TFBS únicos, sino en conjuntos que actúan coordinadamente, los cuales se conocen como módulos reguladores [159].

Un módulo regulador en *cis* (CRM – *Cis*-Regulatory Module) es una secuencia de unos cientos de pares de bases [160] donde hay un conjunto de TFBSs que controlan la expresión de los genes cercanos. Se denominan *cis* en contraposición a *trans*, referido a efectos en genes lejanos o situados en la hebra opuesta de ADN. En un CRM, los factores de transcripción actúan en conjuntos coordinados que cooperan entre sí. A través de este mecanismo se puede entender mejor la gran variabilidad en la expresión de los genes, pues son estas combinaciones tan específicas las que permiten controlar con el nivel de detalle preciso los procesos de transcripción [161].

Al estudio de elementos reguladores se le añade, por tanto, una capa combinatoria. Si se suma la incertidumbre inherente a la predicción de TFBSs en secuencias con la complejidad de la búsqueda de *conjuntos* de TFBSs que actúan de forma coordinada, surge un problema difícil de abordar, especialmente por el tamaño del espacio de búsqueda. A continuación se describen las categorías en las que se pueden agrupar las metodologías de búsqueda de CRMs según el alcance de la información previa requerida para su funcionamiento [162]. Posteriormente se introducen los métodos computacionales de búsqueda de itemsets frecuentes, íntimamente relacionados con las propuestas en este trabajo.

Los enfoques más específicos, aquellos que requieren conocimiento *a priori* sobre la estructura, composición o ubicación del CRM, son de gran utilidad para estudiar procesos de regulación ya conocidos, donde el espacio de búsqueda de secuencias y TFs candidatos se puede reducir ampliamente. Estas metodologías se pueden agrupar en dos categorías, dependiendo del nivel de detalle de la información previa disponible para su uso:

Escáners [163, 164, 165]. Estos enfoques requieren un modelo específico del CRM a buscar y toman muchos parámetros como valores de entrada, como el número o la identidad de los TFs que actúan en el complejo o las distancias entre

sus TFBSs. La utilidad principal de estas metodologías es el estudio de problemas ya conocidos, en los que los actores (TFs involucrados e incluso la distribución de los TFBSs correspondientes) están ya bien definidos.

Constructores. Estas herramientas constituyen una generalización de los escáners mediante una relajación de algunos de sus parámetros. En general requieren para funcionar que el conjunto de secuencias de entrada sea más reducido. De forma análoga a las metodologías de búsqueda de TFBSs *de novo*, no sólo esperan que el conjunto de secuencias de entrada sea menor, sino que estas guarden entre sí alguna relación que pueda verse reflejada en las secuencias reguladoras. En este sentido, la mayoría de estos enfoques funcionan con secuencias promotoras de genes coexpresados o relacionados [166, 167], secuencias conservadas entre especies [168, 169] o una combinación de ambas características [170, 171].

Adicionalmente, existen algunas metodologías que tratan de abordar el problema de la detección de módulos CRM de una forma más general. Sin embargo, la complejidad combinatoria es tan alta que sigue siendo necesario para muchos de los métodos propuestos en un ámbito más general, reducir de algún modo las secuencias de entradas, no siendo posible hasta la fecha encontrar métodos que abarquen todo el genoma no codificante sin restricciones.

Por otro lado, la predicción de CRMs suele ir vinculada a la detección computacional de TFBSs, y el rendimiento de estos métodos vinculado asimismo a la calidad de la predicción de dichos TFBSs.

1.3.5.1. Búsqueda de itemsets frecuentes

En este trabajo se presenta una metodología para la detección de módulos CRMs utilizando una representación basada en patrones o itemsets frecuentes. A continuación se introducen las nociones básicas sobre este tipo de representación.

Los *patrones frecuentes* son modelos que se repiten en un conjunto de datos. Un conjunto de elementos que aparecen con frecuencia juntos en la cesta de la compra (pan, leche y huevos, por ejemplo) es lo que se denomina un *itemset frecuente*. En este apartado se introducen los conceptos básicos relacionados con la búsqueda de

itemsets frecuentes, así como los métodos más frecuentes de extracción de dichos itemsets [61].

Un ejemplo clásico de búsqueda de itemsets frecuentes es el análisis de la cesta de la compra. Este proceso analiza los hábitos de compra de los clientes buscando asociaciones entre los artículos que los clientes ponen en sus cestas de la compra, es decir, los artículos que son comprados simultáneamente. El conjunto de elementos en la compra de un cliente en un momento dado (i.e. lo que aparece en el ticket de compra), se denomina *transacción*. Las transacciones generan dependencias entre elementos. Así, si pensamos en el conjunto de elementos disponibles en la tienda, se le puede asociar a cada uno un valor booleano que represente su presencia o no en una transacción. Estos vectores de booleanos pueden ser analizados para buscar patrones de compra que reflejen elementos que están frecuentemente *asociados* o que son comprados a la vez. Se pueden por tanto representar como *reglas de asociación*. Por ejemplo, podríamos encontrar en una tienda de electrónica una asociación entre la compra de una cámara de fotos y un programa de edición de imagen:

$$\textit{cámara} \implies \textit{software-edición}\ [\textit{soporte} = 4\,\%, \textit{confianza} = 70\,\%]$$

Las medidas de *soporte* y *confianza* son medidas de interés en las reglas de asociación y representan tanto su utilidad como la certeza que representan. Un soporte de un 4% significa que un 4% de todas las transacciones en el análisis muestran que camara y software de edición son comprados juntos. Una confianza del 70 % representa que el 70 % de los clientes que compraron una cámara también compraron el software de edición. Generalmente, las reglas de asociación son de interés si satisfacen un umbral de soporte mínimo y un umbral de confianza mínimo.

Definiciones sobre itemsets y reglas de asociación

Sea $\mathcal{I} = \{I_1, I_2, ...I_m\}$ un conjunto de *items*. Sea D un conjunto de transacciones donde cada transacción T es un conjunto de items o *itemset* $T \subseteq \mathcal{I}$. Cada transacción está asociada a un identificador T_{id}. Un itemset $T = T_1, ..., T_k$ se denomina k-itemset.

Definición 39. ***Soporte de un itemset***. *Un itemset $A \subseteq \mathcal{I}$ tiene un* ***soporte absoluto*** *s, donde s es el porcentaje de transacciones en D que contienen todos los elementos de A.*

Definición 40. *Se dice que una transacción T contiene a A si y sólo si $A \subseteq T$.*

Un itemset que satisface un soporte absoluto mínimo se denomina **itemset frecuente**. Dado el conjunto de transacciones D, se definen a continuación las reglas de asociación.

Definición 41. ***Regla de asociación***. *Una regla de asociación es una regla de implicación de la forma $A \implies B$ tal que $A \subset \mathcal{I}$, $B \subset \mathcal{I}$ y $A \cap B = \emptyset$.*

Para cada regla de asociación T en D, se definen a continuación su **soporte** y su **confianza**.

Definición 42. ***Soporte de una regla de asociación***. *La regla $A \implies B$ tiene en D un* ***soporte*** *s, donde s es el porcentaje de transacciones en D que contienen $A \cup B$, o la probabilidad $P(A \cup B)$. Nótese que esta es la probabilidad de que una transacción T contenga todos los elementos de la unión $A \cup B$, diferente a la probabilidad $P(A \vee B)$.*

Definición 43. ***Confianza de una regla de asociación***. *La regla $A \implies B$ tiene en D una* ***confianza*** *c, donde c es el porcentaje de las transacciones en D que contienen A que también contienen B. Esto define la probabilidad $P(B|A)$.*

Las reglas que satisfacen unos umbrales mínimos de soporte y confianza se dice que son reglas *fuertes*. La confianza de una regla de asociación se puede deducir de los soportes de los itemsets que la componen:

$$c(A \implies B) = P(B|A) = \frac{s(A \cup B)}{A}$$

Es por ello que el problema de extraer reglas de asociación de un conjunto de transacciones se puede reducir a los siguientes pasos:

- **Encontrar todos los itemsets frecuentes**. Vendrán determinados por un umbral de soporte mínimo.

- **Generar reglas de asociación a partir de los itemsets frecuentes**. Estas reglas deben satisfacer unos umbrales mínimos de soporte y confianza.

El primer paso, en el que se centra una metodología de trabajo propuesta en esta tesis, es el más costoso de ambos. El principal problema que provoca encontrar itemsets frecuentes en un conjunto de transacciones es la enorme cantidad de itemsets frecuentes que aparecen, especialmente si el umbral de corte es bajo. Esto se debe a la combinatoria del problema, ya que si un itemset es frecuente, sus sub-itemsets también lo son. Por ejemplo, para un itemset frecuente $\{a_1, ..., a_{100}\}$ de longitud 100, existen $\binom{100}{1}$ 1-itemsets frecuentes, $\binom{100}{2}$ 2-itemsets frecuentes, y así sucesivamente:

$$\binom{100}{1} + \binom{100}{2} + ... + \binom{100}{100} = 2^{100} - 1 \approx 1{,}27 \times 10^{30}$$

Para solucionar este problema se define el *itemset frecuente cerrado* y el *itemset frecuente maximal*.

Definición 44. ***Itemset cerrado***. *Un itemset X es cerrado en un conjunto de datos S si no existe super-itemset $Y \supset X$ con el mismo soporte absoluto que X en S. Un itemset X es un itemset frecuente cerrado en S si es frecuente y cerrado en S.*

Definición 45. ***Itemset frecuente maximal (max-itemset)***. *Un itemset X es maximal en un conjunto de datos S si X es frecuente y no existe Y frecuente tal que $Y \supset X$.*

Sea $\mathcal{C}$ el conjunto de itemsets frecuentes cerrados para un conjunto de datos S que satisface un soporte mínimo. Sea $\mathcal{M}$ el conjunto de max-itemsets en S que satisface ese soporte mínimo. Además, se dispone del soporte absoluto para cada itemset en $\mathcal{C}$ y $\mathcal{M}$. Nótese que $\mathcal{C}$ contiene información completa para todos los itemsets frecuentes, ya que de ellos se puede extraer el set completo.

El algoritmo Apriori

En este apartado se describe el algoritmo básico de extracción de itemsets frecuentes propuesto por Agrawal y Srikant [172]: el algoritmo **Apriori**. Este

algoritmo recibe su nombre del hecho de que utiliza conocimiento *a priori* sobre las propiedades de los itemsets frecuentes. Apriori es un algoritmo iterativo que utiliza los k-itemsets para explorar (k+1)-itemsets. Primero busca los 1-itemsets frecuentes y los filtra en base a un soporte mínimo, obteniendo un conjunto denominado L_1. A continuación, L_1 es utilizado para construir L_2. Para construir cada L_k es necesario un recorrido completo de la base de datos transaccional.

Para mejorar la eficiencia de la generación de niveles se utiliza la **propiedad Apriori**: Todos los subconjuntos no vacíos de un itemset frecuente son también frecuentes.

Por definición, si un itemset I no es frecuente (i.e. $P(I) < s_{min}$), si añadimos un ítem i a dicho itemset, $I \cup i$ no puede ser más frecuente que I, por tanto $P(I \cup i) < s_{min}$.

El algoritmo Apriori consta de dos pasos cada vez que se construye L_k a partir de L_{k-1} ($k \geq 2$).

1. **Unión**. Para encontrar L_k, un conjunto de k-itemsets se genera mediante la unión de L_{k-1} consigo mismo, conjunto al que denotaremos C_k. Los miembros de L_{k-1} se pueden unir si sus primeros[46] $k-2$ elementos son comunes.

2. **Poda**. C_k es un superconjunto de L_k, es decir, todos los k-itemsets frecuentes están en C_k, aunque no todos los elementos de C_k tienen por qué ser frecuentes. Si se escanea C_k filtrando elementos cuyo soporte sea menor que el umbral s_{min}, se obtiene L_k. Sin embargo, C_k puede ser enorme, así que para mejorar la eficiencia de este paso se utiliza la propiedad Apriori. Cualquier k-itemset que contenga un (k-1)-itemset no frecuente tampoco lo será, por lo que se puede eliminar de C_k.

El algoritmo FP-growth

El algoritmo Apriori tiene dos problemas principales. El primero es que puede necesitar generar un número enorme de conjuntos candidatos. El segundo es

[46] El algoritmo Apriori asume como convención que los elementos del itemset están ordenados lexicográficamente.

el coste de analizar la base de datos para obtener el soporte para un conjunto voluminoso de candidatos.

Una forma de resolver este problema es evitar la generación de candidatos, como hace el algoritmo **frequent-pattern growth**, o FP-growth [173]. En primer lugar, FP-growth comprime la base de datos transaccional en una estructura de datos específica denominada **frequent-pattern tree**, o FP-tree, que aglutina la información de asociación de itemsets.

El primer recorrido de la base de datos es igual que en Apriori, se construye L_1, el conjunto de los 1-itemsets frecuentes. A continuación se produce el FP-tree de la siguiente manera.

1. Recorrer D una vez para obtener L_1, el conjunto de los 1-itemsets frecuentes, junto con su soporte. Ordenar L_1 en orden de soporte descendiente.
2. Crear la raíz del FP-tree, T, con etiqueta nula. Para cada transacción t en D:
   a) Seleccionar los items frecuentes en t y ordenarlos de acuerdo al orden en L_1.
   b) Sea la lista ordenada de items frecuentes en T $[p|P]$, donde p es el primer elemento y P es el resto de la lista. Se llama a *insertar-fp-tree([p|P], T)*. La función *insertar-fp-tree([p|P], T)* funciona como sigue. Si T tiene un hijo N tal que la etiqueta de N es la misma que la de p, se incrementa la cuenta de N en 1. Si no, crear un nuevo nodo N con la etiqueta de p, su cuenta inicializada a 1, su enlace al nodo padre enlazado a T, y su node-link enlazado a los nodos en el mismo item-name a través de la estructura node-link. Si P no es vacío, llamar a *insertar-fp-tree(P, T)* recursivamente.

La figura 1.13 muestra un ejemplo de estructura FP-tree obtenido de [61]. Para poder recorrer este árbol se crea para cada 1-itemset un enlace a los nodos del árbol en los que aparece, obteniendo una lista denominada *node-links*. Una vez construído el árbol FP-tree, el algoritmo FP-growth se ejecuta sobre él.

La eficiencia del método FP-growth es aproximadamente un orden de magnitud mejor que la del algoritmo Apriori [61].

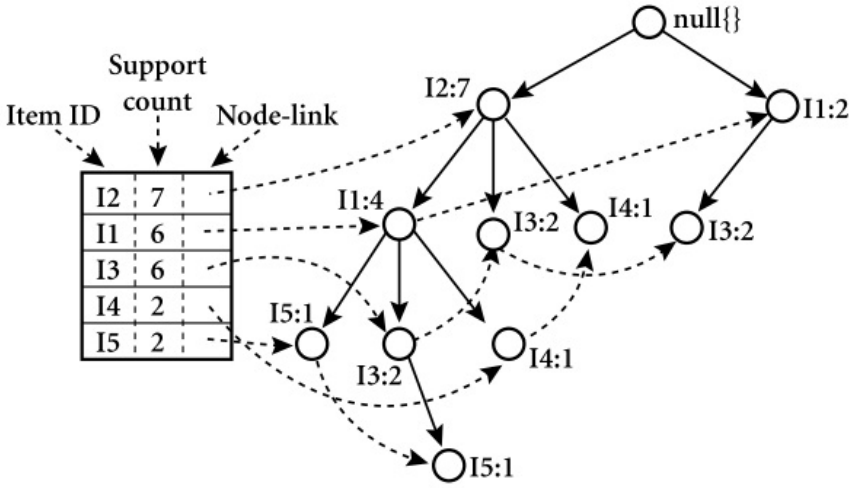


Figura 1.13: ***Ejemplo de estructura FP-tree*** *[61]. Para cada ítem en la tabla de la izquierda encontramos su etiqueta, su frecuencia absoluta y la lista node-links, que enlaza en el árbol de la derecha a todos los nodos con la misma etiqueta. En el árbol, cada nodo tiene un valor asociado su número de nodos hijos, el cual se calcula en el proceso de construcción.*

En los últimos años se han implementado versiones paralelizadas o de *cloud computing* para mejorar la eficiencia de FP-growth, tales como [174], una versión paralelizada del algoritmo FP-growth o [175], la cual utiliza mapreduce, o [176], en Spark.

Algoritmo 1: ***Algoritmo FP-growth****. Extrae itemsets frecuentes de un árbol FP-tree.*

```
1: function <FP-GROWTH>(tree, a)
2:     if Tree contiene un camino simple prefijo then
3:         P = camino prefijo simple en Tree.
4:         Q = trozo multi-camino restante, donde el nodo raíz se reemplaza por
   null.
5:         for β combinación de nodos en el camino P do
6:             generar patrón β ∪ a con soporte = min(nodos en β)
7:         end for
8:         sea set(P) el conjunto de patrones generados.
9:     else
10:        Q = Tree
11:        for a_i en Q do
12:            Generar patrón β = a_i ∪ a con soporte = soporte(a_i).
13:            Construir el árbol condicional de β y después el árbol condicional
   de FP-tree.
14:            if Treeβ ≠ ∅ then
15:                llamar fp-growth(treeβ, β)
16:            end if
17:            Sea set(Q) el conjunto de patrones generados.
18:        end for
19:    end if
20:    Devolver (set(P) ∪ set(Q) ∪ (set(Q) × set(Q))
21: end function
```

Parte II

Contribuciones

Capítulo

2

Priorización en redes biológicas heterogéneas

El desarrollo de las tecnologías de secuenciación de alto rendimiento ha propiciado la aparición de enormes cantidades de datos biomédicos disponibles para su análisis. En el camino hacia la Medicina Personalizada, la extracción de conocimiento de dichas fuentes de información mediante algoritmos integradores juega un papel clave. La información sobre entidades biológicas presente en estas fuentes de información biológica puede ser modelada en forma de grafo o red, donde cada entidad biológica se identifica con un nodo, y las relaciones entre las entidades se identifican como arcos. Estos arcos, además, pueden tener un valor numérico asociado indicando la fuerza de la relación entre las entidades biológicas conectadas. Se ha demostrado que esta representación de información biológica en forma de redes resulta extremadamente útil para integrar conocimiento proveniente de fuentes de información de diversa procedencia y de naturaleza heterogénea [53]. Mediante el análisis computacional de estas redes es posible descubrir relaciones implícitas entre las entidades que conforman el conjunto de datos, ayudando así a formular nuevas hipótesis.

Estas metodologías de análisis se pueden aplicar a redes de un sólo tipo

de entidades (*redes homogéneas*), así como a redes que conectan distintos tipos de entidades (*redes heterogéneas*). Los ámbitos de aplicación son numerosos: desde redes de interacción proteína-proteína, pasando por redes de compuestos farmacológicos, dianas terapéuticas, dominios de proteínas, genes, productos génicos o enfermedades.

Han proliferado multitud de metodologías y enfoques específicos de cada dominio, tanto para casos en los que sólo se estudian las relaciones en un tipo de entidad biológica (*e.g.* redes de interacción proteína-proteína), como para casos en los que se analizan las relaciones entre diferentes tipos de entidades, como por ejemplo las que involucran genes y enfermedades [177, 178]. Sin embargo, en la mayoría de los casos, estas metodologías que integran diversas categorías de entidades biológicas son metodologías *ad hoc* que no permiten su aplicación a otros ámbitos.

Por otro lado, un ámbito de investigación que ha dado lugar a gran cantidad de metodologías *in silico* para la priorización de entidades en redes heterogéneas es el reposicionamiento de medicamentos [179, 180]. El proceso de desarrollo de un nuevo fármaco tiene unos costes temporales y económicos enormes, de alrededor de unos 15 años y hasta 1000 millones de dólares [181], por no hablar de los riesgos en cuanto a la cantidad de estos proyectos que no acaban siendo comercializados debido a fracaso en cuanto a toxicidad o falta de efectividad. Tradicionalmente, los fármacos han sido asociados a los órganos o los síntomas que trataban. Sin embargo, hoy en día se sabe que el componente genético presente en las enfermedades es muy relevante a la hora de tratarlas, lo que permite aventurar que existan tratamientos válidos para una enfermedad que lo puedan ser para otras.

Por todas estas razones, encontrar aplicaciones noveles para medicamentos ya comercializados es un área de investigación que está tomando cada vez más protagonismo, ya que, si bien no elimina los costes de desarrollo en su totalidad, sí reduce en gran parte los costes que involucran las pruebas de seguridad, que ya fueron superadas por el medicamento para su comercialización para su aplicación previa. Además, encontrar aplicaciones nuevas para fármacos

comercializados puede aumentar la cantidad de medicamentos disponibles para tratar una enfermedad, mejorando las perspectivas para la personalización de los tratamientos, al existir un abanico de alternativas mayor, lo que redunda también en beneficio de la Medicina Personalizada.

En este capítulo se aborda el problema de la extracción de conocimiento de redes biológicas heterogéneas, en las que interactúan entidades biológicas de distintos dominios. Existen distintas metodologías en la literatura para la priorización en redes biológicas heterogéneas. Sin embargo, estas metodologías han sido desarrolladas *ad hoc* para un ámbito de aplicación concreto (por ejemplo, priorización gen-enfermedad) y la adaptación de estas metodologías para su aplicación a otros problemas resulta muy costosa, cuando no inviable. Como parte del trabajo de esta tesis se propone ProphTools[1] [182] en la sección 2.2, una metodología general de priorización en redes heterogéneas que se sustenta en una capa de abstracción que encapsula cualquier red heterogénea. Para mostrar el potencial de ProphTools, se proponen dos ámbitos de aplicación de interés: 1) la priorización de mutaciones respecto a enfermedades dando lugar a una herramienta que hemos denominado DiGSNP [89], descrita en la sección 2.2.3; y 2) el reposicionamiento de medicamentos, dando lugar a una herramienta llamada DrugNet [90], descrita en la sección 2.3. DrugNet resulta de especial relevancia por su campo de aplicación, ya que la búsqueda de nuevas aplicaciones para medicamentos ya comercializados resulta de gran relevancia en el marco de la Medicina Personalizada.

[1] http://www.github.com/cnluzon/prophtools

2.1. Metodologías existentes basadas en grafos

En los tremendamente complejos procesos biológicos suelen estar involucradas gran cantidad de entidades diferentes que interactúan simultáneamente. El modelado de dichos procesos como grafos es, en este sentido, eficaz para entender y analizar dichos procesos, lo que explica que en los últimos años hayan aparecido multitud de metodologías que utilizan este tipo de representación para inferir nuevas hipótesis a partir de conocimiento biológico existente [53].

En la actualidad, existe gran cantidad de información disponible sobre numerosas entidades biológicas a partir de la cual se puede extraer información de interacción, como por ejemplo proteína-proteína (Protein-Protein Interaction – PPI), o relaciones entre enfermedades, fenotipos, genes o medicamentos, como se ha detallado en la sección 1.2.

Dadas estas bases de datos, existen algunas metodologías que permiten construir redes a partir de la información contenida en ellas, dando como resultado un grafo (normalmente no dirigido) cuyas aristas pueden tener asociado un peso que indica la fuerza o la significancia de la relación, en base a la evidencia científica que soporta esta relación según dichas bases de datos. En general, la metodología seguida para construir la red de entidades suele formar parte del método de inferencia propuesto, como un paso adicional de preprocesado. Sin embargo, existen algunas herramientas en las cuales la construcción de la red es el objetivo en lugar de la priorización de entidades dada una red. Entre este tipo de herramientas se encuentran ARACNE [183], una metodología para construir redes de genes en base a perfiles de microarrays de expresión, y su reciente sucesora ARACNe-AP [184]. De nuevo, se trata de metodologías específicas de dominio, y encontramos diversas alternativas en el muy poblado campo de construcción de redes reguladoras de genes a partir de datos de expresión: minet [185], un paquete de R/bioconductor para inferencia de redes de transcripción a través de información mutua, GENIE3 [186], C3NET [187], RPNI [188], entre otras.

En general, el interés de estas herramientas de construcción de redes es la estructura de dichas redes, y no tanto la priorización posterior. Paradójicamente,

las herramientas de minería de datos en redes biológicas suelen estar fuertemente acopladas a la base o bases de datos que se utilizaron para construir el modelo sobre el que ejecutan las consultas. En este sentido, gran cantidad de las metodologías de preprocesamiento son *ad hoc* y no puestas a disposición del usuario final, que está limitado a utilizar la herramienta de priorización sobre unos datos prefijados.

Entre los enfoques que permiten hacer análisis automatizados de redes biológicas, como se ha mencionado brevemente en la sección 1.2, existen principalmente dos categorías: en primer lugar se encuentran aquellas que realizan los análisis sobre una única *red homogénea*, o una red de elementos del mismo tipo, en segundo lugar están las herramientas que integran dos o más categorías de elementos biológicos, a las que llamaremos *redes heterogéneas*.

Metodologías de análisis de redes homogéneas

Entre las metodologías de análisis de *redes homogéneas* podemos encontrar metodologías que realizan análisis estructural de la red o aquellas que priorizan elementos con respecto a un conjunto de nodos de entrada.

Entre las metodologías de corte **estructural** existen tres tipos principales de problemas [101]:

1. *Network querying*. Búsqueda de una subred concreta dentro de una red. Entre estas metodologías encontramos enfoques recientes, como NetMatchStar [189].
2. *Alineamiento de redes*. Extracción de subredes similares en dos o más redes biológicas con el fin de encontrar similitudes o diferencias topológicas entre especies [190]. Este es un campo de investigación muy activo, encontramos en él enfoques como [191, 192, 193, 194].
3. *Extracción de motivos en redes* [195]. Búsqueda de módulos o subredes que se repiten en una red más grande, los cuales en el ámbito biológico suelen identificarse con subsistemas que desempeñan algún tipo de funcionalidad. Entre este tipo de enfoques existen metodologías para la búsqueda de

módulos funcionales en redes de interacción proteína-proteína [196].

Por otro lado, entre las herramientas que pretenden **priorizar** al resto de entidades de la red calculando algún tipo de *distancia* entre los nodos, se incluyen metodologías específicas de área de aplicación (*i.e.* proporcionan al usuario una red ya construida, no pudiendo éste utilizar una fuente de información personalizada), como redes de interacción proteína-proteína [103] o redes de genes, como es el caso de SVD-phy, que construye redes de genes basándose en su distribución filogenética [104]. Otras metodologías, tales como DRaWR [105], aumentan las características modeladas en una red permitiendo diferentes tipos de relaciones entre los nodos (i.e. diferentes categorías de aristas). Adicionalmente, herramientas como FunRich [106] ofrecen una gran variedad de bases de datos entre las que elegir para poblar una red sobre la que efectuar análisis de enriquecimiento.

En el caso de redes homogéneas, sí existen algunos enfoques más generales que permiten trabajar sobre una matriz de adyacencia cualquiera. Este es el caso de RANKS [102], que realiza priorización utilizando funciones de *scoring* basadas en funciones kernel. Estas funciones kernel son aplicadas a la matriz de adyacencia de la red en cuestión. RANKS ofrece numerosas funciones kernel distintas y parametrizables para usar sobre la matriz de adyacencia de entrada (lineal, gaussiana, cauchy, polinómica, etc.).

Metodologías de análisis de redes heterogéneas

En cuanto al análisis de ***redes heterogéneas***, estas metodologías suelen estar más vinculadas si cabe al dominio de aplicación. Entre este tipo de herramientas encontramos las muy populares metodologías de priorización gen-enfermedad [107, 177] y proteína-enfermedad [108], las cuales, en algunos casos, incluyen redes de interacción proteína-proteína como intermediaria entre gen y enfermedad [197], o, en otros casos, como medio de relación entre los genes (*i.e.* se utiliza la información de una red PPI para construir la red de genes, quedando esta *implícita* en la red de genes). De nuevo, tanto el método de construcción de dichas redes

como las fuentes de información utilizadas forman parte del método de inferencia, estando éste fuertemente acoplado al dominio de aplicación.

Otro ámbito de aplicación muy extendido para metodologías basadas en redes es el del reposicionamiento de medicamentos. Existen multitud de enfoques computacionales con este fin, los cuales también integran las fuentes de información utilizadas de forma definitiva, no permitiendo desacoplar metodología y herramientas de forma sencilla. Entre estos enfoques encontramos [109], que construye relaciones enfermedad-medicamento, [110, 111], que buscan relaciones entre fármacos y dianas terapéuticas, o herramientas domo DeMAND [198] para la búsqueda de genes relacionados con el mecanismo de acción (Mechanism of Action – MoA) de un compuesto.

2.2. ProphTools: Priorización genérica en grafos heterogéneos

El análisis de redes de información biológica, incluyendo una única red o multitud de ellas, está muy presente en el campo de la bioinformática. Sin embargo, son pocas las herramientas que desacoplan el proceso de inferencia en sí mismo del proceso de obtención, preprocesado o revisión manual de los datos utilizados. Por otro lado, suele existir una limitación en el número de redes o tipos de nodos diferentes que pueden ser tenidos en cuenta en el proceso de análisis, siendo tres el máximo usual cuando se trata de herramientas con fuentes de información prefijadas.

A continuación se describe ProphTools [182], una herramienta *open-source* general y flexible para el análisis de priorización de cualquier tipo de red heterogénea, con un número arbitrario de dominios o tipos de nodos. La metodología de propagación se basa en un Random Walk with Restart similar al algoritmo de Flow Propagation, utilizado por ProphNet [88] y otros enfoques, pero mientras estos enfoques se centran en problemas de priorización particulares, ProphTools añade una capa de abstracción que permite analizar cualquier configuración de redes definida por el usuario.

La figura 2.1 muestra un ejemplo del tipo de problema que se pretende resolver con la herramienta. Se trata de una red heterogénea constituida por las *subredes* A, B y C, entre las cuales existen dos subredes de conexión: la que conecta A y B, y la que conecta B y C. Se trata de calcular cómo de relacionados están los nodos de entrada en la red de entrada (nodos más oscuros en A) con los nodos objetivo en la red objetivo (nodo más oscuro en C). En el caso de priorización en general, este proceso se realiza, uno por uno, para todos los nodos en la red objetivo. Los *scores* generados por este procedimiento sirven para obtener una lista ordenada de entidades de la red objetivo relacionadas con los nodos de consulta.

2.2.1. Algoritmo de propagación

La metodología de ProphTools propaga valores desde un conjunto de nodos de entrada hasta una red objetivo, de la cual se extraen los nodos más relacionados

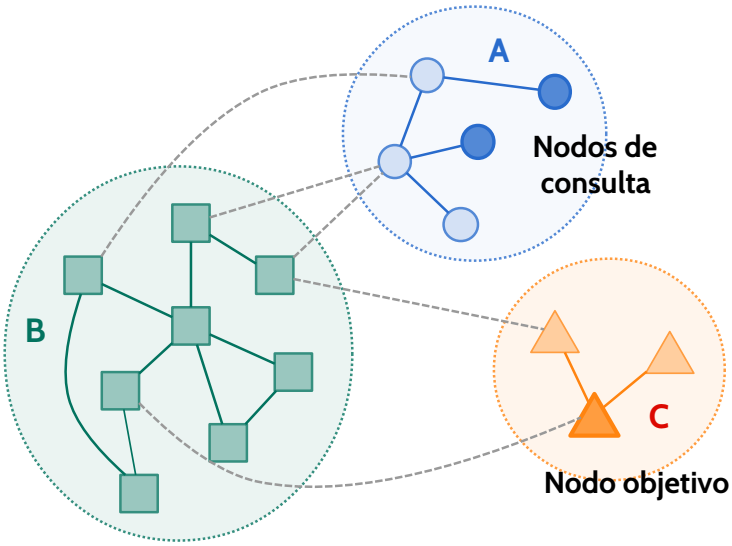


Figura 2.1: ***Priorización en redes heterogéneas****. Se trata de calcular el grado de relación entre los nodos de consulta y los nodos en la red objetivo.*

con las entidades de entrada. A continuación se describe en detalle la nomenclatura utilizada.

Se define el **grafo heterogéneo global** $G = (D,R)$, donde D es un conjunto de subgrafos o **subredes de entidad** y R es un conjunto de **subredes de relación**. Sean $D_q \in D$ la **subred de consulta**, y $D_t \in D$ la **subred objetivo**.

Sea D el conjunto de subgrafos o subredes de entidad, tal que $D_i = (V_i, E_i)$ para $i = 1, ..., n$, donde V_i es un conjunto de vértices que representa entidades biológicas de un dominio específico: $V_i \cap V_j = \emptyset \; \forall i, j$ tales que $i \neq j$. Cada nodo v_{ik} con $k = 1, ..., |V_i|$ en D_i está etiquetado con un valor correspondiente $\psi(v_{ik})$, inicialmente de valor 0, que indica el grado de relación del nodo v_{ik} con el conjunto de nodos de consulta o el de salida, dependiendo de si v_{ik} pertenece a la red de entrada o a la red de destino. Por otra parte, E_i es el conjunto de arcos del grafo ponderado no dirigido, representando relaciones, similitud o interacciones entre los elementos de V_i.

Los valores correspondientes de los nodos son modificados durante el proceso de propagación, mientras que los pesos de las aristas se mantienen constantes.

Sea R el conjunto de grafos de relación. R_{ij} es una red que une nodos de dos tipos biológicos distintos, de tal forma que $R_{ij} = (V_i \cup V_j, C_{ij})$. C_{ij} es el conjunto de aristas no dirigidas que unen elementos de V_i y V_j, tales que $i, j \in 1, ..., n$ y $i \neq j$.

El problema consiste en **calcular el grado de asociación** entre los elementos $Q \subseteq V_q$ y los elementos $T \subseteq V_t$.

Para calcular dicho grado de asociación, se establecen los valores iniciales para el conjunto de consulta: $\psi(v_{qi}) = 1 \, \forall v_{qi} \in Q$, mientras que el resto se iguala a cero: $\psi(v_{qj}) = 0 \, \forall v_{qj} \in V_q - Q$.

Se propaga entonces la información en dos etapas. Primero, se propaga desde la red de consulta hacia la red destino, haciendo uso de dos tipos de propagación: (i) *intra-red*: propaga información dentro de una de las subredes; y (ii) *entre redes*: propaga información de una red a la siguiente en la ruta global de la red de entrada a la de salida. Este proceso se realiza a través de todos los caminos posibles entre la red de entrada y la red de destino.

Para cada uno de los caminos que unen las subredes de consulta y objetivo, se propagan alternativamente los valores desde la subred de consulta, pasando por todas las subredes en el camino que va hacia la subred objetivo. Al final, se propagan los valores de los nodos dentro de la red objetivo.

Por último, la información se propaga desde los nodos objetivo al resto de nodos dentro de la red de destino. Después de estas dos propagaciones, se establece un *score* que relaciona los cojuntos de nodos de entrada y los de salida calculando la correlación entre los valores asignados a los nodos en la red de destino y los alcanzados por los nodos vecinos en las otras redes a través de la propagación desde la red de origen. El uso de la correlación con este fin ha dado buenos resultados en métodos anteriores [199, 107].

Además, como lo que buscamos es una lista priorizada de las entidades de la red objetivo frente al conjunto de nodos de consulta Q, el procedimiento de calcular esta correlación se realiza para todos los elementos $e \in D_t$, como se puede ver en el método general de propagación que se detalla en el algoritmo 2, utilizado por ProphNet [88].

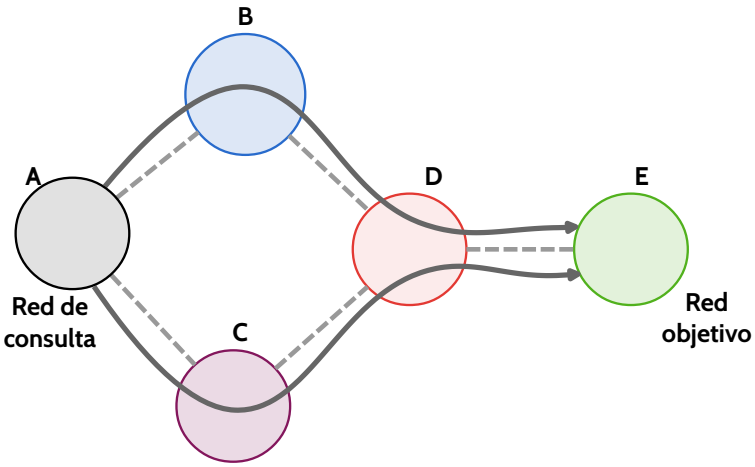


Figura 2.2: *La propagación en dos etapas intra-red y entre redes se realiza a través de todos los super-caminos que unen la subred de consulta a la subred objetivo. En la figura se muestra una red heterogénea con cinco subredes: A, B, C, D y E. La propagación de A a E cuenta con dos caminos que se señalan con flechas continuas.*

A continuación se explican de forma general los procesos de propagación *intra-red*, entre redes y la correlación entre los valores propagados desde la red de consulta y desde la red de destino.

2.2.1.1. Propagación *intra-red*

Existen numerosas metodologías para explorar el grado de relación entre nodos de un mismo grafo [200]. Los métodos que se centran en el vecindario local suelen perder de vista la estructura global[2] de la red [201]. El método de propagación *intra-red* de ProphNet utiliza la información estructural de la totalidad de la red para realizar una propagación de los valores en los nodos de la misma. La propagación *intra-red* se basa en el algoritmo de Flow Propagation [202, 203], que utiliza la matriz laplaciana normalizada para propagar los valores ψ dentro

[2]Nótese que aquí *global* se refiere a la estructura conjunta de una *subred*, no a la red heterogénea total.

Algoritmo 2: ***Algoritmo de propagación****. Prioriza desde la subred de consulta D_q a la subred de destino D_t. G es el grafo global y Q el conjunto de consulta.*

function <PROPAGACIÓN>(G, Q, D_q, D_t)
    Propagación *intra-red* en D_q
    P: lista de caminos de D_q a D_t en G
    for p_i en P **do**
        for subred p_{ij} en el camino p_i de p_{i1} a $p_{i(l-1)}$ **do**
            Propaga valores de p_{ij} a $p_{i(j+1)}$
            Propaga valores dentro de $p_{i(j+1)}$
        end for
        Almacena los valores de $D_{i(l-1)}$ tras la propagación en el camino p_i como $\hat{x}_{i(l-1)}$
    end for
    for $e \in V_t$ en la red objetivo D_t **do**
        Establecer conjunto objetivo $T = e$
        Propagar valores dentro de D_t
        Calcular coeficiente de correlación s_e usando $\hat{x}_{i(l-1)}$ para cada p_i
    end for
    L: Ordena todas las entidades $e \in V$ por sus valores s_e en orden descendente.
    Devolver L
end function

de la red. La normalización toma en consideración el grado de cada nodo para limitar el sesgo hacia nodos muy conectados. Esta normalización también es crítica para la convergencia del algoritmo. Así, el método de propagación es similar a un Random Walk with Restarts, difiriendo en el proceso de normalización de la matriz de adyacencia que guía la propagación [203].

La normalización tiene en cuenta el grado de cada nodo para limitar el impacto de los nodos ricos o *hub* en el proceso de priorización. Cada matriz de adyacencia A es normalizada de la siguiente forma:

$$A_{norm} = D_G^1 * A * D_G^2$$

donde D_G^1 y D_G^2 son matrices diagonales en las que cada componente se calcula como:

$$D^1_{G_{jj}} = \frac{1}{\sqrt{\sum_{j=1}^{c} A_{jk}}}, j = 1, ..., r$$

$$D^2_{G_{kk}} = \frac{1}{\sqrt{\sum_{j=1}^{r} A_{jk}}}, j = 1, ..., c$$

Todas las subredes son normalizadas mediante este proceso antes de iniciar la propagación. Para realizar la propagación dentro de cada subred normalizada D_k, primero se define la información previa Z como aquellos vértices v_{kj} tales que $\psi(v_{kj}) \neq 0$. $Z = Q$ cuando la subred D_k es la subred de consulta. El valor de $\psi(v_{kj})$ en Z, donde $j \in 1, ..., |V_k|$ es entonces:

$$\psi(v_{kj}) = \frac{\psi(v_{kj})}{\sum^{v_{kx} \in V_k} \psi v_{kx}}$$

Sea x_0 un vector con los valores asignados inicialmente a cada nodo en D_k, y $\hat{x}$ un vector con los valores asignados a cada nodo en D_k después de realizar la propagación dentro de D_k. Para calcular $\hat{x}$ es necesario resolver el siguiente problema de optimización:

$$min_{\hat{x}} \sum_{i,j} M_{k_{i,j}} (\hat{x}_i - \hat{x}_j)^2 + \frac{1-\alpha}{\alpha} \sum_i (\hat{x}_i - \hat{x}_{0_i})^2$$

donde M_k es la red de adyacencia normalizada de D_k. La solución a esta ecuación es:

$$\hat{x} = (1-\alpha)(I - \alpha M_k)^{-1} x_0$$

Este sistema lineal se puede resolver de forma exacta. No obstante, existe un algoritmo iterativo más eficiente para resolverlo [204]:

$$x_{i+1} = \alpha M_k x_i + (1-\alpha) x_0, \ \textit{i empezando en 0}$$

Este algoritmo implementa un proceso iterativo donde cada nodo propaga hacia sus vecinos basados en el peso del arco que los conecta. El parámetro α define la importancia que se le da a la información anterior Z. El criterio de parada para el proceso iterativo se establece en $|x_i - x_{i+1}| < \kappa$, con $\kappa = 10^{-3}$.

2.2.1.2. Propagación entre redes

Dada una red D_i cuyos valores ya han sido asignados su valor $\hat{x}_{il}$, estos valores se propagan a la red D_j en el camino actual p_l, con $D_j \neq D_t$. Esta información se propaga desde D_i a los nodos en D_j a través de los arcos en R_{ij} que los conectan. Nótese que las redes bipartitas que conectan las subredes también son normalizadas mediante el mismo método descrito en el apartado anterior. Se le asigna pues a cada nodo $v_{jk} \in D_j$ el valor medio de los nodos en D_i que están conectados con v_{jk}:

$$\psi(v_{jk}) = \frac{\sum^{v_{ix} \in neig_i(v_{jk})} \psi(v_{ix})}{|neig(v_{jk})|}$$

donde $neig_i(v_{jk})$ es el conjunto de los nodos en D_i que están conectados a v_{jk} de acuerdo a R_{ij}. Adicionalmente, se añade un umbral de corte en este paso para eliminar ruido. Un parámetro adicional γ es introducido tal que los nodos con valores $\psi(v_{jk}) < \lceil |V_j|(1-\gamma) \rceil$ se igualan a 0.

2.2.1.3. Correlación entre valores propagados

Después del proceso de propagación desde D_q por uno de los caminos p que conecta D_q con D_t, los nodos en las subredes que son adyacentes a D_t contienen valores que determinan el grado de relación con el conjunto objetivo. Los nodos en la subred objetivo también tienen un valor asignado que representa el grado de relación con el conjunto T. Podemos medir la relación indirecta entre Q y T calculando la similaridad entre los valores de los nodos en la subred objetivo D_t y los valores de los nodos que están conectados a ellos en las redes adyacentes. Esto se puede realizar calculando la correlación de forma simultanea con los valores obtenidos para estos nodos a través de todos los caminos p_i que conectan D_q con D_t. Para cada camino p_i de longitud l se calcula un vector $\bar{x} = S_a \hat{x}_{(l-1)i}$, donde S_a es la matriz de adyacencia normalizada para $R_{(l-1)l}$ y $\hat{x}_{(l-1)i}$ es el vector obtenido después de propagar los valores dentro de la red D_{l-1}.

Definimos el vector $\bar{t}$ como el resultado de concatenar los valores $\hat{x}_t$ como $\bar{t} = repeat(\hat{x}_t, |P|)$, donde $repeat(v, n)$ es una función que repite el vector v n veces.

Definimos también el vector $\bar{x}$ como $\bar{x} = concat(\bar{x}_i)\ \forall i \in 1, |P|$ donde $concat$ significa concatenación de vectores. $\bar{x}$ y $\bar{t}$ tienen el mismo tamaño.

Finalmente, el valor de correlación s se define como $s = corr(\bar{x}, \bar{t})$, donde $corr$ es el coeficiente de correlación de Pearson.

Este proceso de priorización descrito corresponde a una consulta para un conjunto de nodos de entrada Q. No obstante, sea cual sea Q, en el paso final de correlación siempre es necesario calcular la correlación de los nodos dentro de la red de destino, lo cual es una operación computacionalmente costosa que además no depende de la consulta. Por ello, con el fin de acelerar el proceso de cálculo, se precalcula esta propagación intra-red para cada nodo de las redes de entidad involcuradas y se almacena en una matriz de adyacencia precalculada. Este cálculo está incluido dentro de las funciones de preprocesamiento que incluye ProphTools.

2.2.1.4. Extensibilidad del método de propagación

ProphTools no sólo es una herramienta de priorización genérica que permite la integración de diversos tipos de entidades de forma sencilla. Además, es una herramienta modular que permite la inclusión de nuevos enfoques de priorización intra-red y entre redes con el fin de adaptarse mejor a cada tipo de problema.

Esta capacidad se ha podido comprobar ampliando la funcionalidad de la herramienta propuesta. En este sentido, se ha integrado RANKS, la metodología de propagación propuesta por Valentini *et al.* [102] como método de priorización intra-red alternativo.

RANKS es una metodología de análisis de redes. Este método toma como entrada una matriz de adyacencia y la transforma mediante una función kernel. Esta función transforma la matriz de adyacencia de forma que la similitud entre todos los nodos es calculada teniendo en cuenta la estructura global de la red [205]. Dependiendo de las características a tener en cuenta mediante la función kernel, la implementación de RANKS cuenta con los kernels identidad, lineal,

gaussiano, inverso multicuadrático, polinomial, laplaciano y de Cauchy, junto con el kernel Random Walk para un número n de pasos.

Una vez este kernel es calculado, el *score* entre los nodos del conjunto de nodos de entrada $V_C \subset V$ y un nodo cualquiera $v_i \in V$ de la red puede ser calculado de diferentes formas:

- **Media**. Podemos definir el *score* de un nodo i como la media de similitud entre dicho nodo y los nodos del conjunto de entrada V_C.
- **Vecino más próximo (nearest-neighbor).** Si consideramos la distancia mínima entre i y V_C, obtendremos el score de vecino más próximo.
- **K-vecinos más próximos**. La similitud en este caso es la suma de los k vecinos más próximos.

Tanto el cálculo de los kernels como las funciones de scoring de RANKS han sido integradas dentro de la metodología de ProphTools, pudiendo el usuario elegir cómo desea priorizar en su configuración de red heterogénea. Nótese que esta metodología se ha incluido en la etapa de propagación *intra-red*, por lo que la metodología global resultante constituye en sí misma una metodología novedosa de propagación.

Adicionalmente, el cálculo de los kernels se ha incluido en la parte de preprocesamiento que también incluye el cómputo de las matrices precalculadas de ProphTools, con el fin de desacoplar este proceso de los cálculos de priorización. De este modo se reduce el coste de aplicación cuando se desean realizar múltiples consultas a la red. No obstante, el almacenamiento de las matrices kernel puede ser más voluminoso que el de las matrices precalculadas de ProphTools, ya que estas matrices son densas, al contrario de las matrices dispersas utilizadas en el algoritmo de Random Walks original.

2.2.2. Modelo de representación de redes heterogéneas

ProphTools permite la priorización en cualquier configuración de redes heterogéneas mediante un modelo de representación de dicha configuración que incluye una matriz de adyacencia para cada una de las redes de entidades

involucradas, una matriz de adyacencia para cada una de las redes bipartitas que conectan nodos de distintos dominios y una matriz de súper-adyacencia que conecta entre sí los diferentes dominios. A esta información es posible añadir como información adicional etiquetas a los nodos de cada dominio, que facilitarán la comprensión de los resultados.

Dada esta configuración, ProphTools puede ejecutar las metodologías de priorización descritas. Para un conjunto de nodos en una red de entrada, ProphTools devolverá una lista de nodos de la red destino puntuados acorde a su relación con el conjunto de nodos de entrada. Adicionalmente es posible realizar tests leave-one-out (LOO tests) y LOO con validación cruzada para comprobar la capacidad predictiva de la configuración de entrada proporcionada.

Así, ProphTools permite priorizar sobre cualquier conjunto de datos que se le proporcione como entrada. Para ello se define un meta modelo que especifica las subredes que existen en el sistema, las conexiones entre las mismas así como las etiquetas de los nodos de cada red. La figura 2.3 muestra un pequeño ejemplo para una configuración de tres subredes, A, B y C. Además de las matrices de adyacencia de dichas subredes, así como sus correspondientes matrices precalculadas, se incluyen matrices de adyacencia que representan a las subredes que las conectan. Se añade además una matriz de adyacencia a un nivel superior, que indica qué subredes son conectadas por una matriz de adyacencia bipartita y en qué dirección, dado que estas matrices no son simétricas, es decir, si llamamos a la matriz de super-adyacencia S, $S[A,B] = AB$ si las filas de AB representan las entidades de A y las columnas de AB representan las entidades de B. Esto permite garantizar la integridad del modelo.

Además, el formato elegido para incluir esta información utiliza ficheros de tipo `mat` análogos a los generados en MATLAB y Octave, que pueden ser manipulados con la biblioteca libre `scipy.io`, garantizando compatibilidad con otras herramientas con las que se hayan generado las redes de análisis. ProphTools está disponible como paquete de python y puede ser instalado fácilmente a través de pip[3], además de estar disponible su código abierto con licencia GPLv3[4] en

[3] Sistema de instalación de paquetes estándar de python.
[4] GNU Public License v3. https://www.gnu.org/licenses/gpl-3.0.txt

GitHub[5]. ProphTools ha sido desarrollado aplicando Unit Testing y metodologías de integración continua para garantizar su correcta instalación en sistemas Linux. Los requisitos de dicho paquete son muy reducidos a nivel de bibliotecas utilizadas, ya que todas son libres y de acceso gratuito (scipy, numpy, matplotlib entre otras). Además tampoco tiene altos requisitos hardware para funcionar, aunque las exigencias de memoria pueden aumentar con el tamaño de las redes integradas.

2.2.3. DiGSNP: Priorización de mutaciones reguladoras relacionadas con enfermedades

La comprensión de las causas genéticas de las enfermedades es una de los principales objetivos para avanzar hacia la meta de una efectiva Medicina Personalizada. Metodologías como los estudios GWAS o los estudios de expresión con microarrays o metodologías más recientes como RNA-seq permiten obtener evidencia experimental de regiones del genoma relacionadas con enfermedades. Sin embargo, estas tecnologías típicamente devuelven una gran cantidad de resultados asociados con las condiciones bajo estudio.

En este contexto son necesarias metodologías que asistan al investigador, de nuevo, reduciendo el espacio de búsqueda a un conjunto asequible de candidatos de interés. En este sentido y como se ha mencionado con anterioridad, han proliferado las metodologías de priorización gen-enfermedad [177]. Sin embargo, la mayoría de estas herramientas no consideran mutaciones, aunque éstas son conocida causa de gran cantidad de enfermedades, tanto aquellos que se encuentran en zonas codificantes y provocan cambios en la secuencia de aminoácidos [206], como los SNPs en regiones no codificantes [207].

Enfoques como AnnTools [208] o SNPRank [209], basándose en estudios GWAS, relacionan variaciones (SNPs) con enfermedades, pero no profundizan en los mecanismos concretos por los que la enfermedad en cuestión se relaciona con las variantes obtenidas. Adicionalmente, la mayoría de las herramientas disponibles se centran en las regiones codificantes, ignorando las posibles regiones

[5] http://www.github.com/cnluzon/prophtools

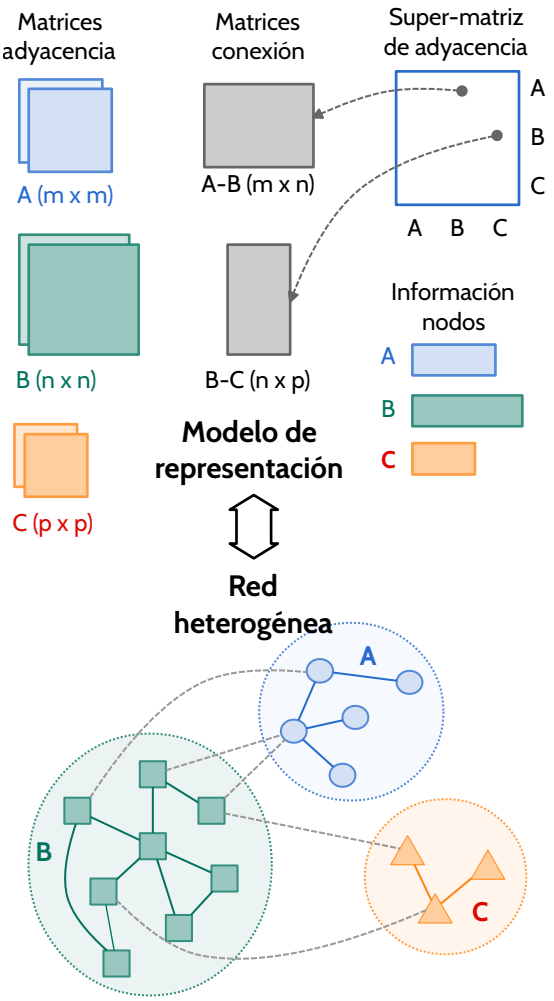


Figura 2.3: ***Modelo de representación de redes heterogéneas de ProphTools****. En la parte inferior del esquema se observa la red heterogénea representada en la parte superior. Consta de de tres subredes, A, B y C, entre las cuales existen dos conexiones, AB y BC. Para A, B y C se guardan matrices de adyacencia, sus precalculadas e información sobre cada nodo. La información sobre qué subredes conecta una subred de conexión se encuentra en la matriz de super adyacencia. Este modelo de representación desacopla el algoritmo de propagación de los datos de entrada, que pueden ser por tanto modificados según el ámbito de aplicación sin afectar al algoritmo de priorización.*

reguladoras presentes en secciones no codificantes del genoma, como por ejemplo, las zonas promotoras de los genes. Como se ha introducido en el capítulo 1, un SNP puede alterar la afinidad de unión de un lugar de unión de factores de transcripción (TFBS), generando una disrupción en la regulación de un gen. Esta variación reguladora, puede, en última estancia, derivar en un malfuncionamiento de la célula, que a su vez provoque un estado de enfermedad en el individuo.

En esta sección se presenta DiGSNP (Disease-Gene-SNP Prioritizer) [89], una herramienta que permite relacionar enfermedades, genes y variaciones en regiones reguladoras del genoma (en particular, aquellas que afectan a TFBSs), de forma simultánea, facilitando la investigación en profundidad de los mecanismos reguladores que provocan algunas enfermedades.

La priorización realizada por DiGSNP funciona en dos niveles. Dada una enfermedad o enfermedades de consulta, DiGSNP obtiene una lista priorizada de genes candidatos a estar relacionados con dicha enfermedad. El segundo nivel busca en las zonas promotoras de estos genes una lista de SNPs potencialmente reguladores dada su presencia en un TFBS. Así, DiGSNP construye una jerarquía enfermedad-gen-SNP que permite vincular SNPs reguladores, priorizados en base a su nivel de influencia en un TFBS putativo, a enfermedades.

La priorización gen-enfermedad es realizada utilizando ProphNet [88], una configuración específica de ProphTools que realiza priorización gen-enfermedad utilizando tres redes: una de fenotipos, otra de dominios de proteínas y otra de genes. La red de fenotipos y la fenotipo-gen están extraidas de OMIM utilizando técnicas de text mining, la red de genes se extrajo de HPRD[6] y la red de dominios de proteínas fue extraida de DOMINE e InterDom[7]. Las redes gen-dominio y dominio-enfermedad fueron extraidas de Pfam[8].

Después de obtener una lista priorizada de genes relacionados con la enfermedad de consulta, cada gen es asociado con una lista de SNPs presentes en sus regiones promotoras. Los SNPs han sido obtenidos de dbSNP[9] [148]. La

[6] Human Protein Reference Database. http://www.hprd.org
[7] http://interdom.i2r.a-star.edu.sg/
[8] http://pfam.sanger.ac.uk/
[9] http://www.ncbi.nlm.nih.gov/projects/SNP

priorización gen-SNP se realiza después basándonos en dos criterios. Primero, los SNPs que se encuentran en un TFBSs son SNPs reguladores candidatos. Segundo, un SNP que causa cambios drásticos en la afinidad de unión de dicho TFBS tiene una mayor probabilidad de afectar a la regulación del gen en cuestión y por tanto también de ser un SNP regulador relacionado con la enfermedad de consulta.

Con la finalidad de averiguar si un SNP está afectando a un TFBS, se aplica IntuitSNP, una metodología computacional que permite predecir el potencial regulador de un SNP basándose en la afinidad de unión de la secuencia en cada alelo. IntuitSNP será descrita en mayor detalle en el capítulo 3. Básicamente, utilizando la medida difusa de afinidad secuencia-motivo SC_{intuit}, se buscan los motivos de las bases de datos JASPAR y TRANSFAC que tienen una afinidad de unión por encima de un umbral de corte de similitud. Los SNPs seleccionados por este primer filtro son ordenados en base a la diferencia de *scores* para los diferentes alelos del SNP, obteniendo una puntuación superior aquellos SNPs cuyos alelos generan una diferencia de afinidad de unión mayor.

La tabla 2.1 muestra los resultados obtenidos para la enfermedad de Alzheimer. Esta disposición jerárquica de los resultados permite al usuario final observar relaciones entre genes, mutaciones y enfermedades. Por ejemplo, la aparición frecuente del motivo Kid3 revela una relación directa entre Kid3 y Alzheimer respaldada por la literatura [210].

La metodología propuesta puede ser aplicada a cualquier método de priorización que pueda puntuar genes relacionados con enfermedadas y SNPs relacionados con genes. Sin embargo, debido a la falta de fuentes de información relacionando mutaciones reguladoras con enfermedades, la validación de este planteamiento es compleja. Las relaciones entre el Alzheimer y genes que aparecen en la tabla 2.1 tales como APP, TREM1 y TREM2 se pueden encontrar en la literatura [211]. Sin embargo, encontrar información específica sobre los SNPs encontrados no fue posible. La razón principal es probablemente el interés de la investigación más centrado en los SNPs codificantes, buscando mutaciones no sinónimas.

El principal interés del presente trabajo es la regulación de los genes, área para la cual hay una menor cantidad de información que en el estudio de regiones

Tabla 2.1: ***Resultados de DiGSNP para Alzheimer.*** *Los cinco mejores resultados de genes para DiGSNP cuando la enfermedad de entrada es Alzheimer, para cada uno de ellos los tres mejores resultados (i.e. los SNPs funcionales más probables debido a su efecto en la afinidad de unión de un TFBS candidato). Cada SNP también se asocia con la región del gen donde se sitúa y con el ID del motivo del TFBS encontrado.*

Enfermedad	Gen	SNP ID	IntuitSNP	Región gen	Motivo
Alzheimer	APP	rs199610454	0.32	5'UTR	Kid3
		rs201528959	0.27	5'UTR	Churchill
		rs200990709	0.25	5'UTR	HNF4
	PSEN2	rs200123803	0.34	5'near gene	ZNF354C
		rs200034334	0.32	5'UTR	Kid3
		rs150618255	0.29	5'near gene	Kid3
	PSEN1	rs202004275	0.32	5'UTR	Kid3
		rs201506908	0.29	5'UTR	Kid3
		rs200531676	0.29	5'UTR	MAFB
	TREM2	rs113167129	0.34	5'near gene	ZNF333
		rs187797067	0.34	5'near gene	C-MAF
		rs138222305	0.32	5'near gene	Kid3
	HD	rs192838728	0.32	5'near gene	Kid3
		rs28616835	0.32	5'near gene	Kid3
		rs398691	0.32	5'near gene	Kid3

codificantes de los genes. Sin embargo, creemos que la metodología propuesta es una prueba del potencial de las metodologías generales de integración entre fuentes de datos biológicos heterogéneos. Este tipo de enfoques nos permiten obtener conocimiento que relacione diferentes entidades, las cuales forman todas parte en el complejo proceso de la regulación.

2.3. DrugNet: Reposicionamiento de medicamentos

El desarrollo de un nuevo medicamento es un proceso largo, costoso y de elevado riesgo, de alrededor de 15 años de duración y con una inversión aproximada de entre 800 y 1000 millones de dólares [181]. Por otro lado, una de las mayores causas de fracaso en el desarrollo de un nuevo fármaco son los controles de seguridad y toxicidad [212]. Debido a estas causas, el campo conocido como **reposicionamiento de medicamentos** ha experimentado un auge en los últimos años. Se entiende como reposicionamiento de medicamentos el proceso de búsqueda de nuevas aplicaciones para fármacos ya comercializados. El relanzamiento de un fármaco ya existente es además unas cinco veces más económico que una reformulación de su principio activo [212]. Por otro lado, existen medicamentos que fueron descartados por falta de eficacia para la dolencia para la que fueron diseñados, pese a que su consumo se ha demostrado seguro al pasar los controles de toxicidad antes mencionados [213].

A pesar de que un reposicionamiento exitoso de un medicamento no elimina la totalidad del coste de una nueva comercialización, ya que son necesarios nuevos ensayos clínicos para garantizar la efectividad del principio activo para su aplicación alternativa, sí se reducen drásticamente el coste y el riesgo de dichos procesos. Las metodologías *in silico* de reposicionamiento de fármacos permiten reducir el espacio de búsqueda de decenas de miles de compuestos a un conjunto asequible de candidatos probables en el que centrar la investigación.

Casos tan sonados como el del compuesto sildenafilo, conocido por su marca comercial, Viagra, inicialmente comercializado para el tratamiento de la disfunción eréctil y posteriormente reposicionado con éxito para el tratamiento de enfermedades coronarias, o el del minoxidil, que pasó de recetarse a pacientes con elevada presión sanguínea a su utilización en el tratamiento de la alopecia androgénica, son ejemplos ya clásicos de lo que se puede conseguir con el reposicionamiento de medicaméntos. Con estos antecedentes, no es de extrañar que en 2009 casi un tercio de los fármacos aprobados fueran reposicionamientos de medicamentos previamente comercializados [214, 215].

Debido a estas razones, el reposicionamiento computacional de medicamentos ha atraído gran atención en los últimos años [181]. El conjunto de los métodos existentes con este propósito se divide principalmente en dos grupos con diferentes perspectivas: mientras los enfoques pertenecientes al primer grupo buscan nuevas aplicaciones para un fármaco a través de su composición química, las posibles dianas terapéuticas de dicho compuesto o cualquier tipo de característica química o molecular del mismo, los segundos buscan nuevas aplicaciones empleando el conocimiento existente sobre la sintomatología y los procesos subyacentes a las enfermedades tratadas por el fármaco a reposicionar. Dicho de otro modo, los primeros enfoques se centran en la **composición** química del medicamento, y los segundos en las **enfermedades** tratadas por el mismo.

Entre las herramientas de reposicionamiento basadas en la composición del medicamento, encontramos numerosas herramientas que relacionan compuestos químicos y dianas terapéuticas a través de propiedades cuantitativas compartidas por ambas entidades [110]. Los perfiles de expresión génica son ampliamente utilizados en esta área [216, 109, 217, 218, 219].

Otra forma de buscar aplicaciones para un compuesto químico es simular su interacción física directa con las proteínas en base a su estructura tridimensional, proceso que es conocido como *docking* [220].

Muchas de estas aplicaciones se basan en interacciones individuales compuesto-diana. Sin embargo, la eficacia de muchos medicamentos reside en su capacidad para interactuar con múltiples dianas al mismo tiempo. El campo conocido como la polifarmacología [221] intenta encontrar relaciones entre compuestos químicos y múltiples dianas terapéuticas.

Otro enfoque muy diferente lo representan aquellas metodologías que pretenden encontrar aplicaciones nuevas para los medicamentos relacionando conocimiento sobre la sintomatología de las enfermedades que tratan, su patología o tratamientos conocidos. Ejemplos de estos enfoques los encontramos en Promiscuous [222] y la propuesta de Agarwal y Yang [223], herramientas que utilizan el conocimiento existente sobre efectos secundarios, ya que éstos suelen estar relacionados en cierto modo con la sintomatología de las dolencias

que tratan. Métodos como PREDICT [224] buscan asociaciones medicamento-medicamento y medicamento-enfermedad mediante técnicas de text mining.

Por otro lado, Chiang *et al.* [225] utilizan el principio de *culpa por asociación*[10]. Este principio establece que la relación entre entidades biológicas que muestran un comportamiento similar o que comparten propiedades es más probable. Este planteamiento no es nuevo en la biología computacional. De hecho, ha sido utilizado para relacionar una amplia diversidad de entidades biológicas, como genes, proteínas, fármacos o enfermedades [226, 227, 228]. En su enfoque aplicado a los fármacos y las enfermedades, Chiang *et al.* consideran que dos enfermedades presentarán rasgos parecidos si se les aplica el mismo tratamiento.

Gran cantidad de estos planteamientos se basan en similaridad y transitividad, conceptos que derivan estructuras en forma de red. Se entiende en general que la compleja naturaleza de la información biológica es intrincada y tiende a adoptar esta estructura de red o grafo, aunque no está claro si alguno de los enfoques anteriormente mencionados es sustancialmente mejor que el resto [229].

Por una parte, los enfoques moleculares son potentes, pero requieren un conocimiento preciso sobre las moléculas que se estudian. La base química de muchas enfermedades y las dianas terapéuticas de muchos fármacos no están completamente establecidas. Las metodologías que aplican *docking* necesitan que la estructura tridimensional de las moléculas participantes en el estudio esté bien establecida y disponible, lo cual no siempre es el caso.

Además, muchas enfermedades afectan a gran cantidad de tejidos y órganos, dificultando su estudio mediante perfiles individuales de actividad molecular. Por último, hay fenotipos que no son fácilmente predecibles basándonos únicamente en características químicas.

Los enfoques relativos a la sintomatología de las enfermedades complementan muy bien a este tipo de enfoques químicos, ya que permiten asimilar cierto desconocimiento acerca de la base molecular que las provoca siempre y cuando exista información clínica sobre pacientes afectados por la dolencia de búsqueda. Estas metodologías son por ello muy utilizadas para estudiar nuevas enfermedades,

[10] *Guilt-by-association* (GBA).

las cuales aún no han sido estudiadas, o enfermedades raras, sobre las que en general sólo existe conocimiento sobre sus síntomas.

Pese a las diferencias existentes entre los enfoques planteados, parece evidente que las herramientas basadas en grafos son idóneas para el análisis de bases de datos biológicas [230]. Este hecho se refleja también en el ámbito del reposicionamiento de medicamentos, como demuestra la existencia de numerosas herramientas basadas en grafos para este fin [110, 109, 217, 231, 111]. Sin embargo, pese a que estos enfoques sugieren nuevas relaciones fármaco-enfermedad, presentan limitaciones cuando se aplican a enfermedades multifactoriales complejas que dependen de múltiples dianas [232]. Además, hasta la fecha no conocemos la existencia de enfoques que tengan en cuenta relaciones entre enfermedades para el reposicionamiento de medicamentos (no se construyen redes enfermedad-enfermedad), a pesar de que se ha demostrado que este tipo de relaciones son clave para la comprensión de dolencias complejas [233].

En esta sección se presenta una herramienta de priorización en redes heterogéneas aplicada al reposicionamiento de medicamentos: DrugNet [90]. DrugNet integra simultáneamente información heterogénea sobre dianas terapéuticas (proteínas), fármacos y enfermedades. El enfoque de DrugNet se basa en ProphTools y la metodología de propagación descrita en la sección 2.2, extendiendo dicha propagación a otros conjuntos de redes: uno constituido por redes medicamento-medicamento y enfermedad-enfermedad y otro que incluye una red de proteínas como intermediaria entre ambas redes.

La validación de esta metodología se ha llevado a cabo mediante la ejecución de tests automatizados y validación cruzada, casos de estudio reales y un estudio comparativo con una herramienta referencia en la materia: PREDICT [224].

2.3.1. Construcción de las redes

Los datos para la red de medicamentos fueron obtenidos de DrugBank 3.0 [234]. La similaridad entre cada par de medicamentos se calculó utilizando la similitud de Lin basada en nodos [235] en las anotaciones para los medicamentos.

Los códigos ATC[11] son un sistema de clasificación para fármacos controlado por el WHOCC[12]. Este sistema divide los medicamentos en grupos diferentes de acuerdo al órgano o sistema en el que actúan o en sus características químicas y terapéuticas. Basándose en la ontología implícita definida por los códigos ATC y dado que un fármaco puede tener múltiples codigos ATC, la similaridad semántica entre un fármaco i y un fármaco j se calcula como:

$$similarity(i,j) = \frac{\sum_{k=1}^{|C^i|} Lin_{dist}(c_k^i, c_z^j)}{|C^i|}, \quad C^i = c_1^i, ...c_{|C^i|}^i$$

donde

$$Lin_{dist}(u,v) = 1 - Lin_{sim}(u,v)$$

siendo Lin_{sim} la similitud semántica de Lin y $z = argmin_{x \in C^j} Lin_{dist}(c_k^i, x)$.

Sólo fueron considerados medicamentos aprobados con al menos un código ATC, conteniendo la red de fármacos resultante 1490 nodos. La red de enfermedades contiene 4517 enfermedades y fue construida utilizando Disease Ontology [92]. Para eliminar redundancia entre nodos de enfermedad, sólo se consideraron nodos hoja. La similaridad entre las enfermedades i y j es calculada como $Lin_{sim}(i,j)$.

La red de proteínas fue extraida de BioGrid 3.2 [236], de la que se obtuvieron 18107 nodos con 136867 interacciones.

La red enfermedad-fármaco se deriva del campo de indicaciones de DrugBank. Una enfermedad se asocia con un fármaco si su nombre está contenido en el campo de indicaciones para el fármaco. Todas las variantes del nombre así como las permutaciones de los tokens que conforman el nombre fueron considerados, obteniendo un total de 1008 relaciones.

Las relaciones fármaco-proteína se extrajeron vinculando símbolos de proteínas y sinónimos de la red de proteínas con las dianas terapéuticas anotadas en DrugBank, obteniendo 4026 relaciones.

[11] *Anatomical, Therapeutic, Chemical*
[12] WHO Collaborating Centre for Drug Statistics Methodology.

Las asociaciones enfermedad-proteína fueron extraidas directamente de DGA[13] [237], obteniendo 11658 relaciones.

2.3.2. Validación del método

DrugNet ha sido probado con dos configuraciones de redes. La primera considerando únicamente fármacos y enfermedades, y la segunda considerando redes de fármacos, enfermedades y proteínas (dianas terapéuticas de los fármacos[14]).

Esta metodología ha sido validada mediante la ejecución de una prueba de validación cruzada con LOO y varios tests LOO: uno clásico y un segundo más estricto que elimina aristas en el vecindario directo para evitar la influencia del mismo en el resultado. Finalmente estudiamos las predicciones de DrugNet para la mejor de las configuraciones con usos de fármacos de descubrimiento reciente (obtenidos de la literatura y de ensayos clínicos) no presentes en nuestros datos.

Las validaciones LOO se ejecutaron para determinar la eficacia de DrugNet en dos configuraciones de redes. Para 1008 tests (uno para cada relación fármaco-enfermedad explícita presente en la red global), dos tipos de LOO fueron considerados: LOO estándar y LOO avanzado. El LOO estándar se ejecutó para cada caso de prueba eliminando el arco fármaco-enfermedad correspondiente y realizando la priorización fármaco-enfermedad. El LOO avanzado no sólo elimina el arco directo fármaco-enfermedad del caso de prueba, sino también todas las relaciones directas del fármaco del caso de prueba a cualquiera de las enfermedades en la red de enfermedades y de la enfermedad del caso de prueba a cualquiera de los demás fármacos. La finalidad del LOO avanzado es ser más estrictos para evitar un posible sesgo por la existencia de comunidades de enfermedades y fármacos relacionados.

[13] Disease and Gene Annotations.

[14] En este trabajo nos referimos con dianas terapéuticas a las proteínas que interactúan con los compuestos de los fármacos presentes en nuestra red, a pesar de que también pueden ser dianas terapéuticas otras entidades sobre las que pueden actuar fármacos, tales como los ácidos nucléicos o los lípidos.

Para cada test LOO se han trazado curvas ROC, representando la proporción de verdaderos positivos entre los positivos contra la proporción de falsos positivos entre los negativos para varios umbrales de corte. Un verdadero positivo ocurre cuando el rank de la enfermedad objetivo está por debajo del rank de corte. Un falso positivo ocurre cuando una enfermedad fuera del test está por encima del umbral de corte. Las curvas ROC [238] son una herramienta muy útil para observar la bondad de un clasificador binario mientras se varía su umbral de discriminación. El área bajo la curva (AUC[15]) se ha calculado para cuantificar los resultados.

La exactitud del ranking se midió para las dos configuraciones de redes y los resultados se resumen en las curvas ROC de la figura 2.4. Ambas configuraciones obtuvieron un valor AUC muy alto en el LOO estándar: 0.9504 para fármaco-enfermedad y 0.9579 para fármaco-proteína-enfermedad. En el LOO avanzado se puede observar la utilidad de la red de proteínas para capturar relaciones indirectas fármaco-enfermedad: 0.8041 para fármaco-enfermedad y 0.8692 para fármaco-proteína-enfermedad. La adición de interacciones proteína-proteína aumenta la robustez del método en la ausencia de relaciones fármaco-enfermedad en el vecindario próximo del par fármaco-enfermedad del test. Las diferencias en términos de valores de ranking entre las dos configuraciones fueron asimismo estadísticamente significativas (p-valores calculados con t-tests y corregidos con Bonferroni). Los resultados muestran por tanto que DrugNet obtiene los mejores resultados para la configuración de tres redes que incluye la de interacciones proteína-proteína como intermediaria. Estos resultados destacan la fuerte relación entre fármacos similares y enfermedades similares compartiendo dianas terapéuticas.

Adicionalmente, se ejecutaron cinco tests de validación cruzada independiente (5-fold) para la red fármaco-enfermedad-proteína, ya que este enfoque fue el que obtenía mejores resultados. Las diferentes pruebas obtuvieron buenos resultados, demostrando la robustez del método. El AUC medio obtenido fue $0{,}9552 \pm 0{,}0015$.

[15] Area Under the Curve

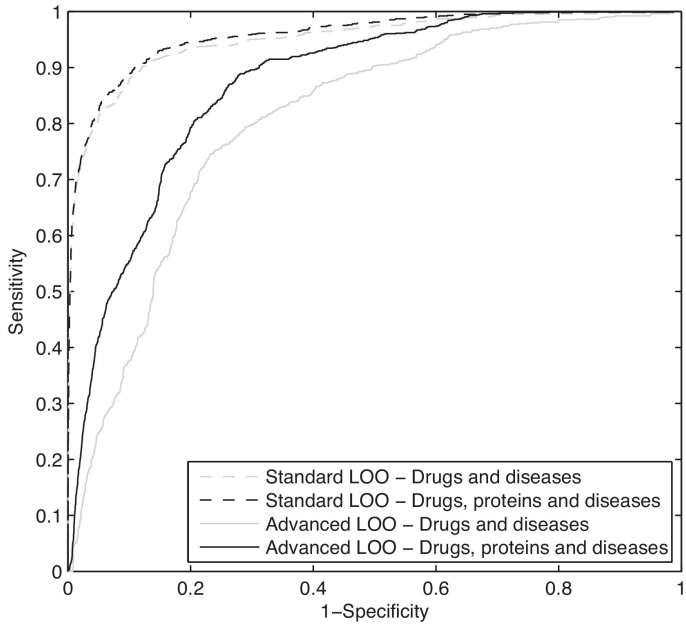


Figura 2.4: ***Resultados de DrugNet para diferentes configuraciones****. Las cuatro curvas ROC presentes en el gráfico representan los resultados para las dos configuraciones probadas en DrugNet. En línea discontinua se muestran los resultados para el LOO estándar para la configuración fármaco-enfermedad y fármaco-proteína-enfermedad. Se puede observar que en este caso la efectividad de DrugNet es similar. Sin embargo, en el caso del LOO avanzado (líneas continuas), la configuración fármaco-proteína-enfermedad obtiene un resultado muy superior a la que únicamente integra las redes de fármaco y enfermedad.*

Ensayos clínicos

Para demostrar la capacidad predictiva de casos reales no presentes en nuestros datos, aplicamos DrugNet a la priorización de relaciones no explícitamente presentes en nuestros datos y derivada de ensayos clínicos en ClinicalTrials.gov. ClinicalTrials.gov es un servicio del NIH[16] con una base de datos de ensayos clínicos realizados en humanos en todo el mundo. Un ensayo clínico es un estudio intervencional donde los participantes reciben intervenciones específicas. En

[16] National Institutes of Health, EEUU.

nuestro caso, un conjunto de fármacos o placebo. Los investigadores monitorizan cada participante para determinar la eficacia y la seguridad del fármaco. Hay cinco fases para cada estudio clínico, desde fase 0 (estudios exploratorios con exposición limitada al fármaco) a fase 4 (estudios en ejecución después de la aprobación del medicamento).

De la base de datos de ClinicalTrials.gov se extrajeron 11992 relaciones medicamento-enfermedad que no estaban presentes en nuestros datos, de las que únicamente 8217 fueron únicas, ya que había pares medicamento-enfermedad repetidos para diferentes fases del estudio clínico.

La tabla 2.2 muestra los resultados y la figura 2.5 muestra la curva ROC. Fármacos en fases más tempranas de estudio clínico tienen un alto riesgo de fracaso devido a toxicidad o baja eficacia. De hecho, una elevada proporción de medicamentos en desarrollo (alrededor del 90 %) fallan durante la fase 1 [239], cuando ya ha sido invertida una gran cantidad de dinero en su desarrollo. Como muestra la tabla, nuestros resultados fueron mejor para ensayos clínicos en fases más avanzadas. Las predicciones hechas por nuestro enfoque, por tanto, parecen fiables y potencialmente podrían reducir los costes de desarrollo de un nuevo medicamento.

2.3.3. Casos de estudio

Para validar los resultados de DrugNet en un entorno real, hemos realizado varios casos de estudio adicionales para tratar de encontrar nuevos usos para fármacos actualmente aprobados. Utilizamos como entidades de entrada medicamentos reposicionados recientemente y observamos el ranking de la enfermedad para la que el medicamento de consulta se ha empezado a utilizar. Nótese que estos nuevos usos no están explícitamente en nuestros datos en forma de relaciones fármaco-enfermedad. Este estudio ha revelado algunos resultados muy prometedores. La tabla 2.3 muestra un resumen de los resultados que se detallan a continuación.

El **metotrexato** es un fármaco utilizado inicialmente en el tratamiento de cáncer. En los últimos años, varios estudios han probado que también es efectivo en

Tabla 2.2: ***Resultados de DrugNet para ensayos clínicos.*** *Resultados obtenidos para el test de priorización utilizando ensayos clínicos recientes en proceso. La primera columna muestra la fase del estudio, la segunda el número de estudios en esta fase y la tercera columna los valores AUC de clasificación para los pares fármaco-enfermedad presentes en dicha fase.*

Fase de ensayo clínico	Núm. casos	AUC
N/A	1993	0.8106
Fase 0	105	0.8976
Fase 1	2050	0.8176
Fase 1–Fase 2	1154	0.8333
Fase 2	3307	0.8232
Fase 2–Fase 3	402	0.8108
Fase 3	1771	0.8652
Fase 4	1210	0.9109
Total	11992	0.8364

el tratamiento de la enfermedad de Crohn [240]. La enfermedad de Crohn obtuvo el puesto número 17 en los resultados de DrugNet.

Por otro lado, el **colesevelam** fue inicialmente desarrollada como una lipoproteína de colesterol (LDL-C) de baja densidad para pacientes con hiperlipidemia primaria. Estudios adicionales han demostrado que el colesevelam puede ser empleado en el tratamiento de la diabetes mellitus de tipo 2 para disminuir la tasa de hipoglucemia [241]. La hipoglucemia apareció como resultado número 16 en los resultados de DrugNet para colesevelam.

La **gabapentina** fue inicialmente utilizada en el tratamiento de la epilepsia y actualmente también se receta para reducir el dolor neuropático. Tras años de estudios se ha mostrado que también es útil para tratar trastornos de ansiedad [242]. DrugNet obtuvo como resultado Trastorno de ansiedad generalizada en el puesto número 7 para gabapentin.

El **cisplatino** ha sido utilizado para tratar diferentes tipos de cancer tales como testicular, ovarios y cáncer de pulmón, pero no para el tratamiento rutinario de cáncer de mama. Sin embargo, estudios recientes han demostrado que este fármaco reduce la expresión de BRCA1 en cánceres de mama triple negativo [243].

Tabla 2.3: ***Casos de estudio de reposicionamiento con DrugNet.*** *Resultados obtenidos para los casos de estudio realizados con DrugNet de reposicionamientos para compuestos con nueva aplicación descubierta no presente en nuestros datos. Para cada compuesto se muestra su aplicación clásica y su nueva aplicación, así como los resultados proporcionados por DrugNet. Nótese que la posición es respecto a la lista total de enfermedades (4517).*

Compuesto	Aplicación clásica	Nueva aplicación	DrugNet rank (enfermedad)
metotrexato	Cáncer	Enfermedad de Crohn [240]	17 (Crohn)
colesevelam	Hiperlipidemia	Hipoglucemia en diabetes mellitus tipo 2 [241]	16 (Hipoglucemia)
gabapentina	Epilepsia, dolor neuropático	Trastorno de ansiedad [242]	7 (Trastorno de ansiedad generalizada)
cisplatino	Cáncer (testicular ovarios, pulmón)	Cáncer de mama triple negativo [243]	11, 12 (Melanoma maligno de mama) y Linfoma de mama)
donepezilo	Alzheimer	Parkinson [244]	2 (Parkinson)
risperidona	Esquizofrenia	Agitación, psicosis y trastorno de sueño en TOC [245]	8 (TOC)

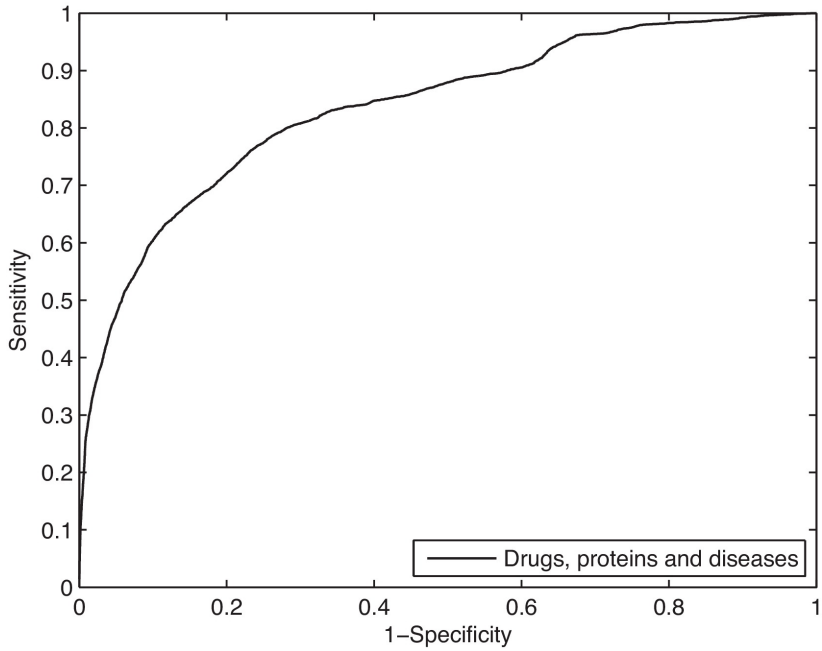


Figura 2.5: ***Resultado de DrugNet para datos de ensayos clínicos****. Curva ROC para la prueba de predicción de fármacos actualmente en proceso de ensayo clínico para la configuración de DrugNet de fármaco-proteína-enfermedad.*

Nuestro método obtuvo en los puestos 11 y 12, melanoma maligno de mama y linfoma de mama, respectivamente.

El **donepezilo** es un inhibidor reversible de la acetilcolinesterasa que ha sido utilizado principalmente para el tratamiendo del Alzheimer, pero que recientemente ha sido aplicado al Parkinson [244]. En nuestra aplicación, Parkinson obtuvo el puesto número 2.

La **rispedirona** es un fármaco atípico antipsicótico utilizado para el tratamiento de la esquizofrenia. Recientemente se ha comprobado que puede ayudar a reducir la agitación, psicosis y trastorno del sueño entre otros síntomas en pacientes con trastorno obsesivo compulsivo [245]. DrugNet obtuvo como resultado la posición 8 para trastorno obsesivo compulsivo.

2.3.4. Comparación con otros métodos

Se ha realizado además una comparación manual de DrugNet con PREDICT [224], un método de reposicionamiento de enfermedades ampliamente citado y de referencia en el campo. Esta comparativa ha entrañado ciertas dificultades, ya que una comparación automatizada exhaustiva no fue posible al no estar PREDICT disponible como código fuente o como herramienta web. En segundo lugar, el uso de diferentes bases de datos para construir las redes (por ejemplo, Disease Ontology u OMIM) produce problemas de terminología, ya que los términos en estas bases de datos no siempre coinciden. Por ello se realizó una revisión manual de la literatura para obtener evidencia que soporte los resultados obtenidos tanto para PREDICT como para DrugNet para un set de consultas limitado.

Primero, realizamos consultas en PREDICT y DrugNet para nuevos usos de fármacos. Por ejemplo, PREDICT solo sugiere nedocromil para tratar dermatitis atópica. Esta hipótesis es validada en la literatura [246] y coincide con la mejor predicción de DrugNet. Sin embargo, DrugNet sugiere otras enfermedades que también pueden ser tratadas con nedocromil. Por ejemplo, bronquitis, y rinitis alérgica, cuyo tratamiento con nedocromil aparece en [247] y [248], respectivamente, están en los puestos 2 y 3 de DrugNet. Algo similar ocurre con el tiludronato. Mientras PREDICT sugiere sólo una variente de acro-osteólisis con osteoporosis como un objetivo potencial para el tiludronato, DrugNet sugiere osteoporosis (posición 7), hipótesis que se sugiere en [249], enfermedad ósea de Paget (posición 1), sugerida en [250], osteoartritis (posición 3), sugerida en [251] e hipercalcemia (posición 5), sugerida en [252], entre otros. DrugNet parece por tanto más capaz de realizar predicciones válidas para consultas de fármacos.

A continuación realizamos consultas en PREDICT y DrugNet para reposicionar fármacos para enfermedades de interés. Por ejemplo, PREDICT sugiere propiltiouracilo, metimazol, meprobamato, tizanidina y ciclobenzaprina para tratar el mixedema [253]. Entre estos, DrugNet sólo sugiere metimazol (posición 3) y propiltiouracilo (posición 7), pero no el resto, para las que no encontramos ninguna evidencia de ninguna relación con mixedema en la literatura. Adicionalmente, otros fármacos sugeridos por DrugNet son la

levotiroxina [254] (posición 1), desmopresina [255] (posición 5) y vasopresina [256] (posición 8).

Una comparación más exhaustiva entre ambas herramientas sigue siendo un problema a resolver, ya que requiere la disponibilidad de PREDICT.

2.3.5. Limitaciones del enfoque

Aunque las fuentes de datos que hemos integrado en DrugNet han sido las más completas que hemos podido encontrar para reposicionamiento de medicamentos, realizamos varios tests para ver si DrugNet presenta limitaciones en forma de sesgos inherentes en estas fuentes de datos.

Primero, comprobamos si ciertas categorías farmacológicas (en nuestro caso códigos ATC) daban mejores resultados que otras. Para conseguir esto realizamos dos tests de validación cruzada (LOO y LOO avanzado) para la red fármaco-proteína-enfermedad y medimos la efectividad de nuestro método para obtener la enfermedad objetivo después de utilizar como entrada un fármaco perteneciente a una categoría ATC específica. Para cada categoría ATC hemos computado una serie de características topológicas básicas de la red para estudiar la influencia de estas características en el resultado de reposicionamiento de los fármacos de estas categorías. Estos resultados se detallan en la tabla 2.4.

Los resultados fueron mejores para fármacos en las categorías *Respiratory system*, *Musculo-skeletal system* y *Nervous system*. Estas categorías mostraron un número más alto de conexinones "enfermedad a fármaco" (número de arcos conectando la enfermedad objetivo con diferentes nodos de fármaco), lo que puede contribuir a esta mejora de resultados. Dado que consultamos al sistema por un fármaco de interés, el flujo de propagación va desde fármacos a enfermedades, así que cuanto mayor número de conexiones presente una enfermedad, la cantidad de información que recopila de la red de fármacos será mayor. Es importante notar que este no es el caso para la característica "fármaco a enfermedad" (número de arcos que conectan el medicamento de consulta con diferentes nodos de enfermedad), ya que un valor más elevado de esta característica implica que la señal está dividida entre más receptores (enfermedades), así que no se correla con mejores resultados.

Tabla 2.4: ***Propiedades y resultados de DrugNet desglosados por categoría ATC.*** *Columnas de izquierda a derecha: Categoría ATC del fármaco de consulta, número de dianas comunes entre el fármaco de consulta y la enfermedad objetivo, número medio de conexiones entre el fármaco de consulta y enfermedades, número medio de conexiones entre la enfermedad objetivo y fármacos, número de arcos fármaco-enfermedad en la cateogría, rank medio para el test LOO simple, rank medio para el test LOO avanzado, y rank medio de ambos tests.*

Categoría ATC	Dianas comunes	Fármaco a enfermedad	Enfermedad a fármacos	Casos	Rank medio LOO simple	Rank medio LOO avanz.	Rank global
Alimentary tract and metabolism	0.0356	14.0044	6.1289	225	301.4933	834.4311	567.9622
Blood and blood forming organs	0.0750	1.8000	4.1500	40	183.4500	592.8750	388.1625
Cardiovascular system	0.0632	4.3684	8.4000	95	307.1789	776.1368	541.6579
Dermatologicals	0.0070	4.2867	11.7972	143	206.4755	632.5734	419.5245
Genito-urinary system and sex hormones	0.2222	3.1944	5.8750	72	286.3194	563.2639	424.7917
Systemic hormonal preparations, excluding sex hormones and insulins	0.0928	27.8866	6.0103	97	565.5361	927.6495	746.5928
Antiinfectives for systemic use	0.0207	3.2207	8.0897	145	152.1586	483.7379	317.9483
Antineoplastic and immunomodulating agents	0.3710	3.7097	5.1371	124	189.4839	734.8871	462.1855
Musculo-skeletal system	0.1159	2.3913	12.4348	69	80.8986	508.0580	294.4783
Nervous system	0.5782	2.6735	7.0000	147	130.5442	304.6599	217.6021
Antiparasitic products, insecticides and repellents	0.1071	3.0000	2.2857	28	136.0714	205.8571	170.9643
Respiratory system	0.0360	3.8777	24.5899	139	74.8058	493.7626	284.2842
Sensory organs	0.0427	4.2137	10.906	117	142.6838	648.8462	395.7650
Various	0.0588	1.5882	9.2353	17	781.8824	1100.5294	941.2059

También ejecutamos un LOO avanzado para evitar la excesiva redundancia que pueda surgir de la existencia de comunidades de asociaciones de medicamentos y enfermedades. Para el test LOO avanzado, las categorías con mejores resultados fueron *Antiparasitic products, insecticides and repellents*, *Nervous system* y *Antiinfectives for systemic use*. Vale la pena mencionar que hubo una diferencia significativa entre los resultados de LOO simple y LOO avanzado. Esto se debe al hecho de que los fármacos en categorías ATC con más conexiones entre fármaco y enfermedad (valores altos en "fármaco a enfermedad" y "enfermedad a fármaco") son más penalizados en el LOO avanzado, ya que estas conexiones son eliminadas de la red. Este es el caso para *Respiratory system*, que mostraba los mejores resultados para el LOO simple y bajó significativamente su posición media en el ranking para LOO avanzado (de una media de ranking de 75 en los LOO simples a una posición media de 493 en los LOO avanzados). Esta categoría mostraba la mayor cantidad de conexiones entre fármacos y enfermedades (una suma media de 28 conexiones), y estas conexiones fueron eliminadas en la validación cruzada ejecutada por los tests LOO avanzados, justificando la bajada en el desempeño.

Por otro lado, la bajada media para *Antiparasitic products, insecticides and repellents* fue solo de 70 posiciones en el ranking del LOO avanzado con respecto a los resultados de LOO simple, resultando esta categoría la que mejor resultados da para los LOO avanzados. Esta categoría sólo presentaba una media de 5 conexiones entre fármacos y enfermedad, por lo que la eliminación de estas conexiones no representó un impacto significativo como en el caso anterior.

Existe también un sesgo contra los fármacos asociados a la categoría ATC *Various*. Esto era de esperar ya que los fármacos en esta categoría son mucho más heterogéneos. Los medicamentos en la categoría *Systemic hormonal preparations, excluding sex hormones and insulins* también obtuvieron resultados de reposicionamiento pobres, probablemente debido al valor significativamente alto de la propiedad "fármaco a enfermedad" (27,8 cuando la media es de 5,72). Esto implica que los fármacos en esta categoría son ampliamente utilizados para muchas enfermedades diferentes y, por tanto, la información está repartida en muchas enfermedades, lo que hace que el reposicionamiento sea peor.

2.3.6. Extensión de la red

Como se ha explicado a lo largo de estas secciones, la metodología propuesta puede ser aplicada sobre una red heterogénea compuesta por distinto número de subredes que conectan entidades de distinto tipo. En este caso, la configuración que proporcionó mejores resultados es la que integra tres subredes de distinto tipo: fármaco-proteína diana-enfermedad. No obstante, podrían añadirse nuevas entidades. Para poder integrar una nueva red de entidades a la metodología ProphTools, es necesario para nuestro motor de inferencia que los nodos de esta red puedan conectarse también a los nodos de otras subredes del sistema, generando al menos un camino entre las redes de consulta y de destino de la inferencia.

En caso de que esto no sea posible, es posible agregar este conocimiento de forma implícita en alguna de las redes ya conectadas. En nuestro caso, DrugNet no agrega conocimiento de diferentes fuentes en la misma red, sino que cada fuente de información deriva su propia subred.

Consideremos ahora el caso de integrar información sobre los efectos secundarios de un fármaco. Hasta donde sabemos, esta información sólo se puede relacionar con fármacos. Esto limita su integrabilidad, ya que incluso si fuera posible construir una nueva subred interconectando efectos secundarios, no sería posible conectar estos a elementos de otras redes. Sin embargo, esta información podría ser integrada en la red existente utilizando información sobre efectos secundarios.

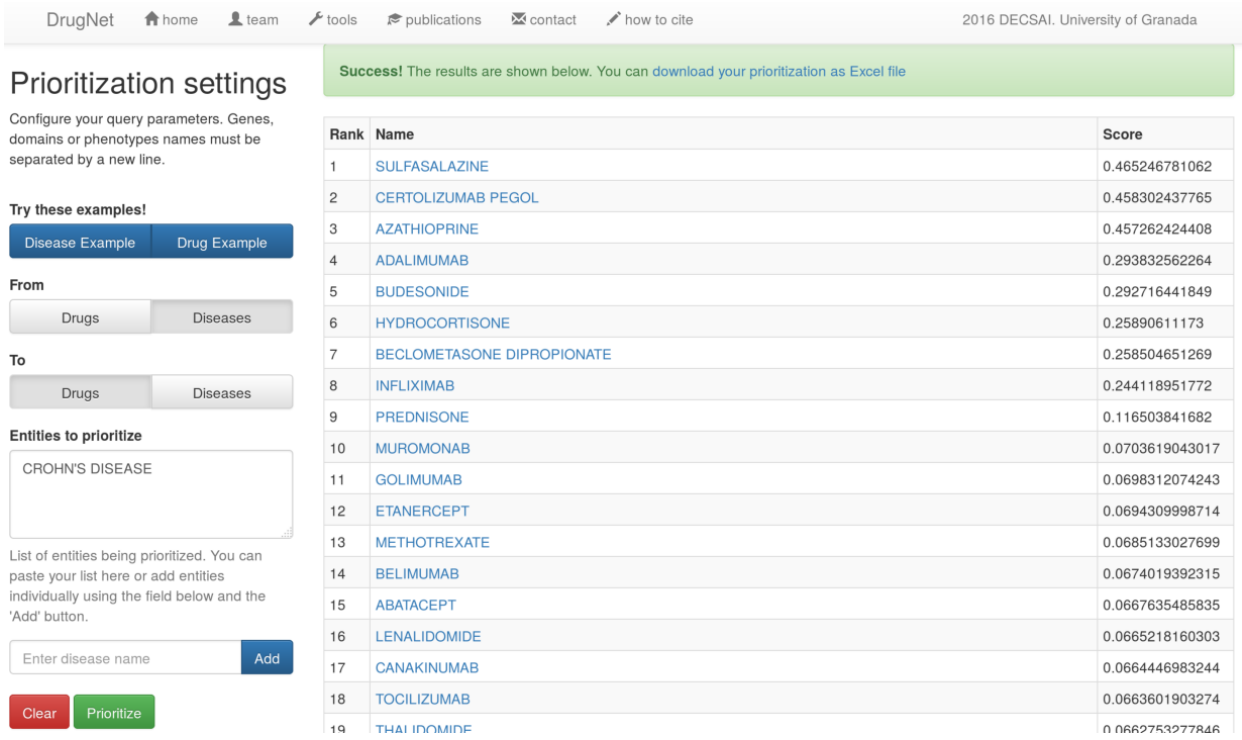
Un enfoque sencillo para acometer este problema sería aumentar la similitud de dos fármacos si comparten un número elevado de efectos secundarios. Este principio de asociación indirecta entre entidades ha sido utilizado en otras herramientas, como [257], en la que se construye la red fármaco-fármaco utilizando similitud de efectos secundarios. Este principio también podría ser aplicado bajo la metodología general de DrugNet para refinar la subred de similaridad de fármacos, agregándola a la medida actual de similitud/distancia entre fármacos.

2.3.7. Servidor web

La metodología propuesta ha sido implementada como una herramienta web a la que se puede acceder en el servidor genome: http://genome.ugr.es:9000/drugnet. La figura 2.6 muestra una captura de la web. La aplicación web ha sido desarrollada utilizando python flask[17], y la interfaz hace uso del framework Bootstrap[18].

La web de DrugNet permite priorizar enfermedad-fármaco y fármaco-enfermedad, así como hacer consultas dentro de la misma red (fármaco-fármaco, enfermedad-enfermedad). Dispone de una función de autocompletado que sugiere nombres de medicamentos y de enfermedades para añadirlos a la lista de entidades de consulta. Además, los resultados se muestran en una lista que incluye enlaces al término correspondiente de Disease Ontology en el caso de enfermedades y a Drugbank en el caso de fármacos. Por último, estos resultados se pueden descargar en formato excel para su posterior análisis detallado *offline*.

[17] http://flask.pocoo.org
[18] http://getbootstrap.com



Rank	Name	Score
1	SULFASALAZINE	0.465246781062
2	CERTOLIZUMAB PEGOL	0.458302437765
3	AZATHIOPRINE	0.457262424408
4	ADALIMUMAB	0.293832562264
5	BUDESONIDE	0.292716441849
6	HYDROCORTISONE	0.25890611173
7	BECLOMETASONE DIPROPIONATE	0.258504651269
8	INFLIXIMAB	0.244118951772
9	PREDNISONE	0.116503841682
10	MUROMONAB	0.0703619043017
11	GOLIMUMAB	0.0698312074243
12	ETANERCEPT	0.0694309998714
13	METHOTREXATE	0.0685133027699
14	BELIMUMAB	0.0674019392315
15	ABATACEPT	0.0667635485835
16	LENALIDOMIDE	0.0665218160303
17	CANAKINUMAB	0.0664446983244
18	TOCILIZUMAB	0.0663601903274
19	THALIDOMIDE	0.0662753277846

Figura 2.6: ***Captura de aplicación web de DrugNet.*** *DrugNet web permite priorizar entre fármaco y enfermedad en ambos sentidos, obteniendo la lista de resultados con enlaces a las entidades de interés y también a la descarga de los resultados en formato excel.*

2.4. Conclusiones

Los últimos años han visto crecer y consolidarse un abanico extenso de metodologías experimentales, las cuales han dado lugar a multitud de bases de datos privadas y públicas con vastísimas cantidades de información de muy diversas características. La extracción de conocimiento de la gran cantidad de datos biológicos disponibles públicamente para la comundiad científica es hoy en día, más que nunca, un reto. Asimismo, no sólo las fuentes de información disponible son múltiples, heterogéneas y extensas, sino también lo son de manera análoga las entidades biológicas representadas por ellas. En este sentido, la representación de esta información en forma de grafo o red resulta conveniente para facilitar el análisis de esta información y la predicción de nuevas hipótesis no explícitamente representadas en los datos.

Esta percepción de la biología como una red de elementos de diversa índole que interactúan entre sí en formas complejas no es nueva, y así lo demuestran la multitud de metodologías que existen para la resolución de problemas concretos, de gran calado en la investigación biomédica y en el marco de la Medicina Personalizada, como pueden ser la priorización gen-enfermedad o el reposicionamiento de medicamentos.

Sin embargo, estas metodologías incluyen en un todo indivisible las fuentes de datos utilizadas para construir la red o redes y el método de priorización utilizado para su análisis. Son, por tanto, metodologías dependientes del ámbito de utilización, limitadas o restringidas a un dominio o problema concreto, por lo que su aplicación a otros problemas resulta inviable o presenta un coste de adaptación y cuya aplicación a otros problemas resulta inviable.

Se ha propuesto una metodología general y flexible que puede integrar un número arbitrario de redes heterogéneas para la priorización de cualquier tipo de entidad representada en estas redes. Esta metodología utiliza mecanismos de propagación de información en redes que fueron utilizados con éxito previamente para la priorización *ad hoc* gen-enfermedad [88].

Utilizando como base estos mecanismos de propagación, se ha propuesto la adición de una capa superior de abstracción que modela cualquier conjunto de

subredes vinculadas en una red global heterogénea. De este modo, se permite al usuario analizar cualquier conjunto de entidades biológicas de su interés, no limitando el conjunto de redes a dos o tres, como es frecuente en muchos de los problemas abordados por estas metodologías.

La metodología propuesta, ProphTools [182], se ha desarrollado en python, utilizando bibliotecas de libre acceso, y su código se ha puesto a disposición de la comunidad científica en forma de código abierto bajo licencia GPLv3 en un repositorio público, GitHub[19], facilitando su uso y extensión futura, tanto por nuestra parte como por cualquier investigador interesado. Prueba de ello es la inclusión en ProphTools de los métodos de propagación de RANKS [102] como alternativas a la propagación *intra-red*, implementados por la metodología general de ProphTools. Estos métodos alternativos de propagación han sido fácilmente integrados en ProphTools gracias a su estructura modular. En el futuro, planeamos integrar nuevos métodos y evaluar comparativamente su eficacia en distintos problemas de priorización.

El valor de esta metodología, además, se ha demostrado en un caso de estudio que integra priorización gen-enfermedad con análisis de secuencia: DiGSNP [89]. Este tipo de aplicación muestra además, que la priorización de entidades en un ámbito de red heterogénea de multitud de tipos de elementos puede ser completada por estudios posteriores, que profundicen en el conocimiento sobre las relaciones ya descubiertas. No sólo podría este planteamiento vincularse posteriormente a metodologías de análisis computacional, sino también a validaciones experimentales, como por ejemplo, ChIP-seq, la cual podría validar algunos de los TFBSs propuestos, y podría ser una línea de trabajo futuro.

El segundo caso de estudio corresponde a un ámbito de gran interés en los últimos años y especialmente en el marco de la Medicina Personalizada: el reposicionamiento de medicamentos. El proceso de desarrollo de un fármaco nuevo es tan largo y costoso que encontrar nuevos tratamientos de enfermedades entre fármacos ya comercializados para otras aplicaciones puede ahorrar gran cantidad de tiempo y de dinero en el proceso, garantizando unos mínimos de

[19] http://www.github.com/cnluzon/prophtools

seguridad y aprovechando los resultados de ensayos clínicos ya realizados en el pasado concernientes sobre todo a toxicidad y seguridad de los medicamentos.

En este contexto, se ha propuesto DrugNet [90], una herramienta de reposicionamiento de medicamentos. DrugNet ha sido testado con dos configuraciones de redes distintas, mostrando una de las ventajas de la metodología propuesta: la capacidad de integrar un número arbitrario de redes conectando distintos tipos de entidades sin alterar los métodos de priorización que analizan estos datos.

Los resultados con DrugNet demuestran que es posible realizar predicciones fiables en las redes utilizadas. En particular, se ha mostrado que DrugNet tiene capacidad de predecir nuevos usos para medicamentos existentes no conocidos ni explícitamente representados en el momento de construir las redes de datos. Del mismo modo, se ha mostrado que los resultados de priorización son mejores para medicamentos que están en fases más avanzadas de ensayos clínicos, mostrando que las predicciones de DrugNet concuerdan con los resultados de ensayos clínicos efectuados sobre medicamentos para nuevas aplicaciones.

La metodología propuesta está disponible públicamente en la web del servidor genome, `http://genome.ugr.es:9000/drugnet`, permitiendo la ejecución de consultas sobre medicamentos y enfermedades.

Los resultados de las pruebas realizadas con DrugNet muestran que este enfoque puede identificar aplicaciones de medicamentos desconocidas anteriormente que son fiables en situaciones reales, y por tanto estos métodos podrían reducir la gran cantidad de recursos requeridos en el complejo proceso de desarrollo de medicamentos. Aparte del hecho de que la seguridad de los fármacos reposicionados no suele ser un problema a resolver, ya que éstos han sido comercializados con anterioridad para otras indicaciones, el método propuesto en esta línea de trabajo generalmente sugiere medicamentos que están en etapas avanzadas de ensayos clínicos. Esto parece sugerir que los medicamentos reposicionados por DrugNet son probablemente tanto seguros como efectivos.

Adicionalmente, DrugNet no sólo puede utilizarse como reposicionador de medicamentos, sino que también puede ser consultado en orden inverso, es decir,

pueden buscarse nuevas indicaciones para fármacos existentes, lo que puede ser un enfoque interesante en muchos casos, como por ejemplo cuando un fármaco seguro ha sido archivado por falta de efectividad.

Por otra parte, la priorización fármaco-enfermedad también puede ser útil en la búsqueda de tratamientos para enfermedades raras, para las que no se ha encontrado un medicamento adecuado, o para mejorar tratamientos existentes.

La ampliación del conjunto de fármacos para el tratamiento de cualquier enfermedad no puede sino jugar en ventaja del paciente, que verá crecer la probabilidad de éxito de su tratamiento, al existir entre un mayor número de alternativas una que se adecúe en mayor medida a sus características individuales.

Capítulo

3

Detección de elementos reguladores en secuencias de ADN

Uno de los principales objetivos de la Medicina Personalizada es tomar en consideración las características individuales presentes en la información genética del paciente para mejorar su diagnóstico y tratamiento. Tanto médicos como pacientes tendrán, en un futuro próximo, fácil acceso a información genética que permita este tipo de atención individualizada. En este sentido, se hace necesario entender la relación entre la variabilidad de los individuos y sus posibles dolencias o tratamientos óptimos. Por ello, los estudios cuyo objetivo es asociar adecuadamente un genotipo a un fenotipo han adquirido gran protagonismo en la actualidad.

Estos enfoques permiten averiguar la propensión a la aparición de ciertas enfermedades o características en un individuo a partir de su información genómica. Dado que los seres humanos compartimos un 99.9% de secuencia en nuestro ADN [12], en estas diferencias reside gran parte de esta información. En cuanto a estas variaciones, los polimorfismos de nucleótido simple (Single Nucleotide Polymorphisms – SNPs) han sido las más extensivamente estudiadas en busca de asociaciones con enfermedades.

Tradicionalmente se ha pensado que las variaciones que más impacto tienen en el fenotipo del individuo son aquellas en regiones codificantes del ADN. Sin embargo, las regiones no codificantes han recibido un mayor protagonismo en los últimos años. Estas regiones son las responsables de controlar el proceso de fabricación de proteínas, y por tanto la aparición de mutaciones situadas en estas regiones puede influir en el funcionamiento del organismo, pudiendo dar lugar, en última instancia, a un proceso patológico o enfermedad. Los SNPs capaces de alterar estas regiones son potenciales candidatos a tener funcionalidad reguladora.

Por otra parte, los estudios GWAS arrojan información sobre SNPs relacionados con enfermedades, pero no explican los mecanismos por los cuales esta relación se produce. Además, como se ha mencionado anteriormente, cerca de una tercera parte de los SNPs relacionados con enfermedades según estudios GWAS recientes pertenecen a regiones no codificantes del genoma [15]. Determinar el impacto funcional de mutaciones en regiones no codificantes asociadas a enfermedades es, por tanto, una tarea de interés, ya que proporciona información sobre los posibles mecanismos por los que un SNP no codificante puede llegar a afectar a un fenotipo.

La regulación génica es un proceso dinámico y complejo. En este proceso los factores de transcripción juegan un papel clave, ya que forman parte fundamental del comienzo y la regulación de la transcripción. Sin embargo, se ha demostrado que dicha actividad además suele estar coordinada formando lo que se conoce como módulos reguladores *cis* (Cis-Regulatory Modules – CRMs) [159]. Así, la complejidad que entraña el estudio de la regulación de los genes aumenta drásticamente si a la búsqueda de motivos individuales en el genoma se le añade una nueva capa de complejidad combinatoria. A dicha complejidad combinatoria se le añade la incertidumbre siempre presente en la predicción de TFBSs y el desconocimiento preciso sobre la distancia o distribución de dichos lugares dentro del CRM, lo que dificulta en gran medida la búsqueda de éstos módulos.

En general, las herramientas existentes para la búsqueda de estas entidades reguladoras se clasifican según su ámbito de aplicación, siendo las más específicas las más desarrolladas en el presente. Estos enfoques suelen requerir conocimiento

a priori sobre los módulos reguladores a buscar, ya que los análisis exploratorios involucran imprecisión y suelen implicar altos costes combinatorios.

El objetivo de esta línea de trabajo es abordar el análisis de secuencias de ADN en búsqueda de elementos reguladores en los dos niveles de complejidad mencionados: por una parte, intentando relacionar la variabilidad en forma de mutaciones en regiones reguladoras con enfermedades; por otro lado, extendiendo dichas regiones reguladoras a aquellas más complejas que representan los módulos reguladores en *cis*. En ambos casos se ha tenido en cuenta la incertidumbre inherente a los datos biológicos manipulados, integrando en las metodologías tecnología difusa.

El conocimiento sobre estos mecanismos reguladores y las metodologías existentes para su estudio se explican en la sección 3.1. A continuación se presenta en la sección 3.2 la metodología IntuitSNP propuesta para la búsqueda de TFBSs afectados por SNPs. Posteriormente, la aportación propuesta en el ámbito del estudio de los módulos reguladores en *cis*, CisMiner [258], es detallada en 3.4. Las conclusiones extraídas en esta línea de trabajo se enumeran en la sección 3.5.

3.1. Análisis computacional y experimental de secuencias reguladoras

El estudio sobre la variación en el genoma es de especial relevancia en lo que concierne al diagnóstico y tratamiento de enfermedades. Así pues, encontrar relaciones entre mutaciones genómicas y posibles trastornos o desórdenes se encuentra entre los campos con más importancia en la actualidad. Entre las posibles mutaciones que pueden aparecer en el genoma de un individuo, las más estudiadas son los polimorfismos de nucleótido simple o SNPs, ya que son las más frecuentes en el genoma, con una frecuencia de aparición aproximada cada 1200 pares de bases en el genoma humano [148]. Así pues, en este trabajo de tesis nos centramos en el estudio de variaciones tipo SNPs, en conjunción con su presencia en las regiones reguladoras del genoma. En esta sección se explica en detalle el contenido de las bases de datos más relevantes disponibles en la actualidad cuyo contenido involucra secuencias, SNPs, información sobre regulación génica y literatura biomédica.

3.1.1. Búsqueda de SNPs funcionales

Actualmente hay pocas propuestas relativas al estudio de las variaciones en las áreas reguladoras del genoma, centrándose la mayoría de las propuestas en el estudio de regiones codificantes, es decir, regiones que se traducen directamente en mARN que posteriormente dará lugar a la secuencia de aminoácidos que conforman la proteína. Cuando un SNP aparece en una secuencia de ADN en la que hay un lugar de unión de un factor de transcripción, dicha mutación puede alterar la afinidad de unión de esta cadena de ADN con el factor de transcripción en cuestión, alterando así el proceso de regulación. En particular, un SNP en un TFBS se puede encontrar en una de cuatro situaciones distintas:

- **Inhibición de un TFBS**. El SNP altera la afinidad del TFBS de forma que el factor de transcripción que se unía a esta secuencia ya no es capaz de unirse. Se dice entonces que el SNP inhibe el proceso de transcripción, silenciando el gen.

- **Creación de un nuevo TFBS**. El SNP produce una alteración en la cadena de ADN de forma que se genera una afinidad con un TF que antes no existía, haciendo que un gen que antes no se expresaba lo haga.
- **Intercambio de TFBSs**. Se produce una combinación de las dos situaciones anteriores, silenciando un proceso de transcripción e iniciando otro.
- **SNP sinónimo**. El SNP no altera la afinidad del TFBS, así que no altera el proceso de regulación.

Es interesante distinguir entre estas cuatro situaciones, ya que de esta manera se obtiene una mayor cantidad de información sobre la probabilidad de que cada SNP altere los mecanismos de transcripción celulares. A continuación se detallan las principales propuestas relacionadas con este enfoque encontradas en la literatura.

Entre las metodologías que tratan de evaluar computacionalmente si un SNP afecta a un posible TFBS, se encuentra **is-rSNP** [152]. Se trata de una plataforma web cuyo objetivo es proporcionar al usuario una lista de motivos TRANSFAC o JASPAR candidatos a estar relacionados con los SNPs indicados. Los candidatos se calculan en base a las matrices PWM, sin tener en cuenta las secuencias que generaron dichas matrices.

Los autores de is-rSNP definen un score básico para una secuencia contra una matriz PWM $W = [w(j,i)]_{4\times n}$ de la siguiente manera:

$$S_n(b) = \sum_{i=1}^{n} w(b_i, i)$$

donde b es una secuencia de nucleótidos tal que $b = (b_1, ..., b_n) \in B^n$ siendo $B = A, C, G, T$.

La propuesta se basa en un algoritmo de ventana deslizante sobre el SNP. Para cada motivo existente en TRANSFAC y/o JASPAR, se desliza una ventana de la longitud del motivo de forma que éste contenga siempre al SNP, y se calcula un score para cada una de las posiciones, de forma que si el motivo tiene longitud n, se generan n scores diferentes, y se conserva el mejor.

El score final se basa en realizar el cálculo de la distribución de todos los posibles scores básicos generados por la matriz PWM. Dicho de otra forma, calcula el score que genera la PWM para todas las posibles secuencias de longitud n y se construye una distribución de scores. En general, este cálculo, reducible al conocido problema de la suma de subconjuntos, es NP duro [259]. No obstante, previo al cálculo de la distribución se realiza una normalización y discretización previa de la matriz PWM. Debido a que el cálculo del score se limita a sumar el valor de la matriz para cada posición, se puede utilizar un algoritmo de programación dinámica para calcular la distribución en un tiempo polinomial. Sin embargo, la operación sigue siendo muy costosa, ya que esta operación se realiza para todas las matrices en TRANSFAC y/o JASPAR.

Una vez realizado el cálculo de esta distribución, se puede obtener un p-valor que indique *cómo de alto* en realidad es el score obtenido. Si la secuencia de consulta tiene un score alto, se pasa a comprobar el alelo mutado, y a calcular de forma análoga una distribución de probabilidad de las posibles diferencias de scores generadas por la matriz, de forma que si, de nuevo, el score obtenido es alto, el SNP se considera un SNP regulador, o rSNP, asociándolo al TFBS indicado por el motivo ligado a la PWM. Esta operación se realiza para todas las matrices de TRANSFAC y JASPAR.

Una de las principales desventajas de este método es la simplicidad del score inicial, el cual tiene en cuenta únicamente los valores de la matriz PWM. El resto de información disponible en TRANSFAC es ignorada. No se tiene en cuenta tampoco una posible interacción entre columnas, ni las secuencias que dieron lugar a la matriz. Los resultados reportados son discretos y muestran lugar a mejora [152].

Otra herramienta disponible para este mismo problema es **sTRAP**[1] [153]. sTRAP permite analizar variaciones en secuencias de ADN con el objetivo de predecir correctamente cambios en la afinidad de éstas para unirse a un factor de transcripción. Es una extensión de la herramienta TRAP [154], la cual combina con un enfoque estadístico para normalizar las afinidades de unión de los diferentes factores de transcripción.

[1] http://trap.molgen.mpg.de/cgi-bin/trap_two_seq_form.cgi – Acceso el 23-12-2016

Para la predicción de las afinidades de unión utilizan un marco biofísico en el que para las afinidades de unión locales en la posición l de la secuencia se evalúa la siguiente expresión:

$$a_1(R_0, \lambda) = \frac{R_0 e^{E_L(\lambda)}}{1 + R_0 e^{E_l}(\lambda)}$$

donde $R_0(W) = 0{,}6W - 6$ y $\lambda = 0,7$ son dos parámetros que fueron ajustados en [154]. W representa el tamaño del motivo. Para cada SNP y cada motivo, se calculan los W pares de afinidades locales que pueden variar entre el alelo mutado y el de referencia. Además, los autores proponen un segundo enfoque en el que consideran no sólo la secuencia correspondiente al motivo, sino la afinidad general A de una secuencia más larga que la contiene. Una vez planteadas estas medidas de similitud, se modela para cada factor, la distribución de afinidades, para así obtener un p-valor. Con este fin se utiliza la siguiente expresión, propuesta por Manke *et al.* en [260], y cuyos parámetros también son ajustados en dicho artículo:

$$logA \sim P(x|a, b, c) = exp(-[1 + a\frac{x - c}{b}]^{-1/a})$$

Como criterio de selección de SNPs reguladores, los autores de sTRAP proponen tanto un log-ratio entre las afinidades o scores de la secuencia de referencia (S1) y la mutada (S2):

$$r_x = log_{10}(\frac{S1}{S2})$$

No obstante, también consideran la posibilidad de proporcionar como candidatos aquellas secuencias cuya afinidad sea alta con el factor de transcripción, independientemente de que ésta varíe mucho o no. Esta última consideración es interesante, ya que contempla la posibilidad de que, bien por incapacidad actual para detectar variaciones significativas en la afinidad, como por dificultades para establecer un umbral de significación para estas diferencias, se pudieran omitir TFBSs que tienen interés regulador. Además, podría ser que los SNPs afectaran a estas secuencias de una forma más compleja, aún no modelada en las medidas propuestas. Sin embargo, al investigador con interés potencial en que utilice estas

herramientas podría seguir interesándole saber que un SNP está ubicado en un TFBS aún en el caso de que éste no parezca afectar a las funciones reguladoras del gen.

Además de is-rSNP y sTRAP, existen otras propuestas relacionadas con la búsqueda de elementos funcionales en el genoma, tales como BCRANK[2] [155], una propuesta que aúna información de experimentos in vivo con técnicas computacionales para determinar potenciales SNPs reguladores. El algoritmo se basa en una estrategia heurística de búsqueda que requiere una estimación inicial. También existen herramientas cuya finalidad, más que proponer métodos para detectar TFBSs, consiste en proporcionar una herramienta web que permita realizar análisis de diversa índole sobre datos relacionados con SNPs. Tal es el caso de PupaSNP [156], que analiza SNPs en exones, regiones reguladoras e intrones, entre otros. Para el caso de SNPs en regiones correspondientes a TFBSs, PupaSNP utiliza la herramienta MATCH [261] proporcionada por TRANSFAC.

[2] predicting Binding site Consensus from RANKed sequences.
http://www.bioconductor.org/packages/release/bioc/html/BCRANK.html – Acceso el 20-12-2016

3.2. IntuitSNP: Influencia de mutaciones en lugares de unión de factores de transcripción

Es sabido que una de las causas de diversas enfermedades son las mutaciones en el genoma del individuo [13]. En este sentido, sería deseable disponer de información relativa a qué mutaciones están asociadas a ciertas enfermedades. La existencia de esta información sería un avance en la predicción, diagnóstico y tratamiento de enfermedades. Los SNPs o mutaciones de sustitución de una sola base son las más estudiadas por ser el tipo de mutación más numerosa y documentada. Aunque existen otro tipo de variaciones, como inserciones y borrados de secuencias de un nucleótido o de mayor longitud, se ha mostrado que los SNPs están muy relacionados con el desarrollo de enfermedades.

Hasta hace algunos años, el estudio de las regiones codificantes del genoma acaparaba la mayor parte de la atención de los investigadores, ya que estas regiones codifican de forma directa la producción de proteínas en la célula y su impacto directo en la expresión genética y el fenotipo de un individuo es más fácil de cuantificar. Sin embargo, recientemente el interés ha recaído también en otras regiones del genoma que también influyen en la producción de proteínas, llamadas regiones reguladoras. Estas áreas son las responsables de iniciar y controlar este proceso de producción, y por tanto la existencia de mutaciones en ellas puede alterar el funcionamiento del organismo, dando lugar a enfermedades.

En la actualidad hay poca información de este tipo disponible, ya que la mayoría de los estudios se han centrado en las anteriormente mencionadas regiones codificantes del genoma. Existen algunas herramientas como las citadas en el apartado anterior, aunque su alcance es limitado. A continuación se describe **IntuitSNP**, una herramienta para profundizar en el conocimiento de la influencia de las mutaciones en regiones reguladoras, con el fin último de estudiar el impacto de las variaciones individuales sobre los mecanismos de regulación genética para contribuir al avance de la Medicina Personalizada.

La figura 3.1 muestra un esquema de los distintos elementos que integran IntuitSNP. Cada elemento se describe en mayor detalle posteriormente.

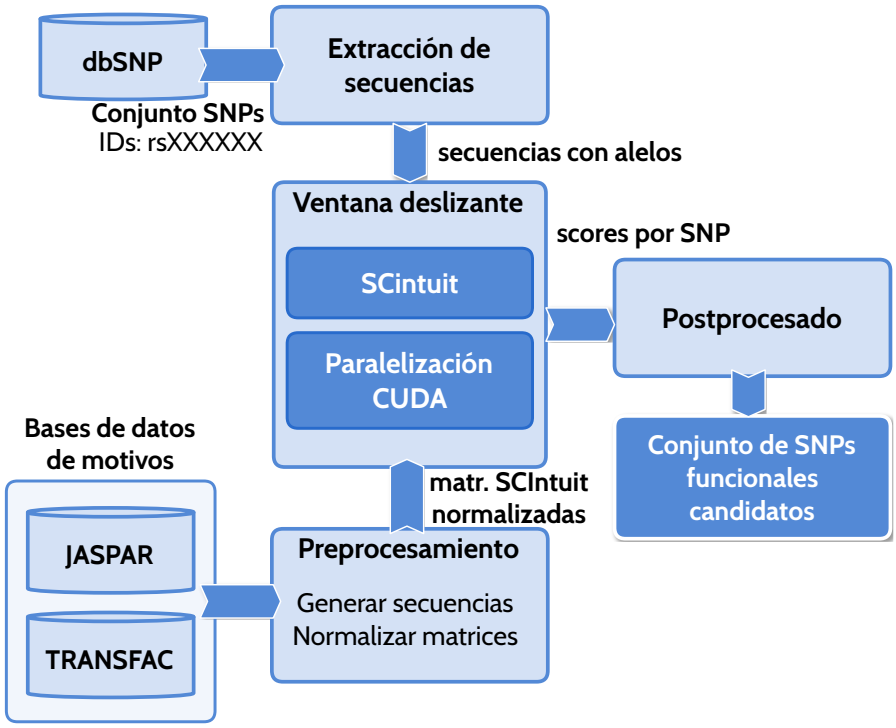


Figura 3.1: *Partes que componen el análisis automatizado realizado por IntuitSNP. Las secuencias alrededor de los SNPs a partir de sus identificadores de dbSNP son extraídas del genoma de referencia. Por otra parte, las matrices de JASPAR y TRANSFAC son preprocesadas para obtener las secuencias que no estén disponibles y normalizar las matrices de frecuencias. Estos datos de entrada son utilizados por el algoritmo de ventana deslizante, el cual hace uso de SC_{intuit} de forma paralela mediante CUDA. Los resultados son posteriormente recuperados y formateados para obtener scores por alelos, secuencias y motivos correspondientes así como la diferencia de score entre alelos.*

3.2.1. Preprocesamiento de los datos

Cuando se trabaja en un campo como la bioinformática, el cual está relacionado con áreas como la medicina y la biología, es frecuente tratar con datos que involucran incertidumbre, imprecisión, valores perdidos, redundancia o problemas de escalas y normalización. No todos estos problemas se pueden eliminar por completo. En casos como la incertidumbre o la imprecisión, muchas veces propiedades inherentes a la materia, será necesario conocer su origen y tratar los datos acorde con el problema que se desea resolver. Como ejemplo en este caso, la incertidumbre y la imprecisión son corregidas en parte a través de la aproximación al problema mediante medidas difusas.

Sin embargo, entre los problemas citados anteriormente, muchos pueden ser evitados o parcialmente reducidos mediante un adecuado preprocesamiento de los datos. Para la realización de este trabajo ha sido necesario un preprocesamiento de las matrices utilizadas en TRANSFAC y en JASPAR, ya que se han encontrado casos de valores perdidos y problemas de escalas que se detallan a continuación.

Diferentes tipos de matrices PWM. En las bases de datos de motivos, las matrices de frecuencias pueden expresarse de diferentes formas. Esto se debe, entre otras cosas, al tipo de experimento que las generó, ya que en determinados casos es imposible o ineficiente proporcionar una matriz de frecuencias absolutas (métodos estadísticos, números muy grandes en términos absolutos). Principalmente se encuentran varios tipos de matrices diferentes:

- **Matriz de frecuencias absolutas, o FSM**. Esta matriz tiene para cada nucleótido y cada posición del motivo el número de veces que apareció dicho nucleótido en esa posición, en términos absolutos.
- **Matriz de frecuencias relativas normalizada**. Contiene para cada nucleótido y posición, la frecuencia relativa de dicho nuncleótido en esa posición. Los valores se encuentran entre 0 y 1 y para cada posición siempre suman uno.
- **Matriz de frecuencias relativas no normalizada**. Contiene para cada nucleótido y posición, una frecuencia relativa, aunque en este caso el

intervalo es en $[0, r]$, siendo r un número real positivo. Para cada fila, los valores siempre suman r.

Este problema se soluciona mediante una normalización de todas las matrices al segundo formato, es decir, el de las matrices de frecuencias relativas normalizadas.

Secuencias generadoras de la matriz PWM no disponibles o incompletas. Para los casos en los que las secuencias generadoras de la matriz PWM no están disponibles, se genera un conjunto de secuencias basadas en la matriz PWM con el fin de poder aplicar SC_{intuit}. La calidad de los scores será peor en estos casos debido a la imposibilidad de utilizar la interdependencia entre columnas de la que se beneficia la medida SC_{intuit}, pero permitirá obtener resultados en caso de que la similitud secuencia-motivo sea más evidente, evitando excluir estos motivos del análisis.

3.2.2. Descripción del método de scoring

Con el objetivo de determinar si el SNP de consulta aparece en un factor de transcripción o no, se plantea un enfoque de ventana deslizante. Para cada motivo de longitud n disponible, la idea es obtener el score SC_{intuit} de todas las posibles secuencias de longitud n que contienen al SNP. Esto equivale a deslizar una ventana de longitud n centrada en el SNP, conservando el mejor *score*.

Además hay que tener en consideración la estructura de doble hélice del ADN. Esto quiere decir que también se compara con la ventana deslizante sobre la secuencia inversa complementaria.

3.2.2.1. Coste computacional del algoritmo de ventana deslizante

Para cada SNP y motivo, se requiere hacer el cálculo de SC_{intuit} $4 \cdot l$ veces, donde l es la longitud del motivo, ya que se realiza la ventana deslizante en la secuencia mutada y sin mutar, para las dos secuencias complementarias. Por tanto, el coste de ejecutar la ventana deslizante para un SNP determinado con un motivo concreto es de $4 \cdot l \cdot C_{intuit}$, donde C_{intuit} es el coste de ejecutar SC_{intuit}. Sea m la

máxima longitud posible de un motivo. Entonces el coste para un SNP y motivo está acotado por $4 \cdot m \cdot C_{intuit}$.

Esta operación además se repite para cada SNP para todos los motivos. Si N es el numero total de motivos en JASPAR y TRASNFAC y S el número de SNPs a procesar, el coste computacional para obtener los resultados será de $O(4 \cdot N \cdot S \cdot C_{intuit})$.

C_{intuit} es una operación cuyo coste depende de la longitud del motivo y el número de secuencias disponibles para este. Para motivos cortos (10bp) no llega al segundo. En general, para motivos de longitud media (15-20bp) el coste ronda los 2, 3 segundos y con los motivos más grandes de TRANSFAC en la actualidad (de longitud 30bp), el coste de ejecución se dispara a cerca de 13 segundos[3]. Además, se ha medido el tiempo de ejecución del algoritmo secuencial para todos los motivos en la versión actual de TRANSFAC. De aproximadamente 2500 motivos, se utilizan únicamente los que pertenecen a la especie humana (un total de 1003 motivos). Ejecutando el algoritmo secuencialmente se obtiene un tiempo medio de ejecución de aproximadamente 38 minutos por SNP[3].

Aunque el coste del proceso es relativamente asequible para evaluar un conjunto reducido de SNPs, puede suponer el coste de horas cuando se compara con la totalidad de motivos en TRANSFAC y JASPAR y especialmente en el caso de cantidades grandes de SNPs. Es por ello que se ha propuesto una paralelización del código secuencial mediante GPU, acelerando de forma considerable la ejecución del software desarrollado. En la siguiente sección se describe brevemente el proceso realizado, así como los tiempos de ejecución.

Paralelización del algoritmo

El uso de procesamiento mediante GPUs en bioinformática no es una novedad. Son candidatas idóneas para procesamiento de grandes cantidades de datos, ya que su bajo coste hace relativamente factible disponer de una cantidad nada despreciable de GPUs. En 2005 se utilizaron para estudios filogenéticos

[3]Estas pruebas fueron sido realizadas en un portátil Sony VAIO con un Intel(R) Core(TM) i 7-2640M CPU a 2.80GHz, 6GB de RAM.

[262]. Desde entonces, se han utilizado en algunas aplicaciones bioinformáticas, incluyendo árboles sufijos [263, 264], e implementaciones de Smith-Waterman [265, 266]. Algunas haciendo uso de la biblioteca CUDA, de NVIDIA [267].

CUDA es una plataforma de computación paralela y modelo de programación creado por NVIDIA, que proporciona extensiones de lenguajes ya existentes (como C/C++ o Fortran) para programación paralela. El objetivo es ayudar a los programadores a centrarse en algoritmos paralelos, y no en los mecanismos del lenguaje de programación paralelo, los cuales abstrae CUDA. También existen bibliotecas para realización de tareas frecuentes ya paralelizadas, como es el caso de procesamiento de imagenes y señales mediante nPP[4] o cuFFT[5], o algebra lineal (cuBLAS[6], cuSPARSE[7]).

El modelo de programación CUDA. Los segmentos de código paralelizado se conocen como *kernels* en CUDA. En un momento dado sólo se puede estar ejecutando un kernel, pero muchas hebras estarán ejecutando el mismo. Una particularidad de CUDA es que las hebras que utiliza son muy ligeras, por lo que se produce muy poca sobrecarga a la hora de crearlas y el cambio es casi instantáneo. Esto permite a CUDA trabajar con miles de hebras. Para poder ejecutar estos *kernels*, se trasladan a la GPU. La GPU es denominada *device* en CUDA, y es donde se ejecutan los *kernels*. La CPU es denominada host. Una vez ejecutadas las operaciones pertinentes, la información necesaria debe volver al host o CPU para poder ser utilizada.

Todas las hebras en CUDA ejecutan el mismo código, y utilizan su propio identificador para obtener los datos adecuados. En general, en un algoritmo paralelo será interesante que se compartan datos entre las hebras, con el fin de disminuir la repetición de operaciones. Sin embargo, si se comparten datos entre todas las hebras, el sistema no será escalable. Por ello CUDA propone un *grid* de computación, el cual básicamente está compuesto de una serie de *bloques* de hebras. Las hebras dentro de cada bloque cooperan entre ellas

[4] Procesamiento de imágenes: http://developer.nvidia.com/npp
[5] CUDA Fast Fourier Transform Library: http://developer.nvidia.com/cufft
[6] Basic Linear Algebra Subroutines: http://developer.nvidia.com/cublas
[7] CUDA Sparse Matrix Library: http://developer.nvidia.com/cusparse

y tienen memoria compartida, pero no entre hebras de distintos bloques. El hardware es libre de ejecutar estos bloques en cualquier procesador de forma transparente al programador. La ejecución de kernels en CUDA tiene una serie de restricciones, que obligan a realizar unos pasos preliminares a la ejecución de los mismos, así como una estructuración posterior del resultado. Una de ellas es que los kernels no pueden acceder a memoria en el *host*. Por tanto, cualquier información que necesiten las hebras debe ser trasladada al device previamente a su ejecución. Para ello hay una serie de llamadas proporcionadas por CUDA: `cudaMalloc`, `cudaMemCpy`, `cudaFree`, las cuales reservan, copian y liberan memoria, respectivamente. Los *kernels*, además, están forzados a ser de tipo *void*, por lo que cualquier valor que tengan que devolver debe ser trasladado también a través de las llamadas de gestión de memoria entre host y device. No obstante, los parámetros de los kernels sí son copiados automáticamente al device. Sin embargo, el número de argumentos tiene que ser fijo, no se permiten variables estáticas ni funciones recursivas[8].

Consideraciones sobre el grid de computación elegido

Para el procesamiento paralelo del algoritmo de ventana deslizante hay que tener en cuenta que el reparto de trabajo debe ser suficientemente equitativo para que el rendimiento de la paralelización sea máximo. Además, dadas las características de la GPU y en particular de CUDA, este reparto debe ser previo a la ejecución. Esto quiere decir que no se puede crear nuevo trabajo dentro de las propias hebras. Por tanto, quedan excluidas bifurcaciones (fork) y recursividad de las posibilidades de implementación. En la nueva arquitectura presentada por NVIDIA, Kepler6, estos problemas son resueltos permitiendo una mayor flexibilidad de programación. Sin embargo, la arquitectura presente en las tarjetas gráficas disponibles en nuestro servidor es anterior y responde a la filosofía ya mencionada. Teniendo en cuenta esto, también hay que tener en cuenta que las hebras de un mismo bloque pueden compartir información entre ellas, pero no con

[8] En las versiones más recientes de CUDA estas restricciones son menores, pero eran las vigentes al momento de implementación de IntuitSNP.

hebras de otros bloques. Además, el grid de computación ha de ser bidimensional, aunque los bloques pueden ser tridimensionales. Cada bloque consta de 512 hebras, donde la dimensión Z indica si la secuencia que se está analizando es mutada o sin mutar. Con respecto al eje X, se reparten las secuencias de consulta a analizar (los SNPs) y a lo largo del eje Y se reparten los motivos de JASPAR y TRANSFAC con los que comparar, de forma que cada hebra efectúa ventana deslizante sobre una secuencia (mutada o sin mutar) y un motivo concreto. Los scores son almacenados en una estructura de datos que después es devuelta a la CPU para su procesamiento. Otra observación relevante sobre la optimización del código reside en la transformación de las estructuras de datos que corresponden a matrices a una única dimensión, de forma que todas las matrices se convierten en vectores. Esto aumenta la eficiencia a la hora de trasladar datos de memoria de CPU a GPU y viceversa.

3.2.3. Caso de estudio

En este capítulo se aplica la propuesta desarrollada a un caso real y se evalúan los resultados obtenidos, comparándolos con los obtenidos por is-rSNP [152] y sTRAP [153]. En primer lugar, se hace una descripción detallada del caso de estudio propuesto, así como de los datos utilizados. A continuación se analizan los resultados obtenidos por la herramienta propuesta, **IntuitSNP**. Posteriormente se analizan los resultados obtenidos con las herramientas is-rSNP y sTRAP respectivamente, los cuales son comparados con los resultados de IntuitSNP.

Con el fin de validar los resultados obtenidos con IntuitSNP comparándolos con los de otras herramientas de reciente aparición como son is-rSNP y sTRAP, se ha utilizado la base de datos ORegAnno [151]. Esta base de datos contiene anotaciones sobre SNPs en regiones reguladoras de ADN. De ORegAnno se extrajeron las anotaciones pertenecientes a la especie *Homo sapiens*, para las cuales existe evidencia experimental de unión de un factor de transcripción a esa secuencia de ADN, siendo éste además especificado en la entrada. Además, se exige también que el SNP especificado aparezca en la secuencia, y su posición e identificador aparecen en dbSNP. Por último, se conoce cierto grado de relación

entre esos SNPs en TFBSs asociados a desarrollo de enfermedades, por lo que son SNPs de interés clínico.

En la tabla 3.1 se detallan los datos utilizados para las pruebas. En ella figura el identificador de la entrada en ORegAnno, así como el identificador del SNP en dbSNP. Además, se proporcionan las coordenadas del SNP (cromosoma y posición), junto con el gen regulado por ellos, el factor de transcripción cuya unión se ha verificado experimentalmente y el cambio de base que representa el SNP.

Tabla 3.1: *SNPs de ORegAnno utilizados para el caso de estudio. Para cada entrada se muestra su identificador en ORegAnno, las coordenadas en el genoma de referencia del SNP mediante la especificación del cromosoma y la posición, el gen al que pertenecen, el factor de transcripción al que afectan, junto con el identificador en dbSNP del SNP y la variación que se produce en él. El identificador de ORegAnno se construye OREGXXXXXXX donde XXXXXXX es el número que aparece en la tabla.*

ORegAnno	Cr.	Posición	Gen	TF	dbSNP ID	SNP
0001971	1	21617245	ECE1	E2F2	rs213045	C/A
0000410	1	159174683	FY	GATA1	rs2814778	T/C
0000092	1	159272060	FCER1A	GATA1	rs2251746	T/C
0002257	1	172627498	FASLG	CEBPB	rs763110	C/T
0000409	1	230849886	AGT	ERE/MLTF	rs5050	A/C
0002195	3	27764623	EOMES	TP53	rs3806624	T/C
0000005	4	74607055	IL8	CEBPB	rs2227306	T/C
0000403	4	156832027	TDO2	YY1	rs16998970	G/A
0000404	4	156832030	TDO2	YY1	rs13434811	G/T
0001935	5	52285077	ITGA2	Sp1 ind.	rs27646	C/G
0001934	5	52285117	ITGA2	Sp1/Sp3	rs28095	T/C
0000090	5	140012916	CD14	Sp fam	rs2569190	T/C
0001924	6	22304204	PRL	GATA-rel.	rs1341239	T/G
0000110	6	31542963	TNF	POU2F1	rs1800750	G/A
0000087	7	55086755	EGFR	SP1	rs712829	G/T

0001933	8	19796671	LPL	Sp1/Sp3	rs1800590	T/G
0000103	8	42072438	PLAT	Sp1/Sp3	rs2020918	T/C
0002196	11	4114539	RRM1	TP53	rs1465952	C/T
0002197	11	62009281	SCGB1D2	TP53	rs2232945	G/A
0002306	11	102745791	MMP12	JUN	rs2276109	A/G
0001932	12	15038919	MGP	JUN	rs1800802	T/C
0000095	12	21525704	IAPP	CREB1	rs11836625	G/A
0000097	12	48302545	VDR	CDX2	rs11568820	G/A
0001909	12	69202580	MDM2	SP1	rs2279744	T/G
0002193	13	111804830	ARHGEF7	TP53	rs1658728	G/T
0001899	16	55511806	MMP2	SP1	rs243865	C/T
0001947	16	56995236	CETP	Sp1/Sp3	rs1800775	C/A
0000313	17	28564346	SLC6A4	TFAP2A	rs25531	A/G
0000007	17	34207780	CCL5	GATA fam.	rs2107538	G/A
0001887	17	56358762	MPO	SP1	rs2333227	G/A
0000089	17	64225529	APOH	TFIID	rs8178822	C/A
0001907	17	71165292	SSTR2	NF1	rs998571	A/G
0002194	18	49864861	DCC	TP53	rs934345	C/G
0000075	19	7733793	RETN	Sp1/Sp3	rs1862513	G/C
0002199	19	40932040	SERTAD1	TP53	rs268682	C/G
0002192	21	46492270	ADARB1	TP53	rs2838769	G/A
0002200	X	12923681	TLR8	TP53	rs3761624	A/G

3.2.3.1. Resultados de la herramienta IntuitSNP

En esta sección se detallan los resultados obtenidos tras la aplicación de nuestra herramienta a los datos presentados con las bases de datos de motivos de TRANSFAC y JASPAR.

Dado que ORegAnno proporciona el factor de transcripción concreto que se asocia a la secuencia, se puede corroborar si un resultado es acertado o no. Por

lo general, los nombres de los motivos se corresponden con los de los factores de transcripción que se unen a ellos. Se ha decidido considerar acierto si el resultado pertenece a la misma familia.

Criterios de evaluación de los resultados. Se han realizado experimentos para distintos umbrales de corte de SC_{intuit}: 0.65, 0.70, 0.75, 0.80[9]. Para cada entrada se tienen en cuenta los siguientes criterios:

- **Acierto.** Si aparece un resultado correcto (motivo perteneciente a la familia) por encima del umbral.
- **Acierto exacto.** Si aparece un resultado exacto (motivo perteneciente exactamente al mismo TF).
- **Número de aciertos relacionados.** Número de resultados de la familia por encima del umbral de corte.
- **Rank absoluto.** Posición en la que aparece el primer resultado correcto.
- **Rank relativo.** Posición en la que aparece el primer resultado correcto normalizado al número de candidatos. Este valor, en $[0, 1]$ será por tanto mejor cuanto más cercano a 0.
- **Rank exacto absoluto.** Posición en la que aparece el primer resultado exacto.
- **Rank exacto relativo.** Posición en la que aparece el primer resultado exacto normalizado al número de candidatos. Este valor, en $[0, 1]$ será por tanto mejor cuanto más cercano a 0.
- **Proporción aciertos/candidatos.** Proporción de resultados correctos con respecto al número de candidatos por encima del umbral.

En la tabla 3.2 se puede observar el número de SNPs correctamente detectados según el umbral de corte de la medida SC_{intuit} junto con una serie de medidas relacionadas, como el número medio de aciertos relacionados para cada SNP, el porcentaje de los aciertos en general que representa y el número medio de candidatos obtenidos por SNP para cada umbral. Se puede apreciar que en los

[9] Para umbrales más altos el número de resultados resulta demasiado escaso para ser significativo.

Tabla 3.2: *Tabla resumen de resultados de IntuitSNP.*

Umbral	#Hits	#Hits rel.	% Hits rel.	# Candidatos
0.65	33	6.81	11.40 %	59.72
0.70	28	5.71	17.15 %	33.29
0.75	27	4.44	28.08 %	15.81
0.80	14	5.07	66.79 %	7.59

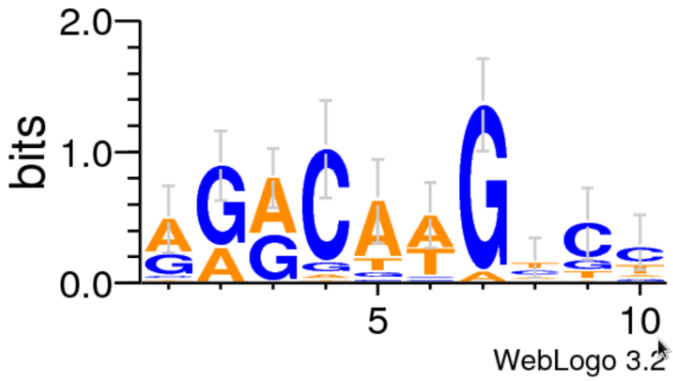


Figura 3.2: *Logo para el motivo M00761 en TRANSFAC.*

valores entre 0.70 y 0.75 está el umbral óptimo, ya que el número de aciertos es elevado ($\sim 70-75\,\%$), el número de aciertos relacionados también lo es y representa casi una quinta parte de los resultados obtenidos. Se trata pues de un conjunto de predicciones de elevada calidad y tamaño manejable para ser estudiado más en profundidad por un investigador interesado en los mecanismos reguladores del SNP en cuestión.

En la tabla 3.4 se puede observar una lista comparativa detallada para cada SNP de la posición del mejor resultado, la del mejor resultado exacto y sus valores relativos, en $[0,1]$, donde la normalización se ha hecho en base al número medio de resultados obtenidos. En la última línea de la tabla se observan los valores medios de estos rankings, teniendo en cuenta que los casos en los que no se ha

obtenido ningún resultado acertado (- en la tabla) no han sido contabilizados.

Observando la tabla 3.4 también se puede observar que, en general, para los resultados que se consideran aciertos en la tabla 3.2 para el umbral 0.75 (sombreado en la tabla 3.4), los aciertos aparecen entre los 10 primeros casos, siendo entre los 10 y los 20 poco frecuente en general, y los resultados que no son aciertos aparecen a partir de la posición 20. Esto sugiere además que los casos en los que no se acierta no se deben a la elección del umbral, sino a que estos casos concretos son difíciles de clasificar. A continuación se observan en detalle algunos ejemplos de este tipo.

Tomando el elemento sexto de la tabla 3.4 (en negrita), se puede observar que es un resultado muy pobre. Por encima del resultado correcto aparecen otros 40 motivos de TRANSFAC y JASPAR. Según SC_{intuit}, de las posibles soluciones acertadas, el motivo que se parece más a la secuencia de consulta es M00761, correspondiente a p53 decamer, con un score de 0.6668. Observando otros resultados, además, como el de `rs934345` (fila 33 en la tabla) y `rs1658728` (fila 25), también se obtienen scores bajos para la misma matriz, que es la que se supone que se parece más al resultado esperado.

Como medida inicial para intentar dilucidar el motivo de estos resultados, se ha realizado una representación gráfica del motivo M00761 de TRANSFAC, el cual se puede ver en la figura 3.2.

A primera vista, es un motivo que no tiene demasiadas posiciones conservadas, a excepción de la séptima posición, donde aparece una G, y las posiciones tercera y quinta, aunque éstas algo menos conservadas. Una posible contrapartida de la medida SC_{intuit} que podría sugerir la visualización de la figura 3.2 es que, si el motivo no es demasiado conservado, los scores obtenidos pueden no ser muy elevados, debido a la propia imprecisión de la información que se posee. Para comprobar esto, se ha realizado un estudio en más detalle que se puede observar en la tabla 3.3.

En la primera fila de la tabla aparece el score obtenido para una secuencia que, a simple vista, es la que más se debería parecer al logo. Viene de tomar la base que más aparece en cada posición, independientemente de si está conservada

o no. Este resultado despeja la primera duda, sobre si para motivos imprecisos SC_{intuit} no obtiene scores altos. Se puede obtener un score alto, incluso máximo, en este caso. Por otro lado, en las siguientes 4 filas aparecen resultados reales obtenidos para los SNPs concretos que aparecen en la descripción. En casos como las filas 2 y 5, son resultados que deberían haber obtenido un score más alto para considerarse aciertos, ya que el factor de transcripción asociado era, efectivamente, el correspondiente al motivo que se analiza (p53).

La primera impresión es que es difícil obtener un score alto en la secuencia en cuestión, ya que en el caso 2 hay 6 bases distintas del mejor resultado, y en el caso 6 cuatro sustituciones. Sin embargo, si se analiza con más detalle el caso 2 con respecto al logo, se pueden encontrar varios detalles importantes. El primero es que los tres primeros cambios (AGA por GAG), así como el cambio en la sexta posición, se realizan por otras bases cuya aparición es relativamente significativa, aunque no sean la base que más aparece. Además, en el caso de las dos últimas posiciones, estas tienen un contenido de información bajo, por lo que la alteración del score debería ser baja en principio.

Para intentar averiguar con más detalle cómo influyen las variaciones en este motivo, se han realizado sustituciones de una única posición en la forma que se muestra en la tabla 3.3 a partir de la fila 6. Las variaciones en la posición más conservada (líneas 10 y 11 de la tabla) modifican considerablemente el score, como era de esperar. Por último, se ha probado a sustituir varias posiciones no muy conservadas a la vez, para estudiar cómo influye su variación. La aparición de varios cambios a la vez, aunque estén en posiciones no demasiado conservadas, generan diferencias de entre 0.35 y 0.45 con el mejor score, obteniendo un resultado inesperadamente bajo cuando se observa a simple vista el logo y la secuencia.

Dado que los resultados experimentales anotados en ORegAnno indican que, aunque las secuencias de las filas 2 a 6 sean relativamente diferentes al logo, el factor de transcripción se sigue uniendo en esa posición, se plantea el debate sobre si las posiciones no conservadas deben influir tanto en el score final. Claramente deben tener cierta repercusión, ya que influye de manera diferente cada variación

Tabla 3.3: *Tabla de variaciones de secuencias para motivo M00761. Las bases que son distintas del mejor caso (primera fila de la tabla) se muestran coloreadas.*

	Secuencia	Score	Dif.	Descripción
1	AGACAAGTCC	0.9937		Secuencia más parecida al logo posible
2	GAGCATGTTC	0.6899	0.303	SNP rs934345
3	AAGCAAGTGC	0.7633	0.230	SNP rs1658728
4	GGGCATGTGC	0.7899	0.204	SNP rs1465952
5	GGGCAAGTTG	0.6668	0.327	SNP rs1465952
6	AGACAAGTAC	0.8393	0.154	Sustitución poco conservada
7	AGACAAGACC	0.9205	0.073	Sustitución casi nada conservada
8	AGACCAGTCC	0.8317	0.162	Sustitución medianamente conservada
9	AGACTAGTCC	0.8633	0.130	Sustitución medianamente conservada
10	AGACTAATCC	0.7082	0.285	Sustitución posición más conservada
11	AGACTACTCC	0.6962	0.298	Sustitución posición más conservada
12	AGGCAAGTCC	0.9605	0.033	Sustitución posición de conserv. media por otra de proporción similar
13	AGACTAGATT	0.6455	0.350	Varias sustituciones poco conservadas
14	AGACCAGAAA	0.5304	0.463	Varias sustituciones poco conservadas

en una determinada posición, aunque en el logo sea menos relevante por tener un contenido de información más bajo. La cuestión aquí parece radicar también en la aditividad del error.

Para ilustrar mejor estos resultados, se ha realizado un pequeño estudio de la matriz VGATAQ6 (M00789) en TRANSFAC. Se puede observar en el logo de la figura 3.3 que las bases más relevantes son las que se encuentran en las posiciones 2 a 5, que dan nombre al motivo. Sería de esperar que la afinidad química de la secuencia con el factor de transcripción en cuestión se debiera principalmente a estas posiciones. Sin embargo, si las posiciones menos conservadas 1, 6 y 7 varían, se pueden obtener scores tan bajos como los que se detallan en la tabla 3.5.

Esto hace pensar que, aunque los resultados con SC_{intuit} resultan coherentes, la acumulación de errores en posiciones muy poco conservadas puede derivar en puntuaciones bajas que no reflejan apropiadamente el parecido de la secuencia al motivo. Se hace necesario por tanto realizar más estudios sobre variaciones en la

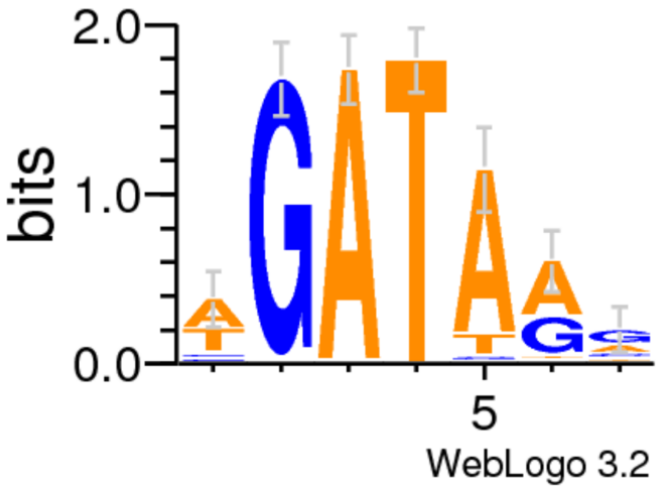


Figura 3.3: *Logo para el motivo M00789 en TRANSFAC.*

medida que permitan optimizar su eficacia.

3.2.3.2. Comparativa de resultados

En la tabla 3.6 se detalla una comparativa entre las posiciones del primer resultado para la aplicación de IntuitSNP, sTRAP [153] y is-rSNP [152]. Se ha añadido una columna que indica el resultado mejor en caso de haberlo, y un guión en caso de empate. Al final aparece un total de SNPs en los que el resultado de la medida propuesta, $IntuitSNP$, es mejor. El segundo total incluye empates. Se puede observar que los resultados de la medida propuesta son algo mejores que los de la medida proporcionada por is-rSNP y muy similares a sTRAP.

Los resultados obtenidos sugieren que, aún siendo ligeramente mejores los resultados de la herramienta propuesta, existe aún margen para la mejora de la medida utilizada, IntuitSNP. En los SNPs en los que la herramienta propuesta no obtiene mejores resultados que el resto, suele haber algún resultado bueno en las otras medidas. Esto hace que resulte interesante plantear si las medidas is-rSNP

Tabla 3.4: *Rank obtenido por el primer resultado para cada SNP. En sombreado los valores considerados como aciertos para el valor de umbral de SC_{intuit} en 0.75.*

		Rank		Rank exacto		
	dbSNP ID	**absoluto**	**relativo**	**absoluto**	**relativo**	**# hits**
1	rs213045	2	0.044	-	-	45
2	rs2814778	1	0.013	1	0.013	78
3	rs2251746	4	0.078	15	0.294	51
4	rs763110	2	0.034	2	0.034	58
5	rs5050	2	0.034	2	0.034	59
6	rs 3806624	41	0.788	42	0.808	52
7	rs2227306	6	0.076	6	0.076	79
8	rs16998970	8	0.098	8	0.098	82
9	rs13434811	18	0.261	18	0.261	69
10	rs27646	6	0.140	6	0.140	43
11	rs28095	2	0.037	2	0.037	54
12	rs2569190	12	0.235	12	0.235	51
13	rs1341239	3	0.039	3	0.039	76
14	rs1800750	28	0.304	-	-	92
15	rs712829	7	0.101	7	0.101	69
16	rs1800590	3	0.050	6	0.100	60
17	rs2020918	5	0.093	5	0.093	54
18	rs1465952	9	0.180	28	0.560	50
19	rs2232945	5	0.070	5	0.070	71
20	rs2276109	1	0.022	-	-	45
21	rs1800802	24	0.471	-	-	51
22	rs11836625	-	-	-	-	70
23	rs11568820	2	0.031	2	0.031	65
24	rs2279744	58	1.000	-	-	58
25	rs1658728	12	0.167	40	0.556	72
26	rs243865	6	0.118	42	0.824	51
27	rs1800775	42	0.894	42	0.894	47
28	rs25531	4	0.095	4	0.095	42
29	rs2107538	1	0.014	1	0.014	72
30	rs2333227	26	0.274	26	0.274	95
31	rs8178822	-	-	-	-	48
32	rs998571	26	0.426	-	-	61
33	rs934345	30	0.625	34	0.708	48
34	rs1862513	7	0.146	-	-	48
35	rs268682	29	0.763	29	0.763	38
36	rs2838769	-	-	-	-	52
37	rs3761624	3	0.056	3	0.056	54
	Medias	13.182	0.236	11.848	0.218	59.730

Tabla 3.5: *Tabla de variaciones de secuencias para motivo M00789. Las bases que son distintas del mejor caso (primera fila de la tabla) se muestran coloreadas.*

	Secuencia	Score	Dif.	Descripción
1	AGATAAG	1.000	-	Mejor secuencia posible
2	AGATAGG	0.930	0.070	Sustitución posición conserv. media
3	AGATAAT	0.837	0.163	Sustitución posición poco conservada
4	AGATAGT	0.779	0.221	Sustitución dos posiciones
5	TGATAGT	0.746	0.254	Sustitución 3 posiciones
6	CGATAGT	0.636	0.364	poco conservadas
7	AGATATT	0.663	0.337	Sustitución 2 posiciones poco conservadas por base menos conserv.

y sTRAP tienen en cuenta alguna característica que IntuitSNP no contempla. Por otro lado, aún en los casos en los que los resultados de la herramienta propuesta son peores, el candidato correcto seleccionado por la herramienta se mantiene en una posición razonable en la lista, esto es, por encima de los 50 primeros en casi todos los casos.

3.2.4. Escalado de matrices PWM para la comparación de afinidad entre diferentes factores de transcripción

Uno de los problemas que se observan en los resultados anteriores es la dificultad para comparar entre afinidades de unión de motivos distintos, dada la heterogeneidad de dichos motivos. Encontramos motivos para los que las medidas de similitud secuencia-motivo tienden a tener puntuaciones más altas y, análogamente, motivos que resultan en puntuaciones más bajas, como se ha observado en el caso de estudio. Además, conceptos como la probabilidad de unión de un factor de transcripción a una secuencia o similitud secuencia-motivo son difíciles de aplicar al modelado de los procesos dinámicos de unión de factores de transcripción. Se ha demostrado que la energía de unión de un factor de transcripción a una secuencia de ADN específica es proporcional a la similitud secuencia-motivo [268], estando esta proporcionalidad escalada por un factor λ.

Tabla 3.6: *Comparativa de posiciones del primer resultado correcto entre la herramienta propuesta, is-rSNP y sTRAP.*

	SNP	Intuit	is-rSNP	sTRAP	Mejor
1	rs213045	2	-	24	**Intuit**
2	rs2814778	1	38	2	**Intuit**
3	rs2251746	4	71	6	**Intuit**
4	rs763110	2	18	1	sTRAP
5	rs5050	2	1	10	rSNP
6	rs3806624	41	2	2	-
7	rs2227306	6	15	9	**Intuit**
8	rs16998970	8	-	12	**Intuit**
9	rs13434811	18	-	55	**Intuit**
10	rs27646	6	13	2	sTRAP
11	rs28095	2	6	8	**Intuit**
12	rs2569190	12	3	24	rSNP
13	rs1341239	3	25	12	**Intuit**
14	rs1800750	28	13	30	rSNP
15	rs712829	7	-	9	**Intuit**
16	rs1800590	3	15	72	**Intuit**
17	rs2020918	5	7	14	**Intuit**
18	rs1465952	9	3	6	rSNP
19	rs2232945	5	1	1	-
20	rs2276109	1	1	1	-
21	rs1800802	24	-	108	**Intuit**
22	rs11836625	-	57	27	sTRAP
23	rs11568820	2	7	22	**Intuit**
24	rs2279744	58	66	111	**Intuit**
25	rs1658728	12	1	1	-
26	rs243865	6	3	107	rSNP
27	rs1800775	42	-	35	sTRAP
28	rs25531	4	59	69	**Intuit**
29	rs2107538	1	-	3	**Intuit**
30	rs2333227	26	5	205	rSNP
31	rs8178822	-	-	215	sTRAP
32	rs998571	26	76	18	sTRAP
33	rs934345	30	29	3	sTRAP
34	rs1862513	7	2	139	rSNP
35	rs268682	29	151	1	sTRAP
36	rs2838769	-	-	24	sTRAP
37	rs3761624	3	199	4	**Intuit**
				Total	17
				(+ empates)	19

En esta sección se propone un método [269] para obtener este parámetro λ con el fin de obtener la energía o la fuerza de la unión a partir de un *score* PWM-secuencia.

En la mayoría de los estudios computacionales, el parámetro λ es ignorado, tomando como único dato el score secuencia motivo obtenido a partir de una matriz PWM y una secuencia concreta. En algunos casos [154, 270] se ponderan estos *scores* utilizando datos de experimentos ChIP-seq. Este enfoque puede generar problemas debido por un lado al ruido presente en los datos ChIP-seq, y especialmente al hecho de que la altura de un pico de ChIP-seq puede no ser proporcional al factor λ, perdiendo la especificidad para encontrar dicho factor [154].

Se describe por ello en esta sección un método para calcular λ utilizando PWMs existentes, tolerancia de error medio máximo y la distribución de scores de las matrices PWM en el genoma de un organismo específico. Este método sólo depende de la secuencia de ADN a comparar y de la composición de ADN de la especie objetivo. λ se calcula mediante una ecuación basada en teoría de energía de *mismatch*.

3.2.4.1. Ecuación de cálculo de λ

En base a la teoría de energía de *mismatch* [268], la energía de desequilibrio para un factor de transcripción en un lugar de unión j de la especie i en el genoma se puede expresar como:

$$E_{mismatch,i,j} = \frac{\Delta S_{i,j}}{\lambda_i} = \frac{S_{max,i} - S_{i,j}}{\lambda_i}$$

donde $S_{i,j}$ es el score de la matriz PWM en la posición j, $S_{max,i}$ es el máximo score posible de la matriz PWM en la especie i y λ_i es el factor de escalado que se pretende estimar. Nótese que esta *energía de mismatch* proviene de teoría de la información, y puede describirse también como *mismatch bits*.

Dada la utilidad de λ para estimar afinidad e incluso tiempo de unión, su cálculo es útil. Derivamos la ecuación basándonos en dos principios:

1. El top 0.1 % de los scores de PWMs en zonas intergénicas son considerados posibles TFBS reales [271]. Este top 0.1 % se ha adoptado en otros estudios [272].
2. La máxima energía de *mismatch* entre el motivo y la secuencia concreta es proporcional al contenido de información de la matriz PWM de su TF correspondiente. El contenido de información del motivo se define como [268]:

$$If = \sum_{k=1}^{L} \sum_{i \in \{A,C,G,T\}} p_{i,k} log_2 \frac{p_{i,k}}{f_i}$$

donde k es el k-ésimo nucleótido en el motivo PWM, f_i es la frecuencia a priori de cada nucleótido y $p_{i,k}$ es la frecuencia ajustada del nucleótido i en la posición k definida como:

$$p_{i,k} = \frac{\nu_{i,k} + f_i \dot{\mu}}{\sum_i \nu_{i,k} + \mu}$$

donde $\nu_{i,k}$ es la frecuencia del nucleótido i en la posición k del motivo.

La cota inferior de los lugares de unión potenciales se calcula a partir del top 0.1 % de los scores de PWM:

$$E_{maxMismatch,i} = \frac{S_{max,i} - S_{top0,1\%,i}}{\lambda_i}$$

donde $E_{maxMismatch,i}$ es la máxima tolerancia de energía de mismatch para la especie i. λ, por tanto, se calcula como:

$$\lambda_i = \frac{S_{max,i} - S_{top0,1\%,i}}{E_{maxMismatch,i}}$$

Este valor $E_{maxMismatch,i}$ se calcula a partir del contenido de información de la matriz PWM utilizando los datos experimentales para los TFs en [ ]:

$$E_{maxMismatch,i} = < E_{maxMismatch} > \times \frac{If_i}{< If >}$$

Donde $< E_{maxMismatch} >$ es la media de tolerancia a la máxima energía de mismatch, la cual es igual a 6 de acuerdo a [273], e $< If >$ representa

el contenido de información medio de las matrices PWM en [273], el cual toma el valor de 13.2 bits. La razón de este cálculo es que si el contenido de información es una buena representación de la especificidad de unión de un factor de transcripción, entonces la diferencia de energía de unión medida en bits entre lugares de unión fuertes y débiles deberá guardar cierta relación con la especificidad de unión de dicho TF. Debido al tamaño reducido de los datos utilizados, asumimos que esta relación es lineal.

3.2.4.2. Caso de estudio

La ecuación de cálculo de λ descrita en el apartado anterior se ha utilizado para estimar dicho valor para diferentes familias de factores de transcripción en diferentes especies. Utilizando los métodos de Berg *et al.* [268], la energía de unión de un factor de transcripción para un lugar de unión específico se puede calcular utilizando la ecuación del apartado anterior. Con la finalidad de mostrar las aplicaciones biológicas del parámetro λ, se muestra en la figura 3.4 un ejemplo de una zona del genoma de *Drosophila melanogaster* con alta densidad de TFBSs. La utilidad de las estimaciones de λ se puede ver en los TFBSs marcados con flechas en la imagen. El segundo TFBS tiene un score mayor, pese a que la energía de unión es más baja al tener en cuenta el factor λ. La tercera flecha muestra dos TFBSs solapantes en los que se produce la misma situación. Pese a que no existe evidencia experimental para la importancia relativa de los lugares de unión de factores de transcripción para este lugar específico, el ejemplo puede servir para ilustrar cómo pueden variar las hipótesis de afinidad de unión al transformar los scores de PWMs mediante un factor λ adecuado.

Este procedimiento ha sido utilizado también para calcular el valor λ para las matrices PWM presentes en JASPAR, agrupadas en tres categorías: los motivos de la especie *Saccharomyces cerevisiae*, *Drosophila melanogaster* y la familia de los vertebrados. Las distribuciones de valores de λ obtenidos son significativamente diferentes entre vertebrados y ambas especies, pero no entre sí.

En general, se ha observado la utilidad del cálculo del factor λ para refinar la estimación de afinidad de unión entre una secuencia y una matriz PWM. La

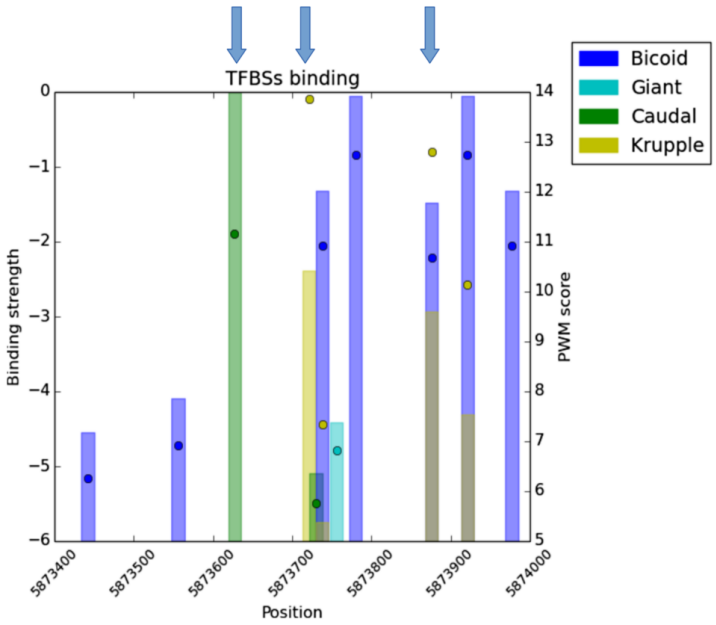


Figura 3.4: *Comparación entre valores de PWM score y energía de unión calculada a través del factor λ en la región del enhancer even-skipped stripe 1 en D. melanogaster. Los círcluos representan los scores PWM, mientras que las barras representan la energía de unión escalada a través del factor λ. Los lugares indicados con flechas representan casos en los que ambos valores difieren lo suficiente como para representar hipótesis alternativas. El color de los círculos y las barras indica el tipo de TFs correspondientes a las matrices PWM.*

aplicación de este método es además sencilla, ya que se basa en una fórmula de fácil aplicación. Por otro lado, este método no es aplicable a motivos de menos de 6 pares de bases, ya que las posibles combinaciones de motivos de longitud menor hacen que el conjunto de valores en el top 0.1 % sea prácticamente igual al valor máximo. Sin embargo, la mayoría de los TFBS en eucariotas tienen una longitud mayor, por lo que esta limitación no debería ser un problema en la mayoría de los casos, siempre y cuando el umbral siga siendo 0.1 %. Otra asunción de esta propuesta es que el rango de tolerancia a mismatch es proporcional al contenido de información de la matriz PWM. Esto quiere decir que para valores extremadamente

altos de contenido de información el valor λ podría ser menor de lo esperado. Es posible, además, que el umbral de corte de 0.1 % deba ser adaptado según el grupo de TFs utilizado. Sin embargo, esto podría ser difícil con la cantidad de datos actual. Por otro lado, este método también genera un contenido de información mayor para motivos más largos. Por último, el factor λ sólo será representativo de la energía de unión del TF al ADN si la matriz PWM es suficientemente fiable. No obstante, el factor λ puede proporcionar información sobre las propiedades de unión de los factores de transcripción. Se hace pues interesante incorporar el cálculo de este factor de ajuste a las medidas de scoring propuestas anteriormente.

3.3. Búsqueda computacional de módulos reguladores

Existen multitud de enfoques *in-silico* para la detección de CRMs, aunque, dada la complejidad combinatoria del problema, estos enfoques se pueden clasificar principalmente en tres categorías de acuerdo a un progresivo aumento del ámbito de aplicación de la herramienta [162].

Escáners. Este tipo de metodologías buscan secuencias que se adapten exactamente a un modelo de CRM previamente definido por el usuario. De este modo, se aplican principalmente a problemas bien caracterizados, en los que el usuario aporta un modelo muy específico del CRM a buscar. Este tipo de herramientas suelen contar con muchos parámetros a fijar, como puede ser el número de TFBSs que componen el CRM, un conjunto reducido de TFBSs entre los que seleccionar o un tamaño de secuencia. Cister [163], Cluster-Buster [164] o Stubb [165] pertenecen a esta categoría. Estas herramientas suelen ser mucho más rápidas, ya que las restricciones impuestas limitan significativamente el espacio de búsqueda.

Constructores. Estas herramientas relajan algunos de los parámetros de los scanners para ampliar el espacio de búsqueda y requieren por contrapartida una serie de secuencias relacionadas que guarden entre sí alguna relación regulatoria. Dado un conjunto de este tipo de secuencias, buscan características similares entre ellas que les permitan construir un CRM putativo. Estos métodos reducen el número de secuencias de entrada principalmente de dos formas: i) limitando el estudio a zonas promotoras de genes co-expresados o relacionados [166, 167]; o ii) centrándose en zonas conservadas desde el punto de vista evolutivo [168, 169]. Sin embargo, la reducción del espacio de búsqueda tiene un coste en ambos casos. Mientras las primeras podrían ignorar regiones menos evidentes pero de importancia reguladora, tales como intrones o secuencias alejadas de los genes regulados [274], las segundas pueden presentar problemas en cuanto al alineamiento de las secuencias, que suele ser requisito necesario para la construcción de CRMs que practican estos enfoques. Por otro lado, se ha demostrado en mamíferos y en especies *Drosophila* que entre uno y dos tercios de

las secuencias reguladoras identificadas no son conservadas incluso entre especies muy relacionadas [275, 276, 277]. Enfoques como INSECT [170] o CORECLUST [171] hacen uso de ambos tipos de relaciones entre secuencias para reducir el espacio de búsqueda.

Screeners. El último tipo de enfoques se ocupan de buscar CRMs por todo el genoma. Al contrario de los enfoques anteriores, este tipo de herramientas no asumen ningún tipo de característica sobre los CRMs, por lo que presentan un campo de aplicación más amplio. Sin embargo, el problema que tratan de resolver es más complejo. Tanto es así, que muchas de las herramientas que han sido categorizadas como *screeners* en [162] no engloban aún por completo el ADN no codificante. Por ejemplo, D-Light [278] no restringe el número de genes pero limita la búsqueda a las zonas promotoras de los genes, obviando el resto de ADN no codificante. Por otra parte, COPS [279] busca coocurrencias de TFBSs en secuencias validadas *in vivo*, requiriendo conocimiento previo y evidencia experimental para la búsqueda. Otras herramientas, tales como TraFac [280], PreMod [168] o EEL [169] también reducen el conjunto de secuencias de entrada a un conjunto de genes co-expresados o a una serie de secuencias ortólogas, por lo que entrarían más en la categoría de constructores.

Independientemente del ámbito al que se apliquen los métodos ya mencionados, todos los enfoques para la detección de CRMs adolecen de una gran complejidad combinatoria, la cual no sólo aumenta con la longitud de las secuencias a analizar, sino también con el número de TFBSs candidatos y el tamaño del CRM. Por ello, algunas metodologías limitan la variabilidad en alguna de las entradas, especialmente en el caso del número de factores de transcripción involucrados en el CRM, llegando en muchos casos a buscar pares de TFs cooperantes [281, 282]. Estos enfoques obvian módulos de regulación mayores en los que los TFs se relacionan de forma más compleja [159].

Otros enfoques tratan de lidiar con esta complejidad combinatoria reduciendo la longitud de secuencia a analizar, restringiendo dichas secuencias a las zonas promotoras de un número reducido de genes co-expresados o a regiones conservadas, como ya se ha mencionado con anterioridad. Con las herramientas

existentes actualmente, un investigador no puede efectuar una búsqueda de módulos reguladores en *cis* sin proporcionar como mínimo un conjunto específico de motivos candidatos como entrada. Esto podría no ser posible si no se tiene conocimiento previo alguno. En este sentido, la probabilidad de elegir el subconjunto adecuado de TFBSs para la búsqueda de CRMs sería equivalente a la de elegir un subconjunto aleatorio de entre todos los TFBSs conocidos del organismo, la cual es muy pequeña. Lo mismo aplica a los enfoques que utilizan secuencias ortólogas a un cierto conjunto de genes o un conjunto de genes co-regulados. Todo este tipo de metodologías son útiles para la investigación en procesos donde ya existe conocimiento previo. Sin embargo, no permiten realizar investigación de tipo exploratorio.

Por otro lado, las herramientas de predicción de CRMs suelen ir de la mano de herramientas de detección de TFBSs. Es por ello que normalmente las herramientas de detección de CRMs requieren como entrada un conjunto de PWMs que puedan ser utilizadas para obtener un conjunto inicial de TFBSs putativos. En este caso, el rendimiento de estas herramientas es muy dependiente de la calidad de la predicción inicial de TFBSs.

Además, la ubicación y longitud exacta de las regiones reguladoras todavía no se conoce con precisión. Se sabe que los CRMs pueden abarcar varios centenares de pares de bases, pero su longitud exacta es variable. Dentro de dichos módulos, la posición de los TFBSs también es vaga e inexacta. Las herramientas *in silico* para la predicción de TFBSs a veces sobrepasan la cantidad aceptable de falsos positivos debido a la complejidad biológica del propio problema, ya que las secuencias a buscar son muy cortas ($\approx 5 - 30bp$) en comparación con el tamaño de los genomas de entrada, y la aparición de secuencias verdaderamente similares a las de los TFBSs buscados que sin embargo no actúan como tales lugares de unión es esperable. Esta falta de especificidad intrínseca en los TFBSs parece sugerir que existe una maquinaria más compleja y reglas más sofisticadas para gobernar los procesos de regulación dirigidos por los factores de transcripción [159].

Cualquier herramienta que pretenda ayudar en la predicción de CRMs debe tener esta problemática en consideración y ser capaz de manejar la imprecisión

y la incertidumbre inevitablemente presentes en los datos. A pesar de que las metodologías difusas han superado en muchas aplicaciones a las metodologías *crisp* equivalentes, no han sido apenas utilizadas en la detección de CRMs.

3.4. CisMiner: Metodología difusa para la búsqueda de módulos reguladores

En este apartado se describe CisMiner [258], una metodología propuesta para búsqueda de CRMs que trata de superar los problemas mencionados, mediante la aplicación de un algoritmo difuso de minería de itemsets frecuentes. CisMiner busca módulos reguladores a través de todo el genoma no codificante de un organismo para módulos de un número arbitrario de TFBSs. Dado un conjunto de TFBSs, CisMiner aplica un clústering difuso de TFBSs cercanos y analiza dichos conjuntos difusos mediante la aplicación del algoritmo Top-Down Fuzzy Frequent Pattern Tree.

3.4.1. Descripción general de la metodología

CisMiner implementa un pipeline de análisis de datos que toma como entrada un conjunto de TFBSs como entrada y obtiene a partir de ellos un conjunto de co-ocurrencias significativas de conjuntos de TFBSs de cualquier tamaño. Con este fin, se ejecutan los siguientes pasos. Primero se ejecuta un clústering difuso de TFBSs en todo el genoma para obtener conjuntos de TFBSs cercanos. Estos conjuntos difusos se incluyen como itemsets en una base de datos transaccional difusa, sobre la que después se ejecutará un algoritmo difuso de itemset mining para obtener conjuntos frecuentes de co-ocurrencias de TFBSs. El algoritmo Fuzzy Frequent-Pattern Tree (Fuzzy FP-Tree) se aplica con este fin, ya que ha demostrado buen rendimiento para grandes volúmenes de datos [283]. El conjunto de TFBSs de entrada que se le proporciona a CisMiner puede ser obtenido tanto *in silico* como *in vivo*, permitiendo a los investigadores combinar esta metodología con cualquier herramienta de predicción de TFBSs para realizar una una búsqueda previa y obtener el conjunto de TFBSs candidatos. La figura 3.5 representa el flujo de procesamiento de datos en la herramienta CisMiner.

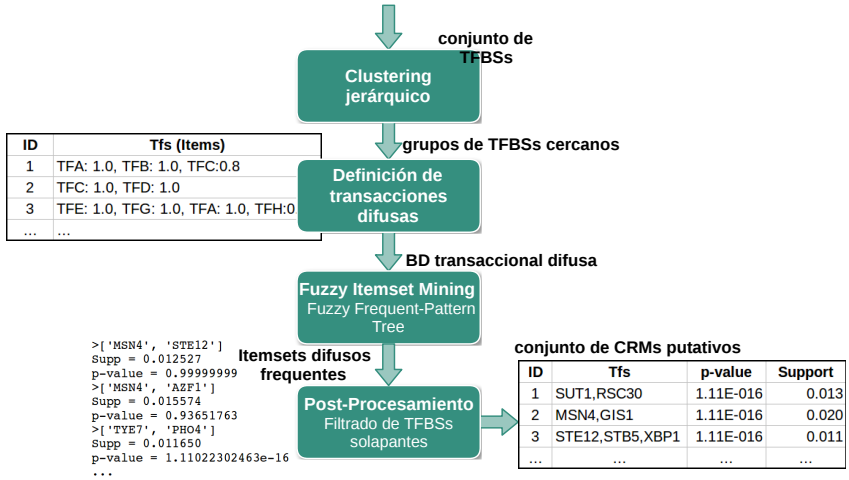


Figura 3.5: *Diagrama representativo de los pasos principales del pipeline de CisMiner. Dado un conjunto de TFBSs, el proceso comienza realizando un clustering jerárquico difuso para obtener conjuntos de TFBSs cercanos entre sí. El resultado de este paso es una base de datos transaccional difusa, que será procesada por un algoritmo de Frequent Itemset Mining difuso (Fuzzy Frequent-Pattern Tree) para obtener un conjunto de itemsets difusos frecuentes. Finalmente, CisMiner realiza una etapa de postprocesamiento para eliminar TFBSs solapantes que aparezcan en cada itemset. Como resultado se obtiene un conjunto de CRMs putativos junto con su p-value estimado y su soporte difuso.*

3.4.2. Clustering de TFBSs y construcción de la base de datos transaccional

Para extraer grupos significativos de TFs que conforman CRMs, el primer paso que ejecuta CisMiner es un clústering difuso para obtener TFBSs que se encuentran cercanos entre sí.

Como ya se ha mencionado en la introducción, los conjuntos difusos fueron propuestos por Zadeh [112] para modelar matemáticamente la imprecisión inherente a algunos conceptos. Brevemente, la teoría de conjuntos difusos permite asignarle a un elemento un grado de pertenecia a un conjunto difuso entre 0 y 1.

El algoritmo de Frequent Itemset Mining fue propuesto por Agrawal [284] para

extraer conjuntos frecuentes dentro de grandes bases de datos. Desde entonces, un vasto número de algoritmos se ha propuesto con la misma finalidad [285]. Dada una base de datos transaccional donde cada transacción es un conjunto de *ítems*, la meta de estas técnicas es encontrar un conjunto de expresiones de la forma $x_1, x_2, x_3, ...x_n$ donde cada x_i representa un ítem. Esta expresión es denominada *itemset*. La probabilidad de que un itemset ocurra en la base de datos se denomina el *soporte* del itemset. Si el soporte de un itemset es mayor que un umbral determinado por el usuario, entonces el itemset es considerado *frecuente*. Por tanto, los algoritmos de *Frequent Itemset Mining* intentan extraer de una base de datos transaccional itemsets cuyo soporte sea mayor que un umbral especificado como parámetro de entrada.

Los algoritmos de Frequent Itemset Mining se han aplicado a numerosas áreas dentro de la informática y han permitido con ello encontrar patrones ocultos en grandes cantidades de datos biológicos [286]. Sin embargo, la información biológica es con frecuencia imprecisa e incierta. Con la finalidad de poder reflejar esta incertidumbre inherente al tipo de datos utilizados, la tecnología difusa ha sido incorporada a la filosofía del Frequent Itemset Mining en el algoritmo Fuzzy FP-Tree utilizado en CisMiner.

Al conectar la tecnología difusa con el concepto de itemset frecuente podemos incorporar la incertidumbre y la imprecisión a nuestro modelo de conocimiento. En este sentido, un *itemset difuso* también es una expresión de la forma $x_1, x_2, x_3, ..., x_n$, con la adición en este caso para cada x_i de un valor en $[0, 1]$ que define el grado de pertenencia de x_i al itemset. El *soporte difuso* mide cuán frecuente es un itemset difuso.

CisMiner combina la teoría de conjuntos difusos con itemsets frecuentes con el fin de encontrar la forma de modelar la incertidumbre presente en los datos biológicos y, en particular, en la información de ubicación de TFBSs. Los clústers difusos de TFBSs cercanos son modelados como itemsets en una base de datos transaccional difusa. Para la construcción de dicha base de datos se ha aplicado un clustering jerárquico de promedio[10] sobre el conjunto de ubicaciones de los TFBSs.

[10] La implementación del clustering jerárquico se obtuvo de la librería python-hcluster.

Un umbral superior de aproximadamente $\approx 300bp$ se utilizó para la terminación de la función de agregación, asumiendo que los CRMs abarcan aproximadamente un tamaño del orden de los cientos de pares de bases [160].

Una vez obtenida la lista de clusters, una *transacción difusa* es definida para cada cluster. El grado de pertenencia de cada item a su itemset difuso correspondiente (esto es, de cada TFBS a su CRM putativo correspondiente) se define por una función trapezoidal, tal como se muestra en la figura 3.6, con los siguientes parámetros: se calcula un centroide C para cada cluster como el valor de la mediana de las posiciones de los TFBSs incluidos en el cluster. La región constante de la función trapezoidal se define en el intervalo $[C-150, C+150]$. Se define entonces una función lineal ascendente en el intervalo $[C-250, C-150]$ y una función lineal descendente en el intervalo $[C+150, c+250]$ para establecer el grado de pertenencia difuso de los elementos en los bordes deĺ cluster. Una vez definida la función de pertenencia, se crea la base de datos transaccional difusa de acuerdo al procedimiento que ilustra la figura 3.6. Para cada TFBS se calcula su función de pertenencia en base a su ubicación relativa al centroide y la función de pertenencia trapezoidal generada. Estos valores son introducidos en la base de datos transaccional para cada item de cada transacción.

3.4.2.1. Extracción de los conjuntos de items frecuentes

Una vez construída la base de datos transaccional, CisMiner procede a obtener conjuntos de coocurrencias significativas de TFs, es decir, CRMs putativos. El procedimiento de Fuzzy Itemset Mining permite a CisMiner eliminar clusters no siginificativos de TFs a escala de genoma completo. Con esta finalidad se implementa el algoritmo Top-Down Fuzzy Frequent Pattern-Growth, desarrollado en [283].

Inicialmente, este algoritmo recorre la base de datos transaccional para obtener una lista ordenada de todos los items frecuentes (i.e. items con un soporte mayor a un umbral especificado). Un item x_i representa a un TFBS y pertenece a una transacción de la forma $x_1, x_2, ..., x_n$ con un cierto grado de pertenencia. El objetivo del algoritmo de Frequent Itemset Mining es extraer itemsets significativos, es

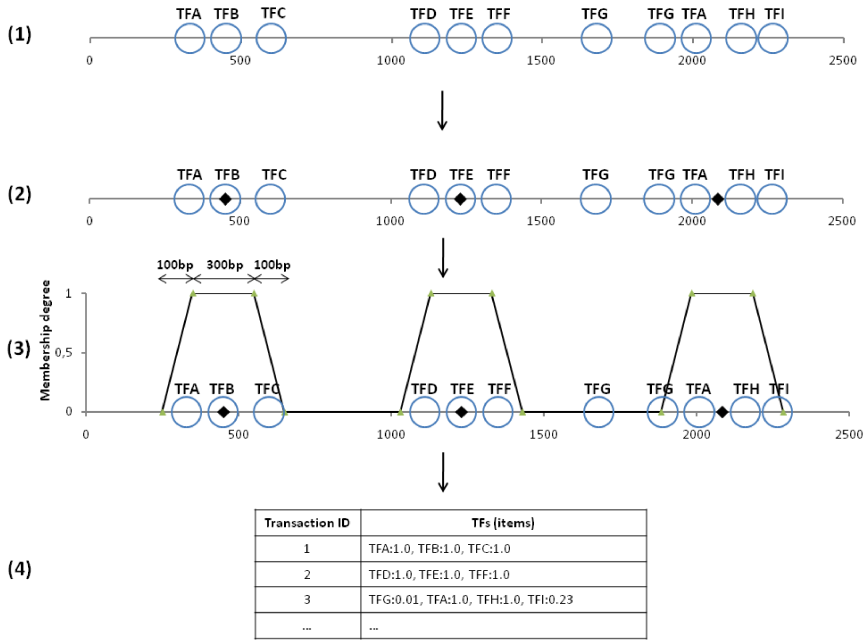


Figura 3.6: *(1) Cada círculo representa un TFBS. Cada TFBS está etiquetado con el nombre del TF que se une a dicha secuencia. (2) Se obtienen tres clusters y los centroides se calculan para cada cluster. (3) Se definen los conjuntos difusos correspondientes para la función de pertenencia de los TFBSs a cada cluster. (4) Se construyen las transacciones difusas a partir de los conjuntos difusos. El valor después del punto y coma indica el grado de pertenencia de cada TF a su transacción correspondiente.*

decir, colecciones de items que aparezcan *frecuentemente* entre las transacciones de la base de datos. Después de que los items con soporte superior al umbral hayan sido seleccionados, los items en esta lista junto con sus soportes son introducidos en la estructura de datos del Fuzzy FP-Tree. La eficiencia de este procedimiento recae en el uso de este árbol, ya que recopila toda la información de las transacciones necesaria para el algoritmo.

Los items de cada transacción que están presentes en la lista de items frecuentes se insertan como nodos en el Fuzzy FP-Tree de acuerdo a su posición en la lista de items frecuentes. Si dos transacciones comparten sus primeros items frecuentes

tambien compartirán el camino hasta el nodo raíz del árbol.

Para cada item I, todos sus nodos están conectados por una lista $side-links$. Un vector asociado a cada nodo almacena los valores de pertenencia de las transacciones que pertenecen al item correspondiente. Además, una tabla de cabeceras H es construida para que cada fila almacene la información asociada a un item I: (item, grados de pertenencia, side-links). Esta tabla permite localizar los nodos que corresponden a cada item en el Fuzzy FP-tree y calcular el soporte de cada itemset.

El soporte difuso para cada itemset se calcula de la forma descrita en [287]. Se calcula además un p-valor para complementar el valor de frecuencia obtenido por la medida de soporte. El procedimiento de [288] para el cálculo del p-valor se adapta al caso difuso. El modelo de cálculo de este p-valor considera una situación sin interés aquella en la que no hay asociaciones entre items, es decir, aquella en la que la ocurrencia de cada item es independiente de los demás en las transacciones. En este sentido, el p-valor representa la probabilidad del itemset de ser *sorprendente* bajo la hipótesis nula.

3.4.2.2. Post-procesado de los CRM putativos

El *clustering* jerárquico utilizado en el primer paso devuelve un conjunto de TFBSs cercanos. Sin embargo, en especial en el caso en el que una herramienta computacional de detección de TFBSs haya sido utilizada para la generación del conjunto de TFBSs de entrada, se deben considerar los efectos del solapamiento de TFBSs para generar los resultados finales. En el caso ilustrado en la figura 3.7, se produce un solapamiento de TFBSs en uno de los resultados. Otros enfoques eliminaban directamente ambos TFBSs en este caso [289]. Esta acción puede llevar a un conteo incorrecto de co-ocurrencia, ya que podría darse un caso en el que una combinación de TFBSs permitiera la unión de ambos. Por ello, CisMiner busca la forma óptima de adaptar una combinación de TFs dada (itemset) en una transacción difusa dada, maximizando el grado de pertenencia del itemset a la transacción. Este encaje óptimo es considerado un CRM putativo.

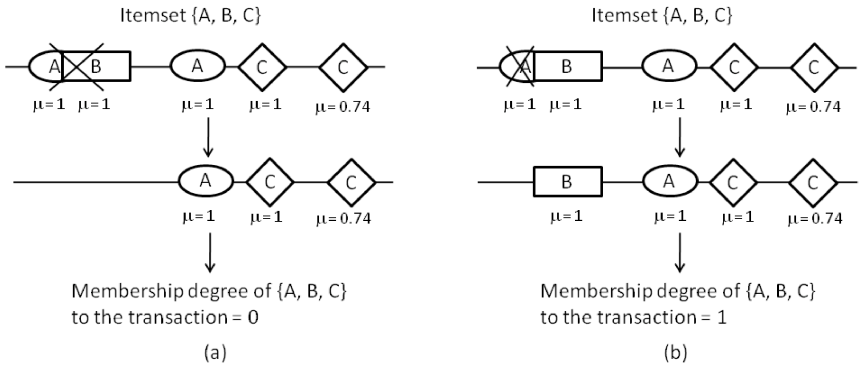


Figura 3.7: *El valor μ indica el grado de pertenencia de cada lugar de unión a su transacción correspondiente. (a) Las parejas de lugares de unión solapantes son eliminadas directamente. (b) Se busca la forma óptima de ajustar el itemset {A, B, C}*

3.4.2.3. Validación de la metodología

Con el fin de comprobar el funcionamiento de la metodología, se aplicó CisMiner a dos conjuntos de datos sobre el genoma de *S. cerevisiae*. Primero, se utilizaron como TFBSs de entrada aquellos presentes en la anotación publicada por Harbison et al [290]. A este conjunto de datos lo llamaremos en el siguiente epígrafe el conjunto de datos de **Harbison**. Posteriormente, para evaluar la capacidad de la metodología propuesta de eliminar resultados espúreos y falsos positivos provocados por herramientas de predicción computacional de TFBSs, se realizó un estudio en el genoma de *S. cerevisiae* para un conjunto de TFBSs predichos computacionalmente. Ambos experimentos de validación son descritos a continuación en detalle.

TFBSs validados de forma experimental. 3328 TFBSs fueron encontrados para 102 TFs en los datos de TFBSs publicados por Harbison et al [290]. Con estos datos se generaron 570 transacciones con una media de 2.79 TFs por transacción y un máximo de 10 TFs en una sola transacción. De estos, CisMiner obtuvo 36 itemsets tras definir los umbrales para soporte difuso y p-valor en 0.01. La frecuencia de aparición de los 96 TFs que fueron incluídas al menos en una

transacción indicaron que hay varios TFs no muy frecuentes con alta probabilidad de generar resultados espúreos.

Los 36 itemsets encontrados por CisMiner fueron contrastados con STRING [291]. Dado un conjunto de genes, STRING busca asociaciones entre estos genes en distintos niveles: distancia en el genoma, coocurrencia de los genes de consulta a través de especies, coexpresión, interacciones proteína-proteína, bases de datos manualmente curadas y text mining. STRING dio evidencia de relaciones entre todos los TFs presentes en los itemsets seleccionados por CisMiner para 32 de los 36 resultados (88,8 %). Además, para 30 de los 36 itemsets, los grafos representando las asociaciones entre los TFs es un grafo conexo (i.e. hay un camino posible entre todos los pares de nodos del grafo). Las 2 restantes de las 32 encontradas en STRING contenían relaciones indirectas con un único TF ajeno al itemset.

La tabla 3.7 muestra los 20 resultados más significativos obtenidos junto con su p-valor. STRING devolvió un grafo conexo para todos los resultados de la tabla excepto los números 13 y 18. Las relaciones entre los TFs presentes en los CRMs predichos son fuertes a diferentes niveles. De los 18 resultados que presentan grafos conexos, en 15 encontramos conexiones en STRING correspondientes a tres clases de evidencia separadas, entre validación experimental, aparición en bases de datos manualmente verificadas, mención conjunta en abstracts de la literatura y predicción de interacción (o interacción experimental validada) de genes homólogos en otras especies. En particular, de los 18 grafos conexos en STRING para los 20 resultados más significativos, para 15 había evidencia experimental de interacción según STRING, para todas las conexiones de dichos grafos, habiendo evidencia según bases de datos y algoritmos de text mining para 18 de los itemsets propuestos.

Algunos de los CRMs propuestos por CisMiner representan procesos biológicos bien caracterizados, como es el caso de la interacción de SWI6 con MBP1 y SWI4 (itemset número 3) para formar los complejos proteínicos MBF y SBF, que cooperan y juegan un papel principal en la progresión de la fase G1 a la S en el ciclo celular [292]. De hecho, es importante mencionar que las combinaciones SWI6, SWI4 y SWI6, MBP1 también están presentes en las posiciones 2 y 7, respectivamente.

Tabla 3.7: *Resultados CisMiner para anotaciones de TFBSs de Harbison et al. Los 20 primeros resultados con el p-valor más bajo. La columna Evidencia muestra (P) si PubMed devolvió resultados en la consulta de los TFs en el CRM putativo, (S) si STRING devolvió un grafo conexo para la consulta, (SP) si ambas condiciones se cumplen y (-) si ninguna.*

ID	CRM putativo	P-valor	Soporte	Evidencia
1	STE12, DIG1	$1{,}11 \times 10^{-16}$	0.059	SP
2	SWI6, SWI4	$1{,}11 \times 10^{-16}$	0.056	SP
3	SWI6, MBP1, SWI4	$1{,}11 \times 10^{-16}$	0.025	SP
4	SKN7, SOK2, PHD1	$3{,}00 \times 10^{-15}$	0.014	S
5	STE12, DIG1, TEC1	$2{,}44 \times 10^{-13}$	0.018	SP
6	SOK2, PHD1	$1{,}20 \times 10^{-12}$	0.027	SP
7	SWI6, MBP1	$3{,}02 \times 10^{-12}$	0.043	SP
8	MBP1, SWI4	$6{,}81 \times 10^{-11}$	0.038	SP
9	RAP1, FHL1	$4{,}66 \times 10^{-09}$	0.016	SP
10	DIG1, SWI4, TEC1	$2{,}02 \times 10^{-08}$	0.011	SP
11	DIG1, TEC1	$4{,}74 \times 10^{-08}$	0.029	SP
12	AFT2, RCS1	$4{,}83 \times 10^{-08}$	0.012	SP
13	PHD1, SUT1	$5{,}92 \times 10^{-08}$	0.011	-
14	STE12, TEC1	$7{,}19 \times 10^{-08}$	0.032	P
15	STE12, SWI6, SWI4	$9{,}72 \times 10^{-08}$	0.014	SP
16	SWI6, DIG1, SWI4	$2{,}13 \times 10^{-07}$	0.012	SP
17	FKH2, NDD1	$3{,}23 \times 10^{-07}$	0.016	SP
18	SOK2, SUT1	$8{,}78 \times 10^{-07}$	0.012	-
19	SKN7, SOK2	$1{,}48 \times 10^{-06}$	0.022	S
20	SWI6, STB1	$2{,}14 \times 10^{-06}$	0.012	SP

Otra interesante relación la representan los itemsets 1, 5 y 14 sobre los factores de transcripción STE12, DIG1 y TEC1. STRING devuelve un grafo completo para estos tres TFs con fuerte evidencia empírica y un artículo reciente de Van der Felden et al. [293] también relaciona estos tres TFs. La coocurrencia de RAP1 y FHL1 (itemset 9) concuerda también con resultados existentes, ya que se ha demostrado que ambos se unen a secuencias promotoras de muchos genes de proteínas ribosomales [294].

Para los itemsets 13 y 18 no se encontró ninguna relación en STRING.

Sin embargo, es de interés mencionar que al consultar por el TF SUT1 en STRING, ninguno de los elementos conectados a él estaba presente en el dataset de Harbison. Por tanto, no era posible encontrar itemsets que pudieran ser confirmados por STRING desde el conjunto de datos inicial.

Además de la validación con STRING, ejecutamos un test utilizando PubMed para obtener validación complementaria buscando en la literatura existente evidencia apoyando los resultados obtenidos. Para los 36 CRMs resultantes, 29 dan resultados al ser buscados en PubMed, representando un 80,55 % de los resultados. Los itemsets 4, 13, 18 y 19 no dieron resultados en PubMed. De estos 4 resultados, es interesante remarcar que 4 y 19 sí estaban vinculados en STRING.

Por último, nótese que la falta de evidencia para el resto de los itemsets implicados (4) no necesariamente implica que no existan dichas relaciones, sino que en el momento de la experimentación no existe evidencia que soporte estas asociaciones. Estudios adicionales serían necesarios para poder descartar totalmente estas asociaciones putativas por validar.

TFBSs detectados computacionalmente. En el experimento descrito con anterioridad se utilizaron como datos de entrada un conjunto de datos de TFBSs ya validados. CisMiner, no obstante, permite su uso tanto con anotaciones de este tipo como con predicciones hechas con cualquier herramienta de predicción de TFBSs. En este sentido, hemos validado CisMiner con un conjunto de TFBSs putativos predichos computacionalmente en el genoma de *Saccharomyces cerevisiae*. Con este fin, se obtuvo el genoma de *S. cerevisiae* de la *Saccharomyces* genome Database (SGD) [295][11] y datos de motivos de dicho organismo de la base de datos JASPAR [296]. Más específicamente, el genoma completo de la levadura se obtuvo de SGD y 177 PWMs fueron obtenidas de JASPAR.

Primero fueron inferidas las posiciones de los TFBSs potenciales. Se utilizaron para este fin las herramientas Patser [297] y Consite [142], ya que su eficacia ha sido probada en repetidas ocasiones [289, 298, 299]. Existen otras metodologías para esta finalidad que utilizan interdependencia entre posiciones [145], pero

[11] http://www.yeastgenome.org/

requieren para ser aplicadas que se proporcionen las secuencias que dieron lugar a las matrices de pesos, y esta información no siempre es proporcionada. En este caso, JASPAR no proporciona dichas secuencias para ninguno de los 177 motivos utilizados.

Comprobamos la capacidad de Patser y Consite para recuperar los TFBSs reales utilizando el dataset de Harbison como benchmark. Primero, ejecutamos Patser sobre el genoma completo de la levadura. Dada una PWM y una secuencia, Patser devuelve una lista de valores en el rango $[0, 15]$ indicando la afinidad del motivo representado por la PWM con cada posición de la secuencia. Por ello, necesitamos escoger un umbral de corte para el score obtenido. Adicionalmente, algunos TFBSs pueden ser más fáciles de detectar que otros debido a la composición de la secuencia, la longitud y a la composición del genoma objetivo entre otros factores. Es por ello que un umbral de corte es necesario para cada motivo [300], ya que las distribuciones de scores pueden variar ampliamente entre motivos. De acuerdo a la documentación de Patser[12] y a nuestra experiencia empírica, no es viable utilizar umbrales inferiores a 7.5. Observamos que por debajo de 7.5 obteníamos una gran cantidad de TFBSs putativos, disminuyendo la significancia de los resultados. Por ello se seleccionó para cada motivo un umbral específico superior a 7.5. La selección de dicho umbral se hizo optimizando el número de TFBSs obtenidos del dataset real de Harbison.

Un procedimiento similar fue llevado a cabo para poner a prueba el rendimiento de Consite. En este caso, para cada PWM, Consite devuelve un valor en $[0, 100]$. Los mejores resultados para Consite se obtuvieron con un umbral por debajo de 70. La figura 4 muestra el número de TFBSs verdaderos recuperados por cada una de las herramientas contra el número total de TFBSs obtenidos. Como se puede ver, Patser tiene un rendimiento superior a Consite en este caso, ya que el número de TFBSs obtenidos por Patser es inferior para el mismo valor de verdaderos positivos. Patser fue por esta razón seleccionado para este test. De los 177 motivos en JASPAR, 66 corresponden con TFBSs en Harbison. Para el resto, los umbrales de corte se establecieron en la mediana del resto de los umbrales.

[12] http://rsat.ulb.ac.be/rsat/help.patser.html

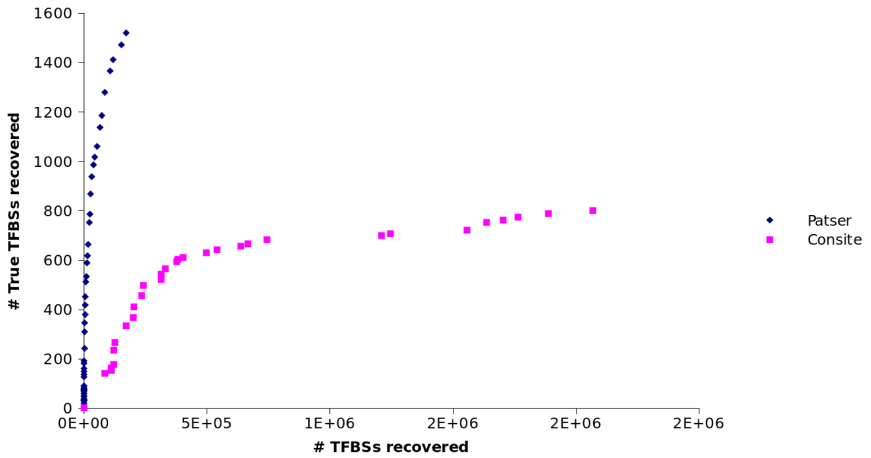


Figura 3.8: *Número de TFBSs del dataset de Harbison [290] contra el número total de TFBSs detectado para Patser y para Consite.*

Con esta configuración, 77921 TFBSs fueron detectados, entre los que se incluyen 1412 de los descritos por Harbison. Estos 1412 sitios conforman el $\approx 50\,\%$ del número total de TFBSs descritos por Harbison. Es digno de mención el hecho de que para un conjunto de motivos (por ejemplo, ARR1, ASH1, BAS1) no se pudo detectar ningún TFBSs, incluso con el umbral de corte de Patser a 0. Igualmente, hubo un conjunto de motivos que también necesitó unos umbrales de corte muy bajos para poder capturar sus TFBSs correspondientes (1.65, 2.25, 0.3 para ABF1, ADR1, AFT2, respectivamente). Toda esta casuística puede reflejar la complejidad biológica del problema. También puede deberse a ciertas diferencias entre los motivos recolectados en JASPAR y aquellos detectados por Harbison. Los tests descritos a continuación son una forma de mostrar la capacidad de CisMiner de superar este tipo de problemas y encontrar CRMs putativos de interés a pesar del ruido presente en los datos biológicos. Sin embargo, cabe recalcar que la búsqueda computacional de TFBSs individuales se encuentra fuera de la finalidad de CisMiner.

Una vez los TFBSs putativos han sido detectados, la base de datos transaccional

difusa es construída. 8716 transacciones se obtuvieron con una media de 8.67 TFs diferentes por transacción y un máximo de 45 TFs en la misma transacción. 255 itemsets fueron obtenidos. Las combinaciones de TFs obtenidas muestran altos p-valores, lo cual indica que muchas de estas combinaciones podrían tener significado biológico. Adicionalmente, los itemsets obtenidos tienen entre 2 y 4 TFs, lo cual encaja con estimaciones hechas en trabajos anteriores [168, 301]. La tabla 3.8 es una muestra de los resultados obtenidos.

En primer lugar, observamos si los resultados obtenidos con TFBSs obtenidos computacionalmente concordaban con aquellos obtenidos del dataset de Harbison. Comparación directa de ambos conjuntos de resultados muestra que sólo el itemset STE12, TEC1 es compartido por ambos. Esto se debe a que el umbral de corte del soporte difuso filtra los resultados del segundo enfoque, ya que si reducimos dicho umbral podemos obtener hasta 13 de los resultados del primer conjunto. Por otro lado, los TFs DIG1, NDD1, SWI6, RCS1 y STB1 no se encuentran en JASPAR, por lo que no es posible encontrarlos en el segundo conjunto de resultados, justificando así la ausencia de 14 de los itemsets que aparecían en el primer conjunto de resultados. Finalmente, el resto de itemsets dejaron de ser significativos en el segundo conjunto de datos.

De nuevo utilizamos STRING para validar los resultados nuevos. En este caso, STRING proporcionó evidencia de relaciones entre los factores de transcripción correspondientes a 23 de los itemsets recuperados. Adicionalmente, para los primeros 100 resultados, se obtuvieron resultados STRING involucrando relaciones directas entre los factores de transcripción de 11 de los CRMs putativos, y relaciones indirectas (involucrando únicamente 1 TF adicional) para 24. Esto indica que para los 100 primeros resultados, alrededor del 35 % representan una relación directa o indirecta entre los TFs propuestos.

La tabla 3.8 muestra algunos de los resultados obtenidos en esta prueba. STE12 y TEC1 (itemset 1) parecen estar fuertemente relacionados, ya que se sabe que ambas proteínas cooperan y regulan varios procesos celulares [302, 303, 304]. La proteína ADR1 aparece relacionada con RAP1, RGT1 y SIP4 (itemsets 2 a 4). Referencias bibliográficas confirman estas asociaciones. ADR1 y RAP1, entre

Tabla 3.8: *Algunos resultados de los devueltos por CisMiner para TFBSs computacionalmente predichos con Patser.*

ID	CRM putativo	P-valor	Soporte
01	STE12, TEC1	$8{,}40 \times 10^{-03}$	0.013
02	ADR1, RAP1	$2{,}39 \times 10^{-08}$	0.014
03	ADR1, RGT1	$9{,}18 \times 10^{-05}$	0.013
04	ADR1, SIP4	$1{,}22 \times 10^{-04}$	0.011
05	AFT2, RAP1	$1{,}17 \times 10^{-13}$	0.012
06	MIG3, RGT1	$3{,}13 \times 10^{-6}$	0.012
07	MSN4, RPN4	$2{,}27 \times 10^{-3}$	0.024
08	MSN4, SKN7	$9{,}24 \times 10^{-6}$	0.013
09	MSN4, GIS1	$1{,}11 \times 10^{-16}$	0.020
10	MIG1, MIG2	$1{,}11 \times 10^{-16}$	0.016
11	STE12, GCR2, STB5, XBP1	$1{,}11 \times 10^{-16}$	0.010
12	ADR1, MIG1	$1{,}11 \times 10^{-16}$	0.020
13	SUT1, MIG1	$1{,}11 \times 10^{-16}$	0.018

otros reguladores de la transcripción, pueden participar en funcionamiento de barrera, bloqueando la propagación del silenciado de la transcripción en la levadura [305]. Del mismo modo, la relación entre ADR1 y RGT1 (YKL038W) también se encuentra en la literatura. Estos dos factores de transcripción están involucrados en la respuesta transcripcional a ciertas perturbaciones transitorias en la fuente de carbono [306]. Finalmente, ADR1 y SIP4 participan en el control transcripcional del metabolismo no enfermante en *Saccharomyces cerevisiae* [307]. El siguiente itemset en la lista (itemset 5) contiene a los factores AFT1 y RAP1, los cuales aparecen relacionados cuando se consulta en STRING[291]. La siguiente combinación involucra a MIG3 con RGT1, previamente mencionado. Hazbun et al. [308] demostraron experimentalmente que los dos TFs se unen a la región promotora de SUC2, que produce enzima invertasa. Después, tres itemsets relacionan MSN4 con RPN4, SKN7 y GIS1 (itemsets 7 a 9), todos ellos describiendo asociaciones conocidas. MSN4, RPN4 y SKN7 participan en la respuesta transcripcional de *Saccharomyces cerevisiae* al estrés impuesto por ciertos

fungicidas y herbicidas [309, 310]. Respecto a la relación entre MSN4 y GIS1, STRING devolvió asociaciones a diferentes niveles: experimental, bases de datos y text mining. Otros autores también describieron con anterioridad que el reguloma de Rim15 está mediado por estos dos factores de transcripción [311]. Por último, MIG1 y MIG2 también parecen estar relacionados a varios niveles. La represión de la glucosa del gen SUC2 depende de MIG1 y MIG2 [312]. Los últimos itemsets en la tabla muestran algunas combinaciones para las que STRING devolvió asociaciones indirectas, y algunos para los que no devolvió asociación.

De forma análoga a la sección anterior, se realizó una búsqueda en PubMed con el fin de obtener más información sobre el set de resultados completo. Los resultados de PubMed fueron manualmente procesados y se encontró relación entre al menos menos 23 combinaciones más.

En nuestro objetivo de proporcional mayor evidencia de la significancia de los resultados, la metodología completa fue ejecutada sobre 100 genomas randomizados. El número medio de TFBSs obtenidos por Patser para este datase fue de 73307.9, número cercano a los 77921 obtenidos en el genoma real. Es interesante remarcar que a pesar de ello sólo unas 11.5 combinaciones significativas de TFs se obtuvieron como resultado de media, cuando en el caso real se obtuvieron 255 combinaciones significativas. Este hecho podría ser evidencia real de la capacidad de la metodología propuesta para eliminar resultados espúreos, y sugiere una tasa FDR[13] menor del 5 %.

Para finalizar esta parte, es importante mencionar que el proceso completo fue vuelto a ejecutar de nuevo después de elevar el umbral de corte de Patser a 8. En este caso se obtuvieron 56 itemsets. Es de interés comentar que en este caso el resultado era un subconjunto de los 255 resultados obtenidos para el umbral más bajo de Patser, y los p-valores similares también a los obtenidos con el umbral de 7.5. Este hecho sugiere también coherencia y robustez.

[13] *False discovery rate.*

3.4.3. Caso de estudio: *Drosophila Melanogaster*

Con la finalidad de comprobar tanto si la metodología propuesta es viable en genomas de mayor volumen y complejidad como si es posible obtener resultados consistentes en dicha situación, se ha aplicado CisMiner al genoma de *Drosophila melanogaster*.

Aunque la interpretación biológica exhaustiva de todos los resultados obtenidos estaba fuera del ámbito de este trabajo, el interés de los resultados es discutido y analizado en esta sección. El genoma completo de *Drosophila melanogaster* fue descargado de FlyBase[14] [313] (release 5.56, marzo 2014). Adicionalmente, 131 PWMs correspondientes a *Drosophila melanogaster* fueron descargadas de JASPAR. De acuerdo a los umbrales de Patser calculados, las secuencias con score superior a 8.1 fueron consideradas TFBSs potenciales. Así, 152338 transacciones fueron obtenidas con una media de 4.34 TFs diferentes por transacción y un máximo de 15 TFs en una misma transacción.

Pese a partir de un conjunto mucho mayor de TFBSs de entrada, sólo 39 CRMs candidatos fueron obtenidos. La tabla 3.9 muestra los 10 mejores resultados obtenidos. En este caso, STRING solo obtuvo una relación directa: Mad, brk. Otros 8 itemsets devolvieron relaciones indirectas cuando se consultó STRING, representando aproximadamente un $\approx 20\,\%$ del total. Por otra parte, 12 de los resultados obtuvieron resultados cuando se hizo una búsqueda en la literatura a través de PubMed. Sin embargo, existe una gran dificultad adicional a la hora de buscar evidencia en la literatura a través de los identificadores, ya que las etiquetas en los genes de *D. melanogaster* tienen identificadores muy poco específicos: *opa, btd, h, D*. También se encontró delación entre los CRMs putativos sugeridos para al menos dos itemsets más al inspeccionar manualmente los resultados obtenidos de PubMed. Este último experimento muestra que es posible aplicar la metodología a genomas más complejos. Además, estos 39 módulos presentan valores de calidad elevados y podrían representar interacciones biológicas reales. Sin embargo, se requiere continuar con la investigación para validar e interpretar todas las combinaciones obtenidas.

[14] http://flybase.org

Tabla 3.9: *Tercer dataset. Resultados obtenidos en el genoma de D. melanogaster para TFBSs putativos predichos por Patser.*

ID	CRM putativo	P-valor	Soporte
01	btd, hkb	$1{,}11 \times 10^{-16}$	0.045
02	btd, Mad	$1{,}11 \times 10^{-16}$	0.041
03	btd, opa	$1{,}11 \times 10^{-16}$	0.033
04	btd, h	$1{,}11 \times 10^{-16}$	0.003
05	Mad, brk	$1{,}11 \times 10^{-16}$	0.029
06	btd, CTCF	$1{,}11 \times 10^{-16}$	0.023
07	Mad, opa	$1{,}11 \times 10^{-16}$	0.023
08	Mad, hkb	$1{,}11 \times 10^{-16}$	0.022
09	brk, opa	$1{,}11 \times 10^{-16}$	0.021
10	Mad, h	$1{,}11 \times 10^{-16}$	0.019

3.4.4. Metodologías difusas frente a sus análogas *crisp*

Para evaluar la contribución de la inclusión de la tecnología difusa al pipeline, implementamos una versión *crisp* de la metodología y comparamos los resultados utililzando los mismos umbrales de corte. El procedimiento para crear la base de datos transaccional *crisp* fue una versión *crisp* de la metodología difusa descrita. Primero, un clustering *crisp* fue ejecutado para identificar las transacciones, donde las fronteras *crisp* fueron definidas en $[C-300, C+300]$ alrededor del centroide. En la base de datos transaccional *crisp*, todos los elementos tienen 1.0 de grado de pertenencia al itemset si se encuentran en el intervalo $[C-300, C+300]$. Después, una versión *crisp* del frequent itemset mining fue aplicada sobre las transacciones.

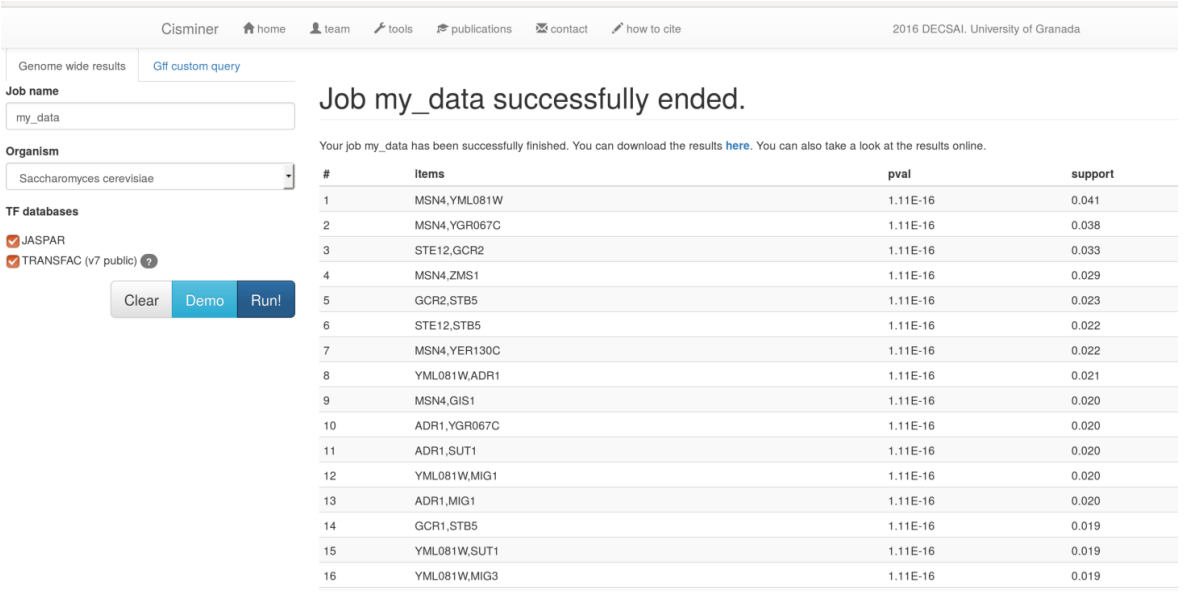
Tal y como se esperaba, se encontraron diferencias significativas entre los resultados difusos y los *crisp*. Primero, el algoritmo *crisp* obtuvo 4217 combinaciones mientras el algoritmo difuso obtuvo 1388, siendo estos un subconjunto de los resultados *crisp*. Para determinar si los valores obtenidos entre ambas metodologías eran significativamente diferentes, se aplicaron dos tests ANOVA (ver tabla 3.10). Los conjuntos difusos han demostrado ser una tecnología superior para modelar particiones con bordes imprecisos. Los resultados obtenidos

Tabla 3.10: *Comparativa de resultados difusos contra resultados crisp análogos.*

	S. cerevisiae	*D. melanogaster*
Soporte medio difuso	0.0084	0.0088
Soporte medio crisp	0.0108	0.0094
p-valor medio difuso	0.0144	0.0133
p-valor medio crisp	0.0070	0.0104
ANOVA signif. soporte	$< 10^{-38}$	$< 10^{-8}$
ANOVA signif. p-valor	$< 10^{-19}$	$< 10^{-8}$

sugieren más p-valores significativos en el conjunto de resultados difuso, pudiendo significar que la poda de resultados espúreos en el algoritmo difuso es efectiva.

Adicionalmente, se siguieron los mismos pasos para comparar los resultados en el genoma de *D. melanogaster*. En este caso se obtuvieron 4388 combinaciones de TFs en el caso *crisp*, mientras que la metodología difusa obtuvo 4272 resultados. La tabla 3.10 muestra que en este caso los comentarios para el caso de *S. cerevisiae* también son ciertos para el nuevo caso, mostrando las ventajas del uso de una metodología difusa para modelar los clústers de TFBSs, mejorando la representación de un CRM al permitir incertidumbre en las fronteras y consiguiendo una efectividad mayor en términos de p-valores.



#	Items	pval	support
1	MSN4,YML081W	1.11E-16	0.041
2	MSN4,YGR067C	1.11E-16	0.038
3	STE12,GCR2	1.11E-16	0.033
4	MSN4,ZMS1	1.11E-16	0.029
5	GCR2,STB5	1.11E-16	0.023
6	STE12,STB5	1.11E-16	0.022
7	MSN4,YER130C	1.11E-16	0.022
8	YML081W,ADR1	1.11E-16	0.021
9	MSN4,GIS1	1.11E-16	0.020
10	ADR1,YGR067C	1.11E-16	0.020
11	ADR1,SUT1	1.11E-16	0.020
12	YML081W,MIG1	1.11E-16	0.020
13	ADR1,MIG1	1.11E-16	0.020
14	GCR1,STB5	1.11E-16	0.019
15	YML081W,SUT1	1.11E-16	0.019
16	YML081W,MIG3	1.11E-16	0.019

Figura 3.9: *Captura de un resultado de CisMiner ejecutado desde la web. Cada itemset frecuente candidato es acompañado de un p-valor junto con su soporte. Además, es posible descargar los resultados en formato .xls para su posterior análisis.*

3.5. Conclusiones

En este capítulo se ha abordado el problema de la búsqueda de elementos reguladores en el genoma en dos niveles de complejidad. Primero, se ha propuesto IntuitSNP, una herramienta que busca SNPs con potencial influencia en la regulación génica por su ubicación en TFBSs. Posteriormente, se ha propuesto CisMiner, una metodología que permite buscar módulos reguladores aplicando una versión difusa de un algoritmo de frequent itemset mining.

En primer lugar, IntuitSNP ha tratado de cubrir las limitaciones existentes en herramientas para la investigación sobre SNPs en regiones reguladoras, no codificantes, del genoma. Este tipo de alteraciones en el genoma están en muchos casos relacionadas con enfermedades. Sin embargo, en los últimos años ha proliferado el estudio de mutaciones en regiones codificantes del genoma, es decir, aquellas que tienen influencia directa en la codificación de la secuencia de aminoácidos que posteriormente dará lugar a una proteína. No obstante, estas regiones no son las únicas que influyen en el resultado final, ya que las regiones reguladoras son las encargadas de controlar el proceso de producción. Alteraciones en estas regiones pueden llegar a ser críticas para el estado de salud del individuo, y su estudio de vital importancia para futuros progresos en tratamiento y diagnóstico de pacientes, en pos de lo que se conoce hoy en día como Medicina Personalizada.

Los resultados presentados son esperanzadores en lo que a precisión se refiere, ya que el porcentaje de acierto es considerable y la herramienta ha probado ser más efectiva que otros enfoques. Además, se ha dado un paso más allá en eficiencia, con la intención de proporcionar los resultados en un tiempo asequible a la hora de analizar grandes cantidades de SNPs. Para ello, se han utilizado técnicas de paralelización mediante GPU, las cuales, además de ser muy efectivas, resultan económicamente asequibles.

En segundo lugar, CisMiner, la metodología computacional para la detección módulos de regulación en *cis*, ofrece importantes ventajas respecto a otras metodologías existentes y resultados prometedores. La utilización de conjuntos difusos para encapsular CRMs crea un modelo más realista de dichos módulos.

Suavizar los bordes de los clusters parece un modelo más adecuado a la realidad que definir particiones *crisp*. De hecho, la tecnología difusa ha demostrado su superioridad a la hora de mejorar la interpretabilidad de este tipo de particiones [314].

Por otro lado, la metodología propuesta no requiere de conocimiento previo a la hora de realizar consultas. CisMiner se ha aplicado al conjunto de secuencias completo de *S. cerevisiae* y *D. melanogaster* sin restricciones en las secuencias a utilizar. Además, tampoco se han establecido restricciones ni sobre el número de TFBSs a incluir en cada CRM ni sobre la disposición de dichos TFBSs en cada módulo. Los resultados proporcionados por la herramienta son fáciles de interpretar, ya que consisten en los itemsets considerados significativos junto con un p-valor y un soporte difuso. Por último, CisMiner está libremente disponible como herramienta web para consulta de la comunidad científica en `http://genome.ugr.es/cisminer`. La imagen 3.9 muestra una captura de su funcionamiento.

Se han proporcionado asimismo resultados experimentales que confirman la consistencia de los resultados proporcionados por la herramienta con el conocimiento existente de la biología de los organismos explorados. En el caso de utilizar predicciones computacionales de TFBSs, se ha observado también calidad en los resultados, a pesar de ser estos fuertemente dependientes de la efectividad de dicha herramienta.

Aunque existen muchos enfoques para el estudio en profundidad de mecanismos conocidos, es nuestro entendimiento que existen menos herramientas que permitan la exploración sin conocimiento a priori de un organismo. En este sentido, la aproximación de CisMiner a genoma completo puede ser de utilidad. Queda pues como trabajo futuro la aplicación de esta herramienta a nuevos genomas de mayor tamaño y más complejos. Adicionalmente, la integración de nuevas fuentes de datos aparte de información puramente secuencial (estructura de la cromatina, información de metilación) puede ayudar a refinar los resultados.

Capítulo

4

ARNs largos no codificantes

Los recientes avances en las tecnologías de secuenciación han demostrado que menos de un 2 % del genoma humano, $\sim$ 30 millones de bases [10], codifica genes. El resto del ADN presente en un organismo ha sido durante mucho tiempo clasificado como *ADN basura*, a pesar de que el conocimiento sobre su funcionalidad ha ido aumentando con el paso de los años. De hecho, pese a que el tamaño del genoma humano es mucho mayor que el de otras especies, su número de genes, alrededor de 20.000, no difiere tanto [16], lo que sugiere que el ADN *intergénico* alberga muchos de los mecanismos responsables de la complejidad de nuestra especie [17].

Con el tiempo se van descubriendo nuevas funcionalidades, entre las cuales se encuentran *loci* capaces de producir transcritos que no se convierten posteriormente en proteínas, así llamados por ello ARNs no codificantes. Recientes estudios han remarcado la omnipresencia del fenómeno de la transcripción a lo largo del genoma. Dicho de otro modo, pese al reducido porcentaje del genoma humano que acaba generando proteínas, más de un 85 % del mismo se transcribe [11].

Mientras otros tipos de ARN no codificantes (e.g. micro ARNs, transfer ARNs) han sido extensamente estudiados, desvelando gran cantidad de subtipos y

funcionalidades [18], los ARNs largos no codificantes (lncRNAs), están siendo objeto de creciente interés en los últimos años. Sin embargo, pese a existir más genes que transcriben lncRNAs (un estudio reciente de Iyer *et al.* [20] sitúa en más de 58.000 la cantidad de *loci* capaces de transcribir lncRNAs) que genes codificantes, los lncRNAs son menos abundantes en términos de volumen de transcripción que otros tipos de ARN [22]. Por ello, este tipo de ARNs no codificantes ha sido menos estudiado en el pasado. La reciente aparición de nuevas tecnologías de secuenciación capaces de capturar la transcripción de ARNs poco abundantes ha permitido un conocimiento mayor sobre estas entidades biológicas.

No obstante, los lncRNAs son a día de hoy poco conocidos, aunque ya se ha demostrado que juegan numerosos papeles en la maquinaria de regulación celular [315] y, por tanto, también están relacionados con enfermedades [26]. Una mejora del conocimiento de los mecanismos tras estas nuevas entidades es por tanto de interés para el campo de la Medicina Personalizada, y además una aportación relevante a un campo en el que todavía existen pocas iniciativas sistemáticas de anotación funcional de lncRNAs.

El problema que se pretende abordar en este capítulo es el análisis sistemático de lncRNAs mediante nuevas metodologías de análisis de secuencia que tengan en cuenta las características específicas del ARN. De la aplicación de estas herramientas surgen anotaciones de elementos relevantes encontrados en el genoma, las cuales podrán ser utilizadas como base para la construcción de análisis posteriores.

En la sección 4.1 se hace una revisión del estado actual de conocimiento de los lncRNAs y las herramientas de análisis existentes para la investigación en este ámbito. A continuación, en las siguientes secciones se detallan dos enfoques diferentes para el análisis de secuencia de ARN. La sección 4.2 detalla el primero, **RNAIntuit** [316]. Esta propuesta se centra en la búsqueda de motivos de unión lncRNA-proteína, haciendo uso de las bases de datos de reciente aparición y explotando la información obtenida de la estructura secundaria del ARN. La metodología propuesta se aplica en la sección 4.3 a Fendrr, un lncRNA relacionado con la oncogénesis [317]. A continuación, la sección 4.4 describe el segundo

enfoque, **MoSI** (Motif Strand Imbalance), un pipeline de análisis de secuencias de ARN que pretende encontrar en ellas motivos *de novo* a partir de una característica propia del ARN: el desequilibrio de hebras. Los motivos obtenidos a raíz de este segundo análisis han dado lugar a un conjunto de anotaciones que esperamos será de utilidad para investigaciones futuras. Finalmente, las conclusiones de ambos trabajos se detallan en la sección 4.5.

4.1. Análisis computacional y experimental de lncRNAs: Estado del arte

Tradicionalmente, de acuerdo con el dogma central de la biología molecular, el análisis de ARN se ha circunscrito a ARN mensajero, es decir, al transcrito a partir de las secuencias que comúnmente llamamos genes, el cual es posteriormente procesado por ribosomas para fabricar proteínas. Sin embargo, en los últimos años se ha evidenciado que el ADN codificante dista de ser el único que es transcrito. Numerosos estudios han hecho patente la existencia de multitud de ARNs no codificantes [318, 319, 320]. Entre estos, han adquirido gran protagonismo los ARNs largos no codificantes, o lncRNAs.

Los lncRNAs son secuencias transcritas de ARN de longitud superior a 200*nt* que no son traducidos en proteínas. Aunque esta distinción de longitud pueda resultar arbitraria, lo cierto es que al momento del descubrimiento de estos transcritos, tanto su longitud como su bajo potencial codificador eran sus características más distintivas, en contraste con su contrapartida, los ya muy estudiados genes codificantes [315].

Hoy sabemos, además, que el volumen de transcripción de los lncRNAs puede ser aproximadamente unas 10 veces menor que el de los genes codificantes [23]. Adicionalmente, este tipo de transcritos presentan una especificidad de expresión mucho mayor que otros genes (aproximadamente un 78 % se consideran específicos de tejido, frente a un 19 % de los mRNAs [23]). Los lncRNAs son además elementos clave en la regulación de multitud de mecanismos celulares, como interferencia transcripcional, activación de factores de transcripción, silenciado epigenético de genes y clusters de genes, regulación de *splicing* o inactivación de zonas promotoras, entre otros [315]. Muchos de estos procesos, como el silenciado de genes o la regulación de *splicing* guardan relación con enfermedades como el cáncer [26].

Este creciente conocimiento de funcionalidades de los lncRNAs ha servido de aliciente para un estudio más extenso de dichas entidades. En particular, la detección y anotación de lncRNAs, tanto experimental como computacionalmente,

ha sido objeto de numerosos avances. La aparición de tecnologías de secuenciación más avanzadas capaces de detectar transcripción en volúmenes más bajos ha sido clave en este aspecto. Entre las metodologías experimentales de secuenciación de ARN podemos destacar RNA-seq [45], la cual permite secuenciar el ARN transcrito en una muestra.

Dado que la transcripción es un proceso dinámico que no sólo varía en el tiempo sino también entre las distintas células de una muestra, hay que tener en cuenta esta variabilidad en los niveles de expresión de las distintas células que componen la muestra. Estas metodologías permiten secuenciar transcritos, dando como salida *reads* que pertenecen a dichos transcritos. Se requieren por tanto metodologías de ensamblado y alineamiento de ARN contra un genoma de referencia. Los reads se mapean contra el genoma utilizando herramientas como TopHat [321], STAR [322] o LAST [323]. Los transcritos de ARN son determinados posteriormente con herramientas como Cufflinks [324] y Scripture [325].

Una vez obtenidas las secuencias de los transcritos mediante dicho alineamiento, es necesario averiguar si un transcrito es traducido o no. Es en este paso donde entran las metodologías *ad hoc* para descubrimiento computacional de lncRNAs. Existen distintas aproximaciones que desempeñan esta función en base al análisis de secuencia, como PhyloCSF [326] y RNAcode [327]. También existen metodologías que se centran en el análisis de potencial codificante, como CPAT [328] y CPC [329], mediante ORFs (Open Reading Frames, secuencias que empiezan con un codón de inicio) o estadísticas de conteo de subsecuencias [330].

La gran cantidad de herramientas disponibles para la anotación y detección computacional de lncRNAs a partir de datos experimentales de tecnologías como RNA-seq ha mejorado sin duda la calidad y la extensión de las anotaciones de lncRNAs. En este sentido, un esfuerzo considerable está siendo acometido para reunir información sobre la localización de los lncRNAs en el genoma en bases de datos públicas. Como resultado de este trabajo, varias bases de datos con anotaciones de lncRNAs humanos han sido construídas y puestas a disposición de la comunidad científica, incluyendo más de 15.000 anotaciones de lncRNAs: GENCODE [28], lncRNAdb [29], LNCipedia [33] y NONCODE v4.0

[31] son algunas de ellas. Por tanto, una localización sistemática a lo largo de todo el genoma está disponible en la actualidad. Sin embargo, a pesar de este creciente conocimiento sobre la localización de los lncRNAs, la comprensión de los mecanismos reguladores en los que forman parte es a día de hoy una cuestión poco estudiada.

Como se ha mencionado con anterioridad, se ha demostrado que los lncRNAs forman parte de la maquinaria reguladora de los genes. Muchos de los posibles vínculos entre los SNPs relacionados con enfermedades y los lncRNAs son una incógnita, pese a que alrededor de una tercera parte de los SNPs relacionados con enfermedades identificados por GWAS pertenecen a regiones no codificantes [15]. Por otra parte, tal como muestra la figura 4.1, se sabe que los lncRNAs interactúan directamente con proteínas, ARNm y ADN [9], pero la funcionalidad o los resultados de dichas interacciones son todavía muy poco conocidos.

En el área de investigación que involucra interacción entre ARNs no codificantes y otras entidades, han surgido muy recientemente tecnologías con la capacidad de detectar la interacción de los transcritos con proteínas en su modificación post-transcripcional [331]. Estas incluyen enfoques como iClip [332], CLASH [333], CLIP [334], el muy reciente PAR-CLIP [335] o incluso CHART [336], un enfoque específico de ARN largos no codificantes, que permite obtener información de interacción lncRNA con proteínas y ADN. Sin embargo, estos enfoques suponen elevados costes económicos y temporales que podrían disminuir si una predicción computacional redujera el espacio de búsqueda de las hipótesis a comprobar. Hasta la fecha, no existe a nuestra disposición una anotación sistemática de TFBSs, SNPs, transcritos o ningún otro resultado de análisis de secuencia realizado para inferir la funcionalidad de los lncRNAs.

Es por ello que junto con estas mejoras experimentales van surgiendo algunas metodologías computacionales para extraer conocimiento de la gran cantidad de datos producida por ellas. Los resultados obtenidos de estos enfoques están siendo a su vez recogidos en bases de datos tales como CISBP-RNA [337], una base de datos de motivos de proteínas capaces de unirse al ARN (RNA-binding proteins, RBPs). Dado que las afinidades de unión de estas RBPs están modeladas

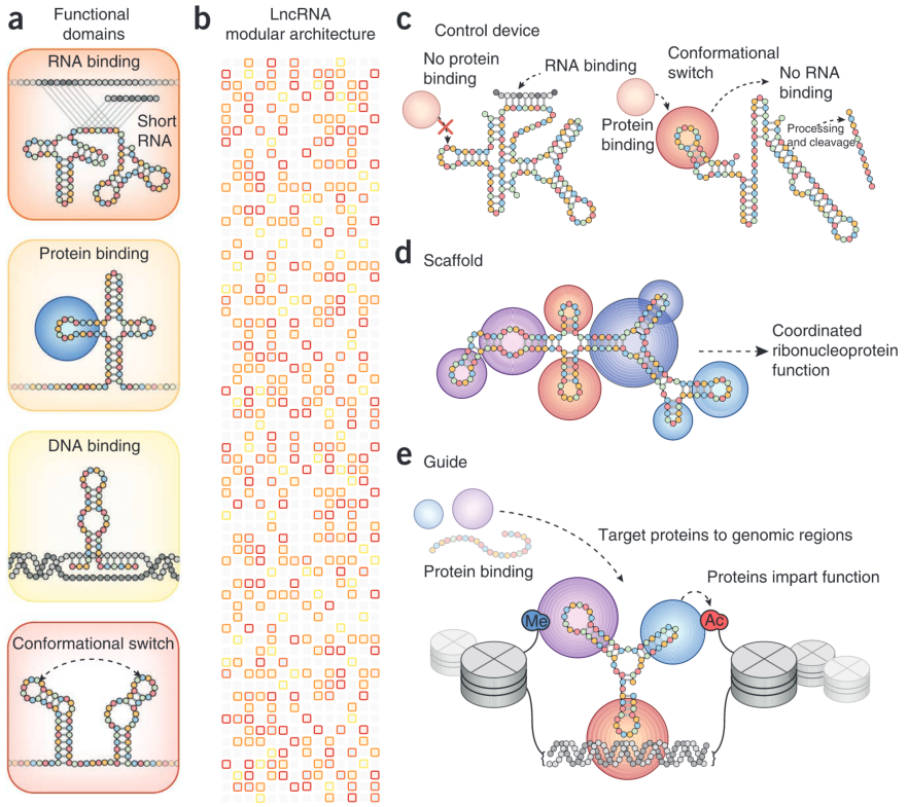


Figura 4.1: *Esquema de funcionalidades posibles de los lncRNA en regulación propiciadas por su estructura modular tomado de Mercer et al. [9]. (a) Los dominios funcionales del lncRNA pueden actuar uniéndose a otros ARN, proteínas o secuencias de ADN. (b) El splicing alternativo que manifiestan muchos lncRNAs hace de la suya una estructura modular, pudiendo encontrar en ellos varias de las funcionalidades mencionadas. Cada fila de la matriz representa un lncRNA y los colores representan las diversas actividades de sus dominios de acuerdo a los colores en a. (c) Los lncRNAs pueden actuar de mecanismo de control dependiendo de entidades externas. Por ejemplo, al unirse a un ARN (en gris en la imagen), puede evitarse que la proteína destinada a esa secuencia no pueda adherirse a ella, mientras que en caso contrario la proteína se podría adherir al ARN (como se muestra en la derecha), generando una conformación estructural distinta. (d) Los lncRNAs pueden actuar de andamiaje para formar estructuras proteicas complejas. (e) También pueden actuar como guía para llevar proteínas a sus lugares de unión.*

en términos de matrices de motivos, las herramientas utilizadas para establecer afinidad de unión ADN-proteína pueden ser adaptadas para predecir afinidad de unión ARN-proteína de una forma similar. Sin embargo, hasta la fecha sólo hemos encontrado disponibles herramientas *de novo* que extraen motivos de conjuntos de datos experimentales [338], tales como MEMERIS [339], RNAContext [340] o GraphProt [341]. La causa de esto probablemente sea la escasez de datos de este tipo hasta fechas muy recientes.

Existen por otra parte herramientas que se sobreponen a la escasez de bases de datos de motivos ARN-proteína buscando patrones en otras características propias del ARN, principalmente lo que se conoce como estructura secundaria del ARN. La figura 4.2 muestra un ejemplo de estructura secundaria para un lncRNA. Los transcritos de ARN tienden a plegarse sobre sí mismos formando esta estructura, la cual se caracteriza por secciones de ARN en forma de doble hebra, es decir, adheridas a otras secciones de los transcritos, y secciones *libres* que se quedan sin unir formando horquillas, bucles y otras estructuras. La mayoría de las herramientas computacionales de búsqueda de motivos *de novo* en secuencias de ARN suelen realizar primero una predicción de la estructura secundaria de todas las secuencias de ARN de entrada y a continuación utilizar dicha estructura para alinear, agrupar o filtrar partes de estas secuencias que cumplen la misma función estructural. De estas subsecuencias agrupadas se extraen los motivos o patrones específicos de ARN.

En general, el conocimiento existente sobre lncRNAs, así como el análisis de los mismos, es mucho más abundante en cuanto a la localización en el genoma del ADN que los transcribe que con respecto a otras características tales como lugares de unión de proteínas. Este hecho afecta a la limitada disponibilidad de anotaciones públicas sobre las que realizar estudios posteriores en profundidad. Por otra parte, existen estudios de lncRNAs concretos relacionados con funciones reguladoras y, en última instancia, con enfermedades. Para comprender los mecanismos por los cuales estos lncRNAs adquieren las funcionalidades mencionadas, un esfuerzo más reciente está teniendo lugar en cuanto al análisis de propiedades de los lncRNAs tales como las posibles interacciones de los lncRNAs

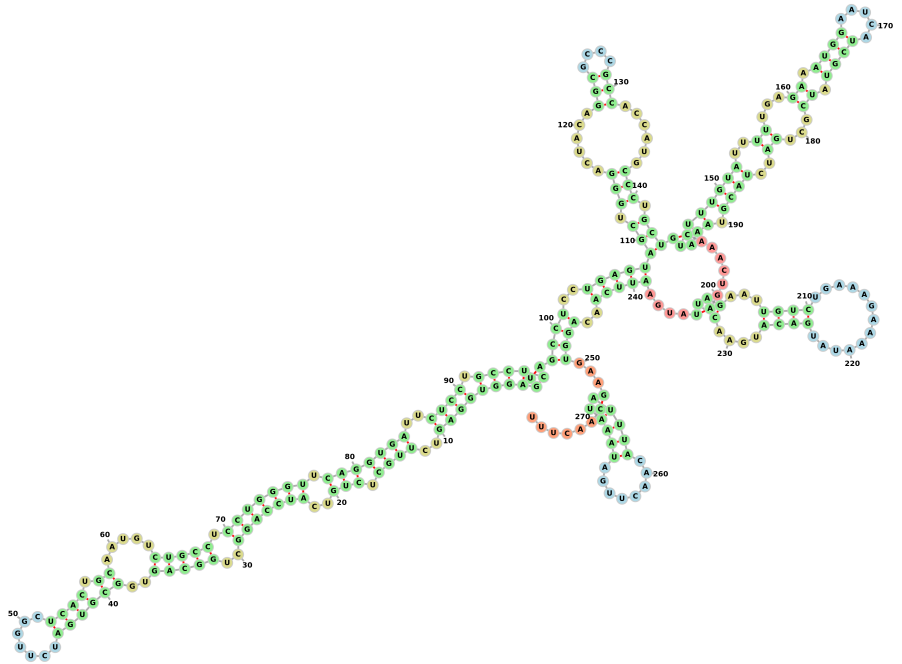


Figura 4.2: *Ejemplo de la estructura secundaria de un lncRNA, calculada usando la herramienta RNAfold [342] y visualizada en forna [343]. Los números indican la posición en la secuencia del transcript y los colores indican tipos de estructura generadas, por ejemplo, el verde representa secciones en forma de doble hélice.*

con otras entidades. Entre estos análisis, la rama experimental ha dado lugar a diversas metodologías para obtener información sobre interacción de lncRNAs con otras entidades como proteínas. Estas metodologías suelen ser específicas de proteína o familia de proteínas, por lo que podrían beneficiarse de herramientas computacionales capaces de seleccionar un conjunto de candidatos plausibles para la interacción.

La mayoría de las herramientas computacionales disponibles para esta tarea son metodologías *de novo* que utilizan como principal fuente de información la estructura secundaria del ARN, también predicha computacionalmente.

4.2. Búsqueda de lugares de unión lncRNA proteína

A continuación se presenta RNAIntuit [316], una metodología cuyo objetivo es obtener una lista priorizada de proteínas candidatas a adherirse a un conjunto de transcritos lncRNA de entrada, bajo la hipótesis de que estas proteínas están probablemente involucradas en vías de regulación post-transcripcional.

Como se ha detallado con anterioridad, la reciente aparición de bases de datos de motivos ARN-proteína favorece el planteamiento de herramientas que adapten lo que se conoce de las medidas de similitud motivo-secuencia para ADN a las particularidades del ARN. En este sentido, adaptamos SC_{Intuit} [147], una medida difusa de afinidad ADN-proteína basada en conjuntos intuicionistas, para evaluar afinidad de unión ARN-proteína. Dicha metodología se valida a continuación en dos casos de estudio, a dos conjuntos de transcritos de conocida relación con el cáncer: HOTAIR [344] y ANRIL [345]. Posteriormente, la metodología propuesta se extiende para aplicarla a un estudio mecanístico de Fendrr, un gen lncRNA relacionado con el cáncer, pero cuyos mecanismos son aún poco conocidos.

4.2.1. Descripción de la metodología

La metodología computacional propuesta pretende descubrir interacciones ARN-proteína mediante el análisis de las secuencias de los transcritos en busca de lugares de unión de RBPs. La figura 4.3 muestra el pipeline propuesto, el cual consta de varios pasos principales:

1. **Extracción de GENCODE del conjunto de transcritos de interés.**

   Primero, los transcritos de interés son extraidos de GENCODE. Cabe recordar que para un mismo gen de lncRNA pueden obtenerse varios transcritos, si estos sufren *splicing* alternativo. Así pues, el conjunto de transcritos generados por el gen de entrada son todos extraidos, para calcular a partir de sus secuencias su estructura secundaria.

2. **Predicción de estructura secundaria del ARN.**

   La razón de realizar una predicción de la estructura secundaria del ARN previamente a la búsqueda de motivos en las secuencias es la evidencia

de que existe una preferencia de las proteínas que se adhieren al ARN de buscar las secciones de ARN que quedan libres tras su plegado [346]. Esta predicción se realiza utilizando la herramienta RNAfold, disponible en el paquete de software Vienna Package [342]. El cálculo de dicha estructura fue realizado utilizando el modelo MFE (Minimum Free Energy-mínima energía libre) mediante el algoritmo de Zucker y Stiegler [347], con los parámetros de energía por defecto del grupod de Turner [348]. RNAfold proporciona una salida en formato texto denominada *notación punto-paréntesis*[1] para la que cada nucleótido de la secuencia se corresponde con un carácter, siendo paréntesis nucleótidos emparejados y el carácter punto los nucleótidos que quedan libres, por ejemplo:

```
GGGCUAUUAGCUCAGUUGGUUAGAGCGCACCCCUGAUAAGGGUG
((((((((((((.........))))))(((((.......)))))
```

En este caso, las secuencias que serían utilizadas serían `AGUUGGUUA` y `CUGAUAA`.

3. **Escaneo de motivos en secuencias de ARN de cadena única.**
   Las secuencias libres extraídas después de la predicción de RNAfold son a continuación escaneadas utilizando una adaptación de SC_{intuit} [147] a secuencias de ARN. Así, sólo es considerada una hebra (en ambos sentidos) en lugar de considerar la doble hélice del ADN. Los scores proporcionados por SC_{intuit} se filtran por un umbral de corte definido por el usuario, el cual, experimentalmente y debido a las características de los motivos de RBPs, suele elegirse por encima de 0.7. El ranking de resultados obtenido es, en última instancia, comparado con los resultados de StarBase para los transcritos lncRNA elegidos.

4.2.2. Validación en lncRNAs relacionados con cáncer

Para validar la metodología propuesta, la aplicamos a un set de transcritos asociados a dos lncRNAs bien caracterizados, cuya relación con el cáncer es

[1] Del inglés dot-bracket notation.

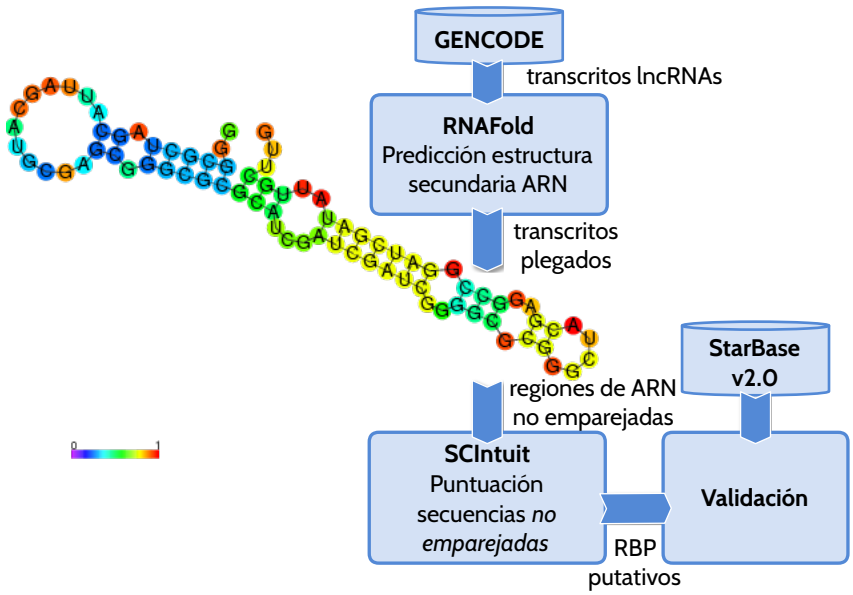


Figura 4.3: *Pipeline de procesamiento de la primera versión de RNAIntuit. Primero, un conjunto de transcritos de un lncRNA de interes son extraidos de GENCODE. Estos transcritos son la entrada de RNAfold, herramienta que efectúa una predicción sobre su estructura secundaria. Esta estructura es procesada para seleccionar las secuencias de cadena única, las cuales se corresponden con los puntos que se muestran en el diagrama. Los paréntesis se corresponden con nucleótidos emparejados. Se muestra también en la imagen un ejemplo de una estructura secundaria de lncRNA predicha por RNAfold, donde los colores representan la probabilidad de unión/libertad de dichas bases. La versión adaptada a ARN de SCIntuit se ejecuta después sobre las secuencias seleccionadas y una tabla de resultados con lugares de unión de RBPs cuyo score sea mayor que el corte (0.75) es generada. Por último, estos resultados son validados consultando StarBase.*

ampliamente conocida: HOTAIR [344] y ANRIL [345]. Las interacciones de estos lncRNAs con proteínas se extrajeron de StarBase 2.0 [100] y este dataset ha sido utilizado para validar los resultados computacionales.

Los transcritos de HOTAIR y ANRIL se extrajeron de GENCODE v21 [28] y se procesaron usando RNAfold. A continuación se seleccionaron las secuencias de cadena única y estas fueron escaneadas por la versión adaptada de SCIntuit, obteniendo así una lista de lugares de unión ordenada por su score de afinidad de unión. Para los 11 transcritos de HOTAIR, de una longitud total de 8762 nucleotidos, 3484 (39.76 %) fueron considerados de cadena única por RNAfold. Al aplicar SCIntuit a estas secuencias se obtuvo una lista de 1659 lugares de unión de RBPs utilizando un umbral de corte de 0.75, basado en la experiencia y resultados previos. De esta lista, 157 (9.46 %) fueron encontrados en StarBase. Adicionalmente, si un lugar de unión candidato es predicho una gran cantidad de veces en un conjunto de transcritos de ARN, es más probable que exista una interacción entre estos transcritos y la RBP correspondiente.

El mismo procedimiento fue aplicado a un conjunto de 17 transcritos del gen ANRIL extraidos de GENCODE v21, con una longitud total de 21813 nucleótidos de los cuales 8801 (40.34 %) eran de cadena única según RNAfold. Una lista de 3841 resultados fue obtenida, de los cuales 353 (9.19 %) se correspondieron con interacciones RBP-lncRNA en StarBase.

Los resultados muestran que las predicciones de interacciones lncRNA-proteína obtenidas mediante la metodología propuesta son coherentes con el conocimiento actual sobre dichas interacciones, pese a la reducida cantidad de datos disponibles. Sin embargo, dicho conocimiento existente es a su vez reducido, por lo que no existe un estándar que nos habilite a hacer un benchmarking exhaustivo de los resultados obtenidos. Con posterioridad al desarrollo de esta metodología, y como se explica en la sección a continuación, esta metodología ha sido refinada y ampliada para su aplicación a un caso mucho más complejo que involucra no sólo lncRNAs y proteínas de unión a ARN, sino también factores de transcripción y posibles interacciones directas entre ARN y ADN.

4.3. Fendrr y su relación con la carcinogénesis

La carcinogénesis ha sido extensamente estudiada desde un punto de vista molecular, y recientemente ha entrado en la era los ARNs largos no codificantes. El gen lncRNA Fendrr está ubicado en el cromosoma 16 (16q24.1). Este gen produce un lncRNA transcrito bidireccionalmente, con el gen del factor de transcripción forkhead box-F1 (FoxF1) en la hebra opuesta.

Existe evidencia contradictoria para apoyar la correlación entre Fendrr y FoxF1. Szafranski *et al.* identificaron que Fendrr regula de forma negativa FoxF1 [349], mientras que Cabili *et al.* mostraron que la expresión y la localización subcelular de los ARNm de Fendrr y FoxF1 están en la misma proporción y compartimentados (núcleo y citpolasma) [23]; por otro lado, Grote y Herrmann indicaron que Fendrr interactúa con el complejo PRC2, el cual inhibe la expresión de FoxF1 [350].

A continuación se describe una metodología computacional integradora específica para un estudio mecanístico del lncRNA Fendrr y su relación con el gen FoxF1 [317]. Dicha metodología aplica la metodología propuesta en la sección 4.2 para predecir interacciones lncRNA-proteína, RNAIntuit, integrándola en un pipeline de análisis más complejo.

4.3.1. Formas de interacción entre lncRNAs y regiones promotoras

Con el fin de comprender mejor los mecanismos moleculares por los cuales Fendrr podría regular la expresión de FoxF1 en cáncer de pulmón, se ha realizado un análisis *in silico* para evaluar posibles interacciones, directas e indirectas, entre los transcritos de Fendrr y la región promotora de FoxF1.

Los lncRNAs, además de poder adherirse a proteínas RBPs, también pueden interactuar con el ADN formando triples ADN:ADN-RNA [351] e incluso reclutar factores de transcripción [352], funcionando como guía de los mismos a sus lugares de unión, sirviendo de componente estructural para complejos más sofisticados o funcionando al contrario como señuelo, evitando que algunos TFs llegen a los lugares de unión a los que estaban destinados, entre otras posibles funcionalidades [24, 9]. Basándonos en este conocimiento biológico reciente sobre cómo ARN, ADN

y proteínas pueden interactuar entre sí, realizamos un análisis computacional de interacción entre los transcritos de Fendrr y la región promotora de FoxF1 a dos niveles. Este estudio se realizó con el fin de encontrar hipótesis plausibles para explicar los mecanismos subyacentes al papel regulador de Fendrr sobre el gen FoxF1. Los niveles de interacción dadas estas entidades biológicas pueden ser dos en términos de complejidad:

1. **Interacción directa.** Los transcritos de Fendrr podrían interactuar con la región promotora de FoxF1 directamente, formando triples ARN-ADN:ADN.
2. **Interacción indirecta.** Los transcritos de Fendrr pueden interactuar con la región promotora de FoxF1 a través de proteínas mediadoras, tales como RBPs o factores de transcripción.

Nótese que estos dos niveles de interacción son asimismo no excluyentes, siendo posible que un transcrito que se una a una sección de ADN formando un triple ARN-ADN:ADN, interactúe a su vez con otras proteínas, funcionando como señuelo, guía o andamiaje.

Descripción de los datos. La anotación de Fendrr y sus 7 transcritos correspondientes fueron extraidos de la última versión de GENCODE al momento de realizar este experimento, GENCODEv23 [320]. La secuencia completa de estos 7 transcritos ha sido utilizada. Las coordenadas que aparecen en GENCODEv23 se corresponden con la versión del genoma GRCh38, referencia de la cual se extrajo la región promotora de FoxF1 de 1428bp de longitud, correspondiente a las posiciones chr16:86509099-86510527. Los motivos de factores de transcripción y RBPs se obtuvieron de varias fuentes. De JASPAR Core 2016 [129] se obtuvieron 519 motivos. 98 motivos de RBP humanos se utilizadon provenientes de CISBP-RNA [337] y 426 motivos de la base de datos HOCOMOCOv9 [353], descargados de su versión MEME en la web de MEME-suite `http://meme-suite.org` para poder ser utilizadas con la herramienta FIMO.

4.3.2. Predicción de interacción directa Fendrr-FoxF1

Los lncRNAs pueden formar triples ADN:ADN-RNA con secciones de ADN afines [351]. Estos triples se vinculan de acuerdo a reglas de formación de triples, las

cuales se basan en las propiedades químicas de los ácidos nucleicos [354]. En los últimos años han aparecido algunas metodologías computacionales para el cálculo de triples ARN-ADN:ADN , tales como Triplexator [355], un método que predice sitios de una longitud dada donde un transcrito de ARN podría unirse; y Triplex Domain Finder [351], una herramienta que toma como entrada un conjunto de regiones promotoras de ADN, las cuales se presume son reguladas por un cierto transcrito de ARN, y busca dominios de unión en el ADN.

Para obtener información sobre la posible unión de los transcritos de Fendrr a la región promotora de FoxF1 y de qué manera, aplicamos Triplexator [355], ya que se centra en el análisis de una única secuencia, mientras TDF requiere un conjunto de regiones promotoras entre las cuales exista una cierta relación de co expresión. TDF utiliza internamente Triplexator para encontrar triples de una forma a una escala más de genoma completo y utiliza los transcritos de Fendrr como control positivo, ya que se ha confirmado que Fendrr forma triples ARN-ADN:ADN en diversas localizaciones del genoma [350]. En su análisis, los autores de TDF declaran que FoxF1 es una zona promotora candidata para formar un triple con un transcript de Fendrr [351]. Sin embargo, no se proporciona información específica sobre la ubicación, longitud y composición de dicho triple. Para obtener esta información específica, aplicamos Triplexator en la región promotora de FoxF1.

Los parámetros utilizados para Triplexator son muy similares a los utilizados en la guía de usuario de la herramienta: longitud mínima de 15 nucleótidos, teniendo como máximo un 20 % de errores, donde no se permiten errores consecutivos, con un porcentaje de guanina de al menos un 20 %. Adicionalmente, se ha desactivado el filtrado de repeats de baja complejidad en la secuencia, ya que se ha sugerido que Fendrr se une justo a ese tipo de repeats [356].

La tabla 4.1 muestra los resultados obtenidos para la región promotora de FoxF1. Además, se ha realizado el mismo test para la longitud completa del gen FoxF1, incluyendo la región promotora (5365 bp), y no se obtuvieron nuevos resultados. Estos resultados solapan además parcial o totalmente con regiones

Tabla 4.1: *Dos transcritos de Fendrr muestran afinidad para formar un triple ARN-ADN:ADN en la región promotora de FoxF1 de acuerdo a Triplexator. Las posiciones son relativas al comienzo del transcrito y de la región promotora, respectivamente. Los cuatro resultados forman dos regiones que solapan con regiones conservadas.*

ID Transcrito	Pos. transcrito	Pos. ADN	Hebra ADN
ENST00000598996.2	1928-1943	1160-1175	+
ENST00000599749.5	146-161	1170-1185	+
ENST00000599749.5	144-159	1253-1268	+
ENST00000599749.5	1967-1982	1160-1175	+

conservadas del genoma, sugiriendo que estas regiones de formación de triples ARN-ADN:ADN podrían tener funcionalidad.

4.3.3. Interacción Fendrr-FoxF1 mediada por proteínas.

En un nivel ligeramente más indirecto de interacción, los transcritos de ARN pueden interactuar con ADN a través de proteínas mediadoras. Hay muchas formas en las cuales los lncRNAs pueden cooperar con proteínas, actuando en diversas funcionalidades. Por ejemplo, los lncRNAs pueden adherirse a proteínas que iban a unirse a secuencias de ADN y evitar este resultado, funcionando como señuelo, o al contrario: pueden funcionar como guía para esas proteínas [9]. Dado que ya se ha sugerido una interacción directa entre Fendrr y la región promotora de FoxF1, es también de interés evaluar si existe además interacción indirecta entre TFs que podrían unirse a TFBSs en la región promotora y el lncRNA.

Factores de transcripción como mediadores. Para predecir factores de transcripción candidatos a actuar de mediadores entre FoxF1 y Fendrr, se realizaron estudios con FIMO, una herramienta para similitud motivo-secuencia [357]. FIMO es una herramienta estadísticamente robusta que convierte un ratio de probabilidades que convierte posteriormente a un p-valor asumiendo una hipótesis nula de orden cero. Todas las matrices en JASPAR Core 2016 vertebrata [129] y HOCOMOCOv9 [353] se han utilizado con este propósito. Los parámetros para FIMO son los parámetros por defecto, pero adaptado a una única hebra para el

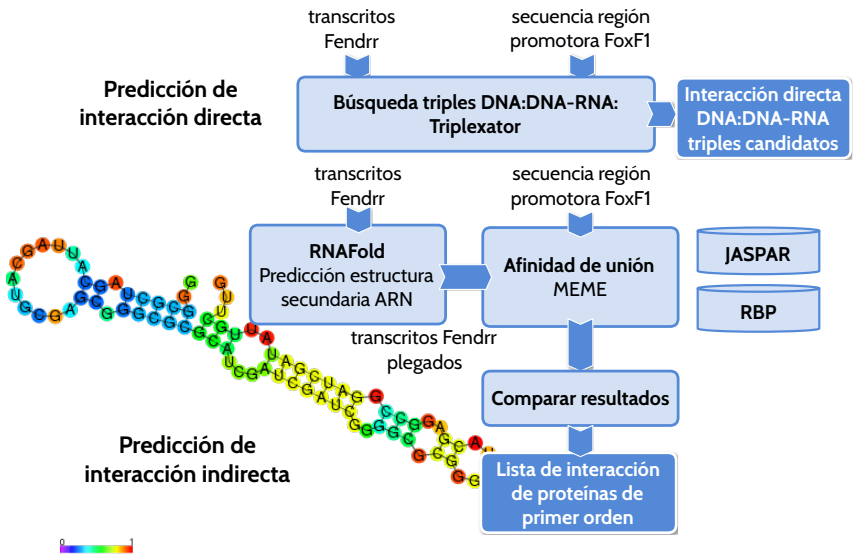


Figura 4.4: *La metodología propuesta para el estudio de Fendrr distingue dos niveles de interacción. El primero calcula mediante Triplexator [355] posibles triples de interacción entre la región promotora de FoxF1 y Fendrr, dado que existe evidencia de que los transcritos se pueden vincular a secuencias de ADN formando triples ARN-ADN:ADN. El segundo nivel explora interacción mediada por proteínas, escaneando los transcritos de Fendrr mediante la metodología propuesta en RNAintuit y la secuencia promotora de Fendrr en busca de lugares de unión de RBPs y TFBSs. Ambas búsquedas son cruzadas para buscar resultados positivos en ambas secuencias. Estos positivos son candidatos a ser mediadores en la regulación Fendrr-FoxF1.*

caso de transcritos ARN. Los transcritos de Fendrr también han sido escaneados en la búsqueda de afinidad con los factores de transcripción en dichas bases de datos bajo las posibles hipótesis en las que Fendrr actúa como señuelo o guía para estos factores de transcripción. En este caso, se ha tenido en cuenta también la estructura secundaria del ARN y se ha calculado dicha estructura usando RNAfold [342] para comparar resultados para transcritos plegados y no plegados.

Una vez plegados, sólo las partes del transcrito que quedan de única hebra se utilizan para la predicción, ya que existe evidencia experimental de que las proteínas que se vinculan al ARN tienen preferencia por estas secciones [358]. Los resultados en los transcritos plegados tienden a ser escasos, debido en parte a la baja cantidad de secuencia, mientras que son muy numerosos cuando la secuencia completa es considerada. Para reducir el número de falsos positivos, buscamos una representación más precisa de la forma en que las proteínas podrían unirse a ARN plegado e implementamos una aproximación intermedia, donde se otorga preferencia al ARN de cadena simple pero no se restringe sólo a dichas secuencias. Para ello, incrementamos las secuencias de cadena única proporcionadas por RNAfold en un número de nucleótidos alrededor de las mismas, considerando que un TF podría unirse parcialmente a estas secuencias [352].

Dado que la longitud media de los motivos de Jaspar Core 2016 es 10.64 y que la longitud media de las secuencias de cadena única dadas por RNAfold en nuestro conjunto de datos es de 3.19, hemos observado que una extensión de 1-2 nucleótidos alrededor de cada secuencia de cadena única aumenta la media a 6.19 y 13, respectivamente, mientras se sigue reduciendo la cantidad de secuencia procesada a un 40.77 % y un 23.51 %, respectivamente, reduciendo así el espacio de búsqueda. Este procedimiento ayuda a encontrar TFBSs que de otra forma serían ignorados, mientras se siguen priorizando las secuencias de cadena única en la estructura secundaria del ARN.

La tabla 4.2 muestra los resultados en la región promotora de FoxF1 que son también encontrados de forma significativa en los transcritos plegados de Fendrr (TFs que muestran alta afinidad de unión tanto al transcrito lncRNA como a la secuencia de ADN, por tanto candidatos a mediación lncRNA-región promotora).

Tabla 4.2: *TFs candidatos a la mediación entre transcritos de lncRNA Fendrr y la región promotora de FoxF1. Los 15 primeros resultados significativos ($pvalor < 10^{-4}$) obtenidos con FIMO (de un total de 26 resultados) que aparecen tanto en los transcritos de Fendrr como en la región promotora de FoxF1, para 519 matrices de JASPAR Core 2016 [129]. Los resultados en negrita aparecen en la literatura biomédica como relacionados con el reclutamiento de las proteínas del grupo polycomb PRC1 y PRC2 [359].*

ID matriz JASPAR	Nombre TF	Posición ARN	Posición ADN	pvalor ARN	pvalor ADN
MA0146.2	Zfx	178-191 +	1119-1132 +	$1{,}34 \times 10^{-6}$	$1{,}10 \times 10^{-5}$
MA0073.1	RREB1	474-493 +	1066-1085 +	$2{,}77 \times 10^{-6}$	$3{,}54 \times 10^{-5}$
MA0119.1	NFIC::TLX1	261-274 -	1213-1226 +	$3{,}13 \times 10^{-6}$	$3{,}81 \times 10^{-5}$
MA0057.1	MZF1(var.2)	151-160 +	883-892 +	$3{,}59 \times 10^{-6}$	$9{,}27 \times 10^{-5}$
MA0138.2	**REST**	**2177-2197 -**	**1241-1261 +**	$1{,}33 \times 10^{-5}$	$5{,}70 \times 10^{-5}$
MA0506.1	NRF1	175-185 +	471-481 +	$2{,}44 \times 10^{-5}$	$2{,}35 \times 10^{-5}$
MA0516.1	SP2	53-67 +	1352-1366 +	$2{,}47 \times 10^{-5}$	$3{,}02 \times 10^{-6}$
MA0597.1	THAP1	52-60 +	475-483 +	$2{,}83 \times 10^{-5}$	$2{,}83 \times 10^{-5}$
MA0149.1	EWSR1-FLI1	424-441 +	895-912 +	$3{,}13 \times 10^{-5}$	$3{,}46 \times 10^{-5}$
MA0162.2	EGR1	84-97 -	1391-1404 +	$3{,}18 \times 10^{-5}$	$3{,}92 \times 10^{-6}$
MA0747.1	SP8	154-165 -	1352-1363 +	$3{,}92 \times 10^{-5}$	$6{,}02 \times 10^{-5}$
MA0517.1	STAT1::STAT2	227-241 -	439-453 +	$4{,}18 \times 10^{-5}$	$1{,}50 \times 10^{-6}$
MA0872.1	TFAP2A(var.3)	172-184 +	464-476 +	$4{,}47 \times 10^{-5}$	$8{,}57 \times 10^{-6}$
MA0039.2	Klf4	706-715 -	25-34 +	$4{,}49 \times 10^{-5}$	$2{,}12 \times 10^{-5}$
MA0471.1	**E2F6**	**2268-2278 +**	**1261-1271 +**	$4{,}53 \times 10^{-5}$	$2{,}66 \times 10^{-6}$

Las secuencias de ARN de cadena única fueron expandidas en dos nucleótidos de acuerdo a la longitud media de los motivos de JASPAR. Posteriormente se realizó una búsqueda manual en la literatura para comprobar que los TF destacados en la tabla, REST y E2F6, se relacionan con expresión génica a través del reclutamiento de las proteínas de grupo polycomb PRC1 y PRC2 [359]. Esto, sumado a la proximidad de sus lugares de unión putativos en la secuencia de ADN (como se muestra en la figura 4.5) y la interacción directa ARN-ADN:ADN predicha por Triplexator los postulan como los actores principales en la maquinaria reguladora Fendrr-FoxF1.

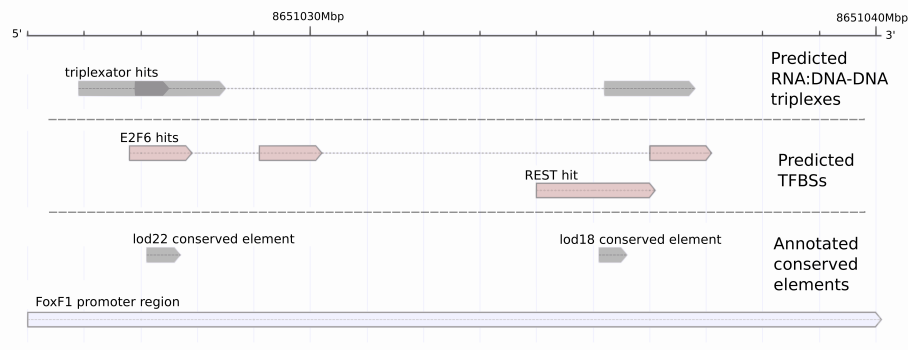


Figura 4.5: *Posiciones relativas entre los elementos reguladores putativos involucrados en la interacción entre el lncRNA Fendrr y la región promotora de FoxF1. La región ilustrada corresponde a una zona de 150bp de longitud en la región promotora de FoxF1 localizada en las coordenadas chr16:86510250-86510400 del genoma. La anotación en la parte superior se corresponde con los triples ARN-ADN:ADN predichos por Triplexator entre los transcritos de Fendrr y la región promotora de FoxF1. La segunda línea de anotaciones corresponde con los TFBSs de los factores de transcripción E2F6 y REST predichos por FIMO. Estas dos regiones próximas del genoma solapan con dos elementos conservados en la pista de conservación de 100 vertebrados de Genome Browser [360]. Las anotaciones fueron dibujadas utilizando AnnotationSketch[361] y vectorizadas manualmente.*

RBPs como mediadoras. Los transcritos de ARN interactúan a menudo con RBPs [362], proteínas que pueden tomar parte en muchas funciones celulares, tales como el transporte del ARN y la localización celular o biogénesis [363]. El mismo

enfoque utilizado para encontrar factores de transcripción que podrían unirse tanto a los transcritos de Fendrr como a la región promotora de FoxF1 se aplicó a los motivos presentes en la base de datos CISBP-RNA [337], y los resultados fueron comparados. Tal y como se muestra en la tabla 4.3, tras una búsqueda manual de estos resultados en la literatura biomédica, se muestra que la mayoría de los resultados positivos de RBPs están directamente relacionados con funcionalidad específica de ARN.

En esta sección se ha aplicado una metodología computacinoal para explorar dos niveles de posible interacción molecular entre un par de entidades biológicas de interés: los transcritos del lncRNA Fendrr y FoxF1, un gen cercano que se cree relacionado con dichos transcritos a través de un vínculo regulador. La conclusión que hemos alcanzado es que esta interacción es posible en ambos niveles, tanto directo como indirecto.

La abundancia de entidades biológicas que normalmente interactúan en la maquinaria de una célula presenta un reto cuando el objetivo es un estudio mecanístico. En este sentido, creemos que la metodología aplicada puede ser utilizada para numerosos casos similares. Sin embargo, la aplicación de esta metodología sí requiere un conocimiento *a priori* sobre la relación entre las entidades iniciales, esto es, Fendrr y FoxF1. Por otra parte, otras capas adicionales de información, tales como conservación, marcas de histonas u otras marcas epigenéticas pueden ser añadidas para encontrar entidades mediadoras con mayor precisión.

Adicionalmente, podría explorarse un segundo nivel de interacción más indirecta introduciendo redes de interacción proteína-proteína, lo que podría dar información sobre complejos proteínicos más sofisticados como mediadores de la regulación.

Tabla 4.3: *RBPs candidatas a la mediación entre transcritos de lncRNA Fendrr y la región promotora de FoxF1. Los motivos significativos de RBPs encontrados por FIMO tanto en la región promotora de FoxF1 como en los transcritos de Fendrr (plegados con RNAfold y extendidos en 2 nucleótidos). Resultados repetidos fueron filtrados. Los resultados destacados no se han encontrado relacionados directamente con función específica de ARN.*

ID Matriz	Nombre RBP	Posición ADN	Posición ARN	pvalor ADN	pvalor ARN
M050_0.6	RBM4	1269-1275 +	9-15 +	$2{,}96 \times 10^{-5}$	$2{,}96 \times 10^{-5}$
M109_0.6	RBM4B	1269-1275 +	9-15 +	$2{,}96 \times 10^{-5}$	$2{,}96 \times 10^{-5}$
M044_0.6	**PPRC1**	1342-1348 +	217-223 +	$2{,}96 \times 10^{-5}$	$8{,}88 \times 10^{-5}$
M151_0.6	HNRNPH2	32-38 +	89-95 +	$3{,}60 \times 10^{-5}$	$7{,}21 \times 10^{-5}$
M126_0.6	SRSF4	1353-1359 -	68-74 +	$3{,}60 \times 10^{-5}$	$3{,}60 \times 10^{-5}$
M043_0.6	**PCBP2**	682-688 -	650-656 +	$4{,}39 \times 10^{-5}$	$4{,}39 \times 10^{-5}$
M086_0.6	SRSF12	1179-1185 +	44-50 +	$5{,}35 \times 10^{-5}$	$5{,}35 \times 10^{-5}$
M177_0.6	**PCBP1**	1142-1148 +	2850-2856 +	$5{,}35 \times 10^{-5}$	$5{,}35 \times 10^{-5}$
M027_0.6	HNRNPL	16-22 -	499-505 +	$6{,}51 \times 10^{-5}$	$6{,}51 \times 10^{-5}$
M169_0.6	hnRNPLL	16-22 -	499-505 +	$6{,}51 \times 10^{-5}$	$6{,}51 \times 10^{-5}$

4.4. Motivos específicos de ARN mediante desequilibrio de hebras

Los lncRNAs se consideran moléculas modulares formadas por dominios funcionales y motivos. Sin embargo, sólo una pequeña cantidad de ellos ha sido descrita en detalle. En esta sección se presenta una anotación a nivel de genoma completo de motivos específicos de lncRNAs basado en un pipeline de análisis de secuencia de ARN llamado **MoSI** (Motif Strand Imbalance - Motivos de desequilibrio de hebras). A diferencia de otras herramientas de identificación de motivos en secuencias de ARN, MoSI utiliza la **dirección de transcripción** del ARN.

En este capítulo se han presentado enfoques computacionales para la búsqueda de elementos reguladores presentes en secuencias concretas de ARN, más específicamente, en lncRNAs. Dicha búsqueda se ha realizado adaptando metodologías específicas de ADN a ciertas particularidades del ARN como puede ser su estructura secundaria. Además, se han utilizado datos concretos de proteínas para las cuales existe evidencia experimental de su adherencia a secuencias de ARN (lugares de unión de RBPs). A continuación se presenta una metodología computacional para la búsqueda de motivos específicos de ARN que prescinde del uso de matrices de motivos como información de entrada, es decir, se trata de una búsqueda de motivos específicos de lncRNAs *de novo*.

La problemática que entraña la búsqueda de motivos *de novo* en secuencias de ARN no sólo engloba toda la casuística que involucra el mismo tipo de búsqueda en secuencias de ADN. Se añade además una nueva dificultad: dado que la secuencia de ARN es transcrita a partir de una cadena de ADN que hace de plantilla, el contenido de secuencia es el mismo, por lo que es difícil distinguir entre motivos de ADN y motivos propios de ARN. La forma de distinguir ambos tipos de motivos radica en el uso de características extra que diferencien al ARN de su ADN de origen. De esta forma, la ya mencionada estructura secundaria del ARN es parte clave para gran cantidad de metodologías de búsqueda de motivos de ARN [338]. En estos enfoques, primero se realiza una predicción de la estructura secundaria de

todas las secuencias de ARN de entrada y a continuación se utiliza dicha estructura para alinear, agrupar o filtrar partes de estas secuencias que cumplen la misma función estructural. De estas subsecuencias agrupadas se extraen los motivos o patrones específicos de ARN.

En la presente sección, sin embargo, se utiliza otra característica inherente al ARN: la **hebra de transcripción**. El ARN tiene una direccionalidad que el ADN no posee. Los genes, y al igual que ellos, los lncRNAs, se transcriben a partir de una de las dos hebras de ADN. La hipótesis sobre la que trabaja el enfoque propuest es que la hebra de transcripción no es algo aleatorio. Esto quiere decir que, dado un conjunto de secuencias de ARN, es de esperar cierta *direccionalidad* en su composición. Hasta la fecha, se ha utilizado lo que se conoce como **equilibrio de hebras**[2] en el ámbito de la evaluación de calidad en experimentos de secuenciación NGS. Al contrario que el ARN, el ADN tiene forma de doble hélice, y en una secuenciación a genoma completo se espera que exista la misma proporción de *reads* de una hebra que de la complementaria. Dicho de otro modo: el estudio del equilibrio de hebras en un conjunto de *reads* puede sacar a relucir sesgos o artefactos en los experimentos de secuenciación incluso en ausencia de un genoma de referencia [364, 365].

De modo análogo a las metodologías que recurren al equilibrio de hebras para evaluar la calidad de un experimento de secuenciación o corregir artefactos, nosotros pretendemos utilizar el desequilibrio de hebras esperado en secuencias de ARN para encontrar motivos específicos de ARN en un conjunto de transcritos de entrada.

4.4.1. Medida de desequilibrio de hebras

Dado el alfabeto de entrada $L = \{A, C, G, T\}$, se define un **k-mer** como una palabra de longitud k en L: $K = x_1 x_2 ... x_k$ donde $x_i \in L, \forall i \in \{1, ..., k\}$. Se define K^* como el conjunto de todas las palabras posibles de longitud k en L, de forma que $|K^*| = 4^k$.

[2] Traducción del concepto en inglés strand balance.

Para todo k-mer k, se define k^- como su reverso complementario, es decir:

$$K^- = x_k^- x_{k-1}^- ... x_1^-$$

donde x_i^- es la base complementaria a x_i de acuerdo a la unión de las bases en la doble hélice del ADN, es decir, A y T son mutuamente complementarias, al igual que lo son C y G. Por ejemplo, dado el kmer $K = ACCG$ donde $k = 4$, su complementario será $K^- = CGGT$. Por claridad designaremos una pareja constituida por un k-mer y su complementario K^+ y K^-, siendo K^+ por convenio el k-mer sobrerrepresentado.

Dada una secuencia $S = s_1 s_2 ... s_l$ de longitud l en L tal que $l \geq k$, se define la *frecuencia* f de K en S como el número de veces que se puede encontrar K en S.

En base a esta frecuencia, en la presente sección se definen tres medidas para el cálculo del índice de desequilibrio de hebras de un k-mer K. Nótese que no tiene sentido evaluar el desequilibrio de hebras en k-mers que son simétricos respecto al reverso complementario, por lo que de aquí en adelante, se excluirán del análisis en K^* los $4^{\lfloor k/2 \rfloor}$ k-mers tales que $K^+ = K^-$.

Dado un k-mer K^+ se define su índice de sobrerrepresentación sc_{fold} como:

$$sc_{fold} = \frac{f(K^+)}{f(K^-)}$$

Si asumimos el equilibrio como la hipótesis nula que pretendemos rechazar en nuestro experimento de desequilibrio de hebras, podemos esperar que la probabilidad de equilibrio para un k-mer y su complementario siga una distribución binomial, $P_{balance} \sim Binom(f(K^+) + f(K^-), 1/2)$. Por tanto, podemos establecer la el p-valor correspondiente como:

$$P_{balance}(K^+) = C_{f(K^+)}^{f(K^+)+f(K^-)} \, p^{f(K^+)} (1-p)^{f(K^-)}$$

Este p-valor será más cercano a cero cuanto mayor el desequilibrio, proporcional a la cantidad de experimentos, esto es, a las frecuencias del k-mer en cuestión y su complementario. Esta probabilidad sirve para evaluar una pareja de k-mers, ya que ambas frecuencias suman la totalidad de los experimentos del

test binomial y p vale 1/2, ya que se espera equilibrio, por lo que $P_{balance}(K^+) = P_{balance}(K^-)$.

Para conseguir una medida de sobrerrepresentación que tenga en cuenta también las frecuencias de ambos k-mers, no sólo la proporción entre ellos, de la misma forma que el test binomial mencionado anteriormente, pero a la vez con capacidad para puntuar tanto k-mers sobrerrepresentados como aquellos que están infrarrepresentados, se define la función de **índice de desequilibrio de hebras** $score_{SI}$ como una aplicación:

$$sc_{SI} : K^* \rightarrow \mathbb{R}$$

en forma de combinación lineal de la frecuencia $f(K^+)$ y el índice de sobrerrepresentación sc_{fold}:

$$sc_{SI}(K^+) = \alpha(f(K^+)) + \beta sc_{fold}(K^+)$$

Los parámetros α y β permiten modificar el nivel de importancia de la frecuencia de un k-mer con respecto a sc_{fold}. Nótese que cuando $\beta = 1$ y $\alpha = 0$, $sc_{SI} = sc_{fold}$.

4.4.2. Estudio de desequilibrio de hebras en lncRNAs intergénicos

A continuación se pretende demostrar que la señal de desequilibrio de hebras, medida en los términos detallados en el apartado anterior, es mayor en un conjunto de transcritos de lncRNAs de entrada que la que se obtendría de un conjunto de secuencias de ADN. Primero se describirá el conjunto de datos de lncRNA seleccionados así como su preprocesamiento. A continuación se describirá un experimento de randomización de **montecarlo** con el que rechazamos la aleatoriedad de la señal de desequilibrio de hebras en nuestro conjunto de datos. Por último, se utilizará un conjunto de hexámeros con función conocida y específica de ARN para evaluar la capacidad predictiva de las medidas sc_{fold} y sc_{SI}.

Conjunto de lncRNAs intergénicos de GENCODE. Con el fin de evaluar la hipótesis propuesta, se ha extraido un conjunto de 5888 genes lncRNAs

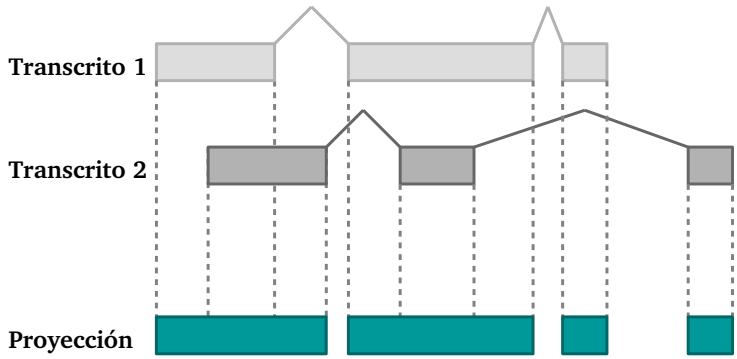


Figura 4.6: *Proyección de transcritos generados por el mismo gen lncRNA. Los intervalos de las coordenadas de los exones de ambos transcritos se proyectan. Los exones que solapan se unen, y los que no solapan son añadidos, creando un transcrito representativo en composición de todas las isoformas del lncRNA. Este proceso se realiza para evitar sesgo hacia genes con un número elevado de isoformas, ya que los exones compartidos generan repetición de secuencia.*

intergénicos (lincRNAs) de GENCODE v19 [320] en base a su ubicación en el genoma y su potencial codificador. La finalidad de este filtrado es evitar lncRNAs que solapan con genes de otro tipo y puedan introducir ruido en nuestra señal. Este conjunto de 5888 genes engloba a su vez un conjunto de 9139 transcritos, ya que una alta proporción de estos genes poseen diversidad de *isoformas*. Para evitar sesgo hacia genes con *splicing* alternativo, se genera un transcrito modelo para cada gen, el cual se construye proyectando sus exones sobre el genoma (ver figura 4.6). Por último, la secuencia de los transcritos proyectados es obtenida del genoma de referencia GRCh37 excluyendo repeats. Los repeats enmascarados se han obtenido también del genoma de referencia, tal y como aparece en Genome Browser. El motivo de la exclusión de los repeats también es el sesgo que estos introducen hacia los k-mers que contienen estos repeats, introduciendo desequilibrio de hebras artificial. Como resultado, se obtuvieron 5888 transcritos, con un total de 6609637 nucleótidos de longitud, excluyendo un 35 % de repeats.

A continuación se contabilizaron todos los k-mers de longitud 6 (hexámeros) sin solapamiento. Las frecuencias de cada K^+ fueron comparadas con su complementario K^-. Para cada uno de estos pares, se realizó el test estadístico para

desequilibrio de hebras basado en el test binomial, obteniendo un total de 2016 p-valores. Nótese que un total de 64 hexámeros simétricos (como por ejemplo AAATTT) fueron excluidos de este test. Cuando se realizan tests estadísticos a un número elevado de casos, es necesario corregir los p-valores obtenidos, ya que por azar se pueden obtener p-valores significativos. Por ejemplo, si nuestro corte de significancia es $\alpha = 0{,}01$ y hacemos 100 tests, podríamos obtener 1 falso positivo. Este fenómeno lo corregimos mediante el método de Benjamini-Hochberg [366] para corrección de tests múltiples. Los p-valores corregidos muestran un desequilibrio muy fuerte para un subconjunto muy amplio de parejas de hexámeros (25 % de los hexámeros sobrerrepresentados obtuvieron un p-valor $< 0{,}01$). La figura 4.7 muestra una adaptación de un gráfico MA[3] para las frecuencias normalizadas de los hexámeros sobrerrepresentados con respecto a las frecuencias normalizadas de sus secuencias complementarios. La idea principal es realizar un gráfico donde se tengan en cuenta la diferencia en proporción de pares de valores y también su abundancia media. El gráfico 4.7 muestra los k-mers sobrerrepresentados, ya que el p-valor para el test binomial de desequilibrio se calcula sólo para esos k-mers. En el eje y, el valor M representa la proporción $f(K^+)/f(K^-)$ para el conjunto de hexámeros sobrerrepresentados en escala logarítmica:

$$M = log_2(f(K^+))$$

A representa una transformación logarítmica de las frecuencias normalizadas:

$$A = \frac{1}{2}(log_2(f(K^+)) + log_2(f(K^-)))$$

.

Los puntos coloreados en rojo representan hexámeros cuyo p-valor es inferior al umbral de 0,01, y las líneas discontinuas representan los percentiles número 99.

Validación por el método de Montecarlo. Este conjunto de secuencias de lincRNAs fue comparado con un conjunto de experimentos aleatorios en

[3] *Mean Average plot*. Los gráficos MA se suelen utilizar para analizar resultados de microarrays de expresión, donde M es la proporción rojo/verde y A es la intensidad media [367].

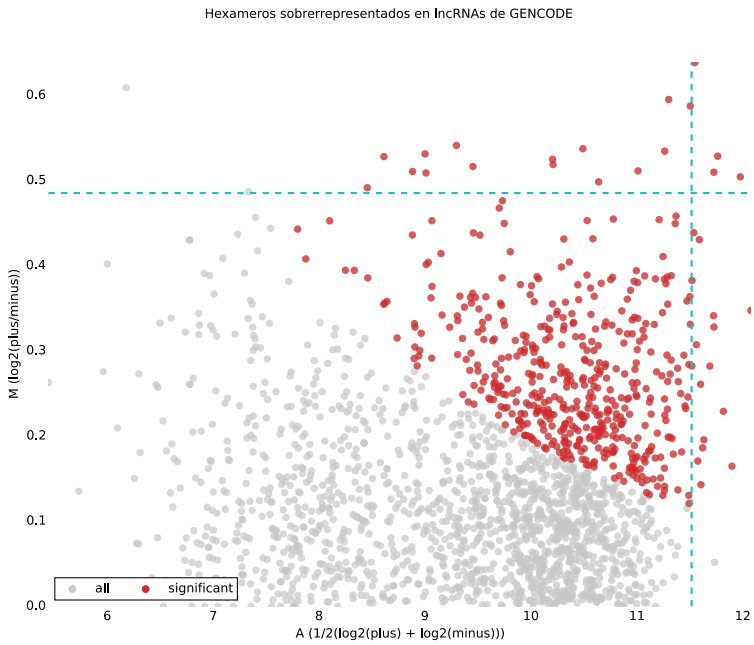


Figura 4.7: *Adaptación de un gráfico MA para las frecuencias normalizadas de los hexámeros sobrerrepresentados con respecto a las frecuencias normalizadas de sus reversos complementarios. M representa la proporción* $f(K^+)/f(K^-)$ *para el conjunto de hexámeros sobrerrepresentados y A representa una transformación logarítmica de las frecuencias normalizadas. Los puntos coloreados en rojo representan hexámeros cuyo p-valor es inferior al umbral de* 0,01*, y las líneas discontinuas representan los percentiles número 99.*

un método de montecarlo con n=1000 iteraciones. Para cada interacción, las coordenadas genómicas del dataset real fueron movidas aleatoriamente a lo largo de todo el genoma (repeat masked) utilizando BedTools [368]. Las secuencias resultantes fueron analizadas del mismo modo que el conjunto de datos real. Cada hexámero fue contado y comparado con su reverso complementario mediante un test binomial, posteriormente corregido el p-valor resultante mediante Benjamini Hochberg. Para estos 1000 tests, en ninguno de los conteos de hexámeros hubo un número mayor de hexámeros significativamente sobrerrepresentados en el conjunto de datos aleatorio comparado con el conjunto de datos real, obteniendo así un p-valor de montecarlo de 0.000999, con una cantidad media de hexámeros sobrerrepresentados por experimento aleatorio < 1 (0.156). Probando de forma individual por pareja de hexámeros, 1444 resultados muestran mejor p-value (i.e. más bajo) en el conjunto de datos real para > 997 iteraciones, obteniendo un p-valor de montecarlo (corregido para multiple tests) $< 0{,}01$.

Esta randomización muestra una señal de desequilibrio de hebras muy fuerte en el dataset real con respecto al aleatorio. Sin embargo, utilizar únicamente este p-valor para clasificar los hexámeros, nos daría un conjunto muy elevado de motivos de interés, por lo que obtendríamos con toda seguridad gran cantidad de falsos positivos. En las secciones subsecuentes utilizaremos otras medidas de desequilibrio de hebras junto con un conjunto de hexámeros específicos de ARN obtenidos de bases de datos existentes y de la literatura.

Conjunto de hexámeros reales. En este apartado se describe en detalle el conjunto de hexámeros reales extraidos de la literatura. Estas secuencias se han utilizado para evaluar la capacidad predictiva de las medidas de desequilibrio de hebras propuestas, así como para estimar un umbral de corte adecuado para seleccionar un conjunto de hexámeros funcionales candidatos.

Se ha obtenido un conjunto de 483 hexámeros pertenecientes a cuatro categorías:

- **Señales de poli-adenilación**. La poliadenilación consiste en la adición de una cola de adeninas en la sección 3' del transcrito, importante para la exportación nuclear, la traducción y la estabilidad del ARN [369].

Es también frecuente en ARNs no codificantes de organismos eucariotas [370]. Este proceso necesita la existencia en los transcritos de una señal de poliadenilación, la cual generalmente se encuentra también cerca del extremo 3' del transcrito. En este conjunto de hexámeros se encuentran tanto la señal AAUAAA como sus variantes obtenidas de Beaudoing *et al.* [371].

- **Exonic Splicing Enhancers**. Motivos de 6 bases encargados de dirigir el proceso de splicing. El conjunto de motivos potenciadores de splicing frecuentes en ARN han sido obtenidos de Fairbrother *et al.* [372].
- **Exonic Splicing Silencers**. Los motivos silenciadores de splicing están relacionados con el splicing alternativo de los transcritos. Sus motivos fueron obtenidos de Wang *et al.* [373].
- **Targets de microRNAs**. Las secuencias de los microRNAs se obtuvieron de miRbase [98]. Entre ellos, se seleccionaron el 5 % de mayor nivel de expresión en cualquiera de los tejidos en miRmine[4]. A continuación, se extrajeron las secuencias complementarias a la región seed de los microRNAs como se define en [374] como miRNA matches: la región complementaria reversa a los nucleótidos 2 a 7 (inclusive) del extremo 5' del miRNA. Dado que estas secuencias tienen longitud 7, 2 hexámeros fueron obtenidos de cada una de las secuencias.

La tabla 4.4 muestra la composición de hexámeros del conjunto de prueba. La figura 4.8 muestra la disposición del conjunto de hexámeros conocidos dentro del gráfico MA generado anteriormente.

Capacidad predictiva de las medidas propuestas. El conjunto de hexámeros con funcionalidad propia de ARN manualmente recolectado de la literatura y de bases de datos específicas fue a continuación utilizado como conjunto de elementos positivos para evaluar la capacidad predictiva de los scores formulados en el primer apartado, sc_{fold}, sc_{SI}, así como los p-valores calculados en el conjunto de datos reales para comprobar el desequilibrio de hebras.

[4] http://guanlab.ccmb.med.umich.edu/mirmine

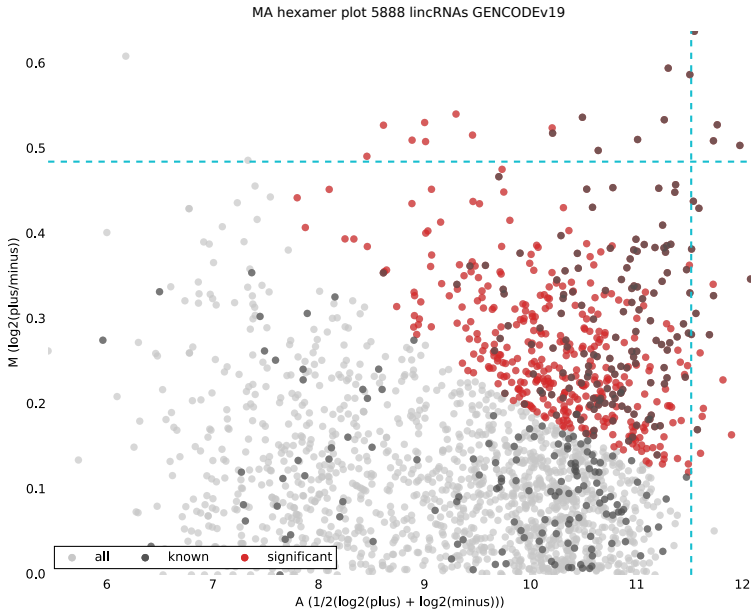


Figura 4.8: *Gráfico MA donde se muestran los hexámeros con función específica de ARN conocida. Los puntos de color gris oscuro muestran los hexámeros conocidos en el conjunto de hexámeros sobrerrepresentados.*

Tabla 4.4: *Fuente y número de hexámeros únicos obtenidos para cada uno de los tipos de elementos específicos de ARN utilizados.*

Fuente	**Función**	**Número**
Fairbrother [372]	Exonic Splicing Enhancers	238
Wang [373]	Exonic Splicing Silencers	73
miRbase [98] + miRmine	miRNA seed matches	169
Beaudoing [371]	Señal de poliadenilación	10
	Total (únicos)	483

Utilizando sc_{fold} y sc_{SI} para diferentes valores de α y β y el conjunto de hexámeros conocidos, podemos aproximarnos a este problema como a un caso de clasificación binaria, donde cada hexámero puede ser considerado como *positivo* o *negativo*. El conjunto de hexámeros *positivos* son aquellos que tienen funcionalidad específica de ARN. Las dos medidas de scoring propuestas ofrecen como resultado una lista ordenada de hexámeros en orden de interés. Dada esta lista ordenada y el conjunto conocido de positivos, es posible trazar curvas ROC para ambos scores. Las curvas ROC (del inglés Receiver Operator Characteristic) se utilizan para estimar la efectividad de los clasificadores. Básicamente, se generan puntos de la curva ROC moviendo un umbral de corte para seleccionar los N primeros items en el ranking. Para cada umbral de corte se calculan la tasa de verdaderos positivos y la tasa de falsos positivos (FDR, FPR), y dichos valores conforman los valores x e y del gráfico. Como medida general de calidad del clasificador, se suele utilizar el valor AUC (del inglés Area Under de Curve), que representa el área bajo la curva ROC. En el caso del clasificador perfecto, el área bajo la curva es igual a 1, y el clasificador aleatorio da un área bajo la curva de 0.5. Es interesante notar que el peor resultado posible es 0.5, ya que para valores inferiores a 0.5, se podría construir el clasificador inverso (aquel que da valor negativo a los casos positivos y viceversa) y el valor de AUC_{nuevo} sería $1-AUC$.

En la figura 4.9 se muestra la curva ROC para sc_{fold} y tres variantes de sc_{SI}, donde los valores (α, β) son (0.50, 0.50), (0.75, 0.25) y (1.00, 0.00), respectivamente. Nótese que cuando $\beta = 1$ y $\alpha = 0$, $sc_{SI} \equiv sc_{fold}$. Además, se puede ver cómo el mejor resultado se obtiene para $\alpha = \beta = 0{,}5$. Tiene sentido pensar que cuando el desequilibrio es alto y además lo es la frecuencia de los hexámeros, se espera más funcionalidad específica de ARN que cuando estos simplemente son abundantes o cuando hay una gran diferencia pero la frecuencia de aparición de K^+ y K^- es demasiado baja.

Comparación con intrones. La metodología propuesta se ha aplicado a los intrones de los mismos 5888 genes lncRNAs para comprobar la diferencia en el potencial predictivo del conjunto de hexámeros seleccionados por sc_{fold} y sc_{SI}. Los intrones se han calculado con respecto a los transcritos proyectados, es decir, sólo

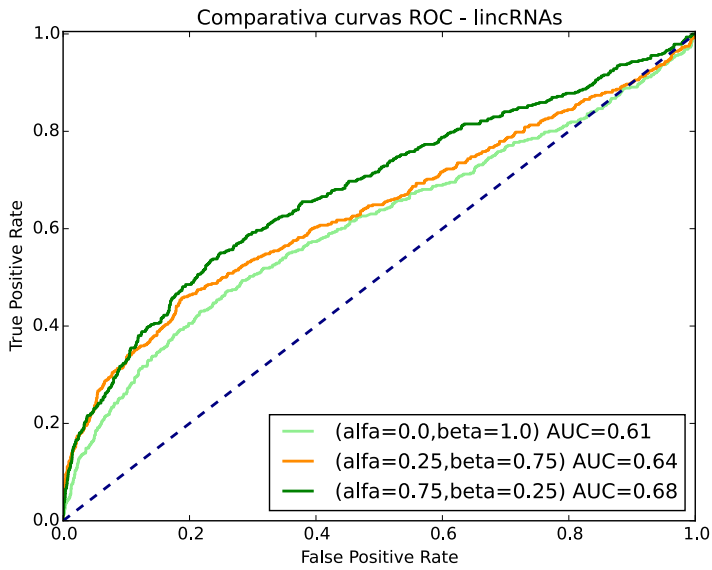


Figura 4.9: *Gráfico ROC para las funciones de scoring sc_{SI} para distintos valores de α y β. La línea verde claro corresponde a sc_{fold}, donde $\alpha = 0$ y $\beta = 1$. Se puede observar que sc_{SI} mejora la predicción de kmers conocidos.*

se ha incluido secuencia que no está en *ningún intrón* de los lncRNAs pertenecientes al conjunto de datos. El conjunto de secuencias es mucho mayor, ya que el tamaño de los intrones es mayor que el de los exones en nuestro conjunto de datos. La tabla 4.5 muestra las proporciones de nucleótidos, repeats y nucleótidos.

La figura 4.10 muestra la curva ROC para el mismo conjunto de hexámeros propios de RNA conocidos y los hexámeros ordenados por sc_{SI} y sc_{fold}. Se puede observar que el valor de AUC es mucho más bajo en el caso de la medida sc_{fold}, mientras que sc_{SI} parece discriminar peor entre exones e intrones cuando se tiene más en cuenta la abundancia de hexámeros únicamente. (A y B, respectivamente). Se confirma así también que los hexámeros seleccionados en lincRNAs, pese a abundar también en las secuencias intrónicas, en estas el desequilibrio de hebras no es informativo para propiedades específicas de ARN.

Tabla 4.5: *En la siguiente tabla se muestran los tamaños de los conjuntos de datos de intrones-exones para el set de 5888 genes lincRNA seleccionado. En ambos casos se elige una proyección, es decir, cada gen lincRNA es representado por una proyección de todos sus exones, considerándose intrones las secuencias complementarias. Se consideran pues como intrones las secuencias que no están en* ***ningún*** *exón del conjunto de prueba. Se observa que en la proporcionalidad de bases hay cierta similitud, aunque en el caso de los intrones la diferencia AT-GC es más acentuada que en exones.*

	Exones	**Intrones**
Núm. secuencias	19218	13749
Longitud total	4236796	80497221
Longitud media ($\pm\sigma$)	343.929 (± 649.297)	12169.966 (± 27527.943)
% repeats	35.89 %	51.89
% A (excl. repeats)	28.19 %	30.56
% C (excl. repeats)	22.39 %	18.78
% G (excl. repeats)	22.78 %	19.25
% T (excl. repeats)	26.63 %	31.41

4.4.3. Pipeline de extracción de motivos específicos de ARNs largos no codificantes

La metodología de análisis de motivos mediante desequilibrio de hebras propuesta en este apartado se ha automatizado en un pipeline para extracción de motivos específicos de ARN. La figura 4.11 ilustra todo el proceso que se realiza a partir de las anotaciones de lncRNAs de GENCODE. A continuación se describen en más detalle los pasos automatizados del pipeline MoSI (Motif Strand Imbalance).

1. **Preprocesado.**

   En primer lugar, se realiza un preprocesado de los datos que incluye extraer un subconjunto de lncRNAs candidatos intergénicos, es decir, aquellos que no solapan con otros genes u otras entidades transcritas, para evitar interferencia con motivos específicos de ARN codificante. Estas anotaciones de transcritos de lncRNAs son proyectadas en un único transcrito para evitar que haya sesgo hacia genes lncRNA con splicing alternativo. De esta manera cada exón está representado una única vez. Además, las secuencias

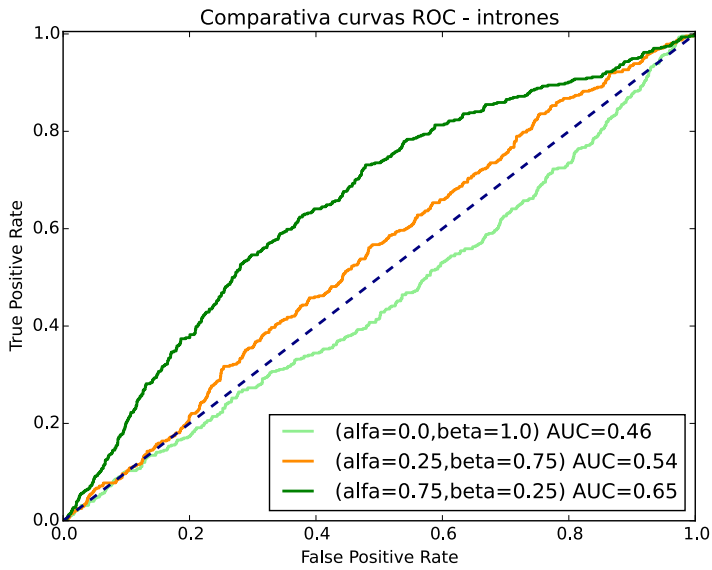


Figura 4.10: *Gráfico ROC para las funciones de scoring sc_{fold} y sc_{SI} en el conjunto de intrones. En este caso, sc_{SI} también predice mejor los hexámeros conocidos par altos valores de α.*

repetitivas son enmascaradas para evitar sesgo hacia la composición de los repeats. Para ello se extraen las coordinadas de las anotaciones proyectadas de GENCODE del genoma de referencia correspondiente ya enmascarado por Repeat Masker obtenible desde el Genome Browser.

2. **Conteo y emparejado de k-mers.**

   En base a un parámetro k, se contabilizan todos los k-mers en las secuencias de entrada. Además, existe un parámetro de solapamiento que permite excluir k-mers iguales y solapantes, con la finalidad de evitar priorizar k-mers cíclicos, como por ejemplo `GCGCGCGC` incluiría dos veces el hexámero `GCGCGC` si se permite solapamiento, sólo una sería contada en caso contrario.

3. **Scoring de desequilibrio de hebras.**

   A partir de los conteos realizados en el paso anterior se calculan los índices

de sobrerrepresentación y desequilibrio de hebras sc_{fold} y sc_{SI} descritos en secciones anteriores. Estos valores permiten obtener una lista ordenada de k-mers.

4. **Selección de k-mers por score.**

   Se selecciona el conjunto de k-mers más sobrerrepresentados a partir de un umbral de corte. Este umbral de corte, así como el scoring más adecuado para el conjunto de k-mers, es seleccionado en base a los resultados de predicción con respecto al conjunto de datos de validación recogido de la literatura biomédica y las bases de datos miRmine y miRbase.

5. **Anotación centrada en transcrito y a nivel de genoma.**

   Los k-mers seleccionados sirven de entrada a un proceso de anotación que escanea todos los transcritos de GENCODE para generar dos ficheros de anotación, uno a genoma completo y otro relativo al transcrito. Los k-mers que aparecen en juntas de splicing son también anotados debidamente divididos en el fichero de genoma completo. Estos ficheros son proporcionados con la finalidad de que análisis posteriores arrojen luz sobre la posible funcionalidad específica de estos k-mers seleccionados que se presumen candidatos a tener funcionalidad específica de ARN, y más en particular, de lncRNA.

El pipeline MoSI propuesto es modular y susceptible de ampliarse con posteriores análisis como conservación y clustering de los motivos obtenidos como candidatos. Adicionalmente, está completamente automatizado y parametrizado, de forma que numerosos estudios extra se pueden realizar variando dichos parámetros, desde el tamaño de k-mer al conjunto de lncRNAs de entrada, para el estudio por ejemplo de conjuntos de lncRNA coexpresados, o aquellos conjuntos de lncRNAs que guarden algún tipo de relación conocida. Por otro lado, versiones más recientes de GENCODE se analizarán para proporcionar nuevos resultados. La reproducibilidad del análisis realizado está pues garantizada, tanto para actualizar los resultados en caso de actualización de las fuentes de datos de entrada, como para explorar conjuntos de parámetros no explorados previamente.

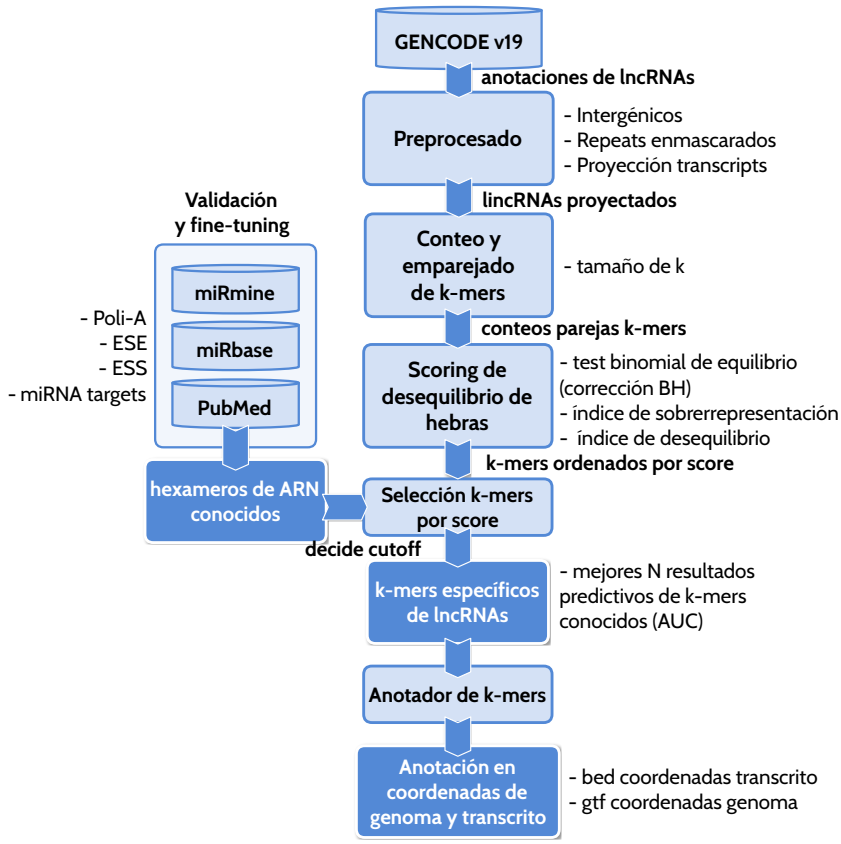


Figura 4.11: *Pipeline de análisis de k-mers MoSI. Primero se hace un preprocesado de los transcritos de GENCODE: filtrado en base a su ubicación intergénica y potencial codificador, se enmascaran los repeats y se obtiene un transcrito representante de cada lncRNA mediante proyección de exones. Estas secuencias son procesadas para contabilizar los k-mers en base a un parámetro k. A continuación, los k-mers son puntuados utilizando los scores sc_{fold} y sc_{SI}. Las listas ordenadas de k-mers resultantes son validadas mediante el conjunto de hexámeros conocidos construido a partir de la literatura biomédica y de las bases de datos miRmine y miRbase. Por último, se elige un umbral de corte y se anotan los k-mers seleccionados. La anotación se realiza a nivel de genoma y a nivel de transcrito, incluyendo en ella todos los transcritos de GENCODE. En azul oscuro están representadas las salidas del pipeline que son resultados de salida del pipeline MoSI. Todo este proceso, desde selección hasta anotación, está automatizado y parametrizado, pudiéndose volver a ejecutar para otras configuraciones (distintos tamaños de k) y distintas versiones de anotaciones y genomas de referencia.*

4.5. Conclusiones

El campo de la biomedicina es una área de conocimiento en continua evolución. Como se ha podido ver a lo largo de este capítulo, los ARNs largos no codificantes son una entidad que ha adquirido gran protagonismo en los últimos años. Este protagonismo incipiente ha generado cantidad de herramientas, enfoques y bases de datos que se han ido paulatinamente poniendo a disposición de la comunidad científica. En este entorno de continuo crecimiento e innovación hemos desarrollado dos propuestas, las cuales a su vez han ido evolucionando a la par que el conocimiento existente en este campo.

En este capítulo se han propuesto dos metodologías de estudio de ARNs largos no codificantes, teniendo en cuenta dos características propias del ARN: su estructura secundaria y su direccionalidad de transcripción.

La primera propuesta, RNAIntuit, aborda el estudio computacional de posibles hipótesis de interacción de ARNs largos no codificantes con otras entidades biológicas. Estas interacciones pueden ser tanto directas como indirectas, con elementos como ADN, factores de transcripción o proteínas RBP. El estudio de estos mecanismos de interacción deja clara la complejidad de los procesos de regulación en los que se involucran los lncRNAs, haciendo necesarias medidas integradoras que tengan en cuenta numerosas entidades biológicas. Sólo así será posible entender los mecanismos concretos por los cuales un lncRNA puede regular un gen.

El estudio de estas interacciones se ha realizado teniendo en cuenta la estructura secundaria de plegado del ARN, la cual condiciona la interacción de los transcritos de lncRNAs con otras entidades. En este sentido, se aúna una predicción computacional de motivos ya validada con anterioridad en otras aplicaciones con una capa adicional de información de localización.

Los resultados obtenidos por esta metodología se muestran coherentes con el conocimiento existente sobre las entidades involucradas al momento de desarrollo del trabajo. Sin embargo, la novedad del interés existente en los ARNs largos no codificantes es la causa de que no sea este conocimiento muy extenso, al igual que

los datos de entrada utilizados no eran exhaustivos. El hecho de que a pesar de ambas circunstancias los resultados muestren consistencia resulta prometedor. No obstante, se requiere trabajo futuro para introducir nuevas fuentes de información que de seguro surgirán a corto plazo.

Adicionalmente, la validación experimental de las hipótesis resultantes establecidas a partir del estudio computacional del lncRNA Fendrr están en proceso, siendo en sí mismas un ejemplo de la utilidad de las herramientas computacionales predictivas en el ámbito de la biología molecular. La reducción de espacios de búsqueda que pueden resultar combinatorialmente intratables a un conjunto asequible de candidatos ahorra costes económicos y temporales permitiendo a los laboratorios centrarse en las entidades de verdadero interés. Asimismo, la colaboración interdisciplinar entre expertos en computación, biología y medicina se hace inevitable, siendo este estudio una prueba palpable de esto.

La segunda mitad de este capítulo describe MoSI, una metodología destinada a ampliar el conjunto de anotaciones sobre ARNs largos no codificantes concretos, el cual ahora mismo consta principal y casi exclusivamente de la ubicación en el genoma de estas entidades. Mientras en el estudio inicial nos centrábamos en lncRNAs conocidos, de relación demostrada con enfermedades, en la segunda propuesta nos centramos en la búsqueda de motivos específicos de ARNs largos no codificantes de forma exploratoria, proporcionando como resultado final un conjunto de anotaciones, a partir del cual se espera que puedan surgir múltiples estudios de mayor especificidad.

Con la finalidad de ser capaces de distinguir entre motivos específicos de ADN y motivos específicos de lncRNAs, se ha utilizado una medida de desequilibrio de hebras basada en la dirección de la transcripción, una propiedad propia del ARN. Formulando esta característica en términos matemáticos y aplicando metodologías de Inteligencia Artificial hemos sido capaces de demostrar que esta señal caracteriza a un conjunto amplio y representativo de lncRNAs como son los lncRNAs intergénicos. Además, esta medida es capaz de predecir un conjunto de motivos de funcionalidad específica de ARN extraido de la literatura y de bases de datos relacionadas con la funcionalidad del ARN. Esta capacidad predictiva

de motivos conocidos indica que el conjunto de motivos seleccionados como de interés podría confirmar en un futuro nuevas funcionalidades. Es por ello que estos motivos seleccionados han sido anotados en todos los lncRNAs y estas anotaciones, junto con la metodología proporcionada, son contribuciones del trabajo presentado en esta tesis.

Por último, la reproducibilidad de los métodos utilizados es imprescindible, incluso cuando se trata de metodologías exploratorias en un conjunto amplio de datos. Es necesario poder repetir el estudio para nuevas versiones de la información de entrada, y con este fin se detalla un pipeline de análisis de datos, a través del cual seremos capaces de crear nuevas versiones de las anotaciones realizadas para nuevas versiones de GENCODE o del genoma de referencia.

Parte III

Conclusiones

Capítulo

5

Conclusiones y trabajo futuro

5.1. Conclusiones

El presente trabajo realiza varias aportaciones al estudio de elementos biológicos presentes en los procesos de regulación de los genes, haciendo especial hincapié en la integración de fuentes de datos heterogéneas y en el análisis de secuencias. En particular, se han propuesto metodologías computacionales para abordar tres problemas: i) priorización de entidades biológicas en redes heterogéneas, y su aplicación en las áreas de predicción de SNPs reguladores en enfermedades y reposicionamiento de medicamentos; ii) análisis de secuencias de ADN en búsqueda de elementos y conjuntos de elementos reguladores; y iii) análisis de ARNs largos no codificantes, entidades de creciente interés científico y gran potencial regulador.

En el marco de la Medicina Personalizada, estas aportaciones son de interés, ya que la Medicina Personalizada sólo se puede abordar desde un mejor entendimiento de los mecanismos de regulación de los genes. Las metodologías propuestas en esta tesis incluyen distintos aspectos clave para progresar hacia la Medicina Personalizada: i) la capacidad de integrar numerosas y continuamente crecientes cantidades de información, ii) la capacidad de representar y analizar

sistemas complejos en los que interactúan distintos tipos de entidades biológicas iii) la capacidad de análisis de variaciones individuales y de elementos reguladores en el genoma para identificar su papel en la regulación genética y su potencial influencia en el fenotipo del individuo.

Priorización en redes biológicas heterogéneas

La primera línea de trabajo descrita en esta memoria abarca la inferencia de conocimiento en redes biológicas heterogéneas construidas a partir de múltiples fuentes de información. Este ámbito de aplicación tiene gran relevancia en un campo tan dinámico como la biología, donde la constante aparición de nuevas técnicas experimentales y la masiva producción de datos biológicos a partir de las mismas requieren de metodologías escalables y robustas que permitan representar el conocimiento existente e inferir nuevas hipótesis de interés.

En esta línea se ha propuesto **ProphTools** [182], una metodología general de análisis de redes heterogéneas, que se ha aplicado a dos ámbitos de interés: i) la búsqueda de elementos reguladores relacionados con enfermedades, mediante **DiGSNP** [89], una metodología que combina la priorización gen-enfermedad con el análisis de potencial regulador de mutaciones, y ii) el reposicionamiento de medicamentos, mediante **DrugNet** [90], una metodología de aplicación de inferencia en redes heterogéneas donde se incluyen enfermedades, fármacos y proteínas (dianas terapéuticas).

ProphTools [182] es una metodología general y flexible de priorización en redes heterogéneas, compuestas por varias subredes que representan diferentes dominios biológicos de interés. Esta metodología está basada en dos mecanismos de propagación: una propagación intra-red que implementa una versión del paradigma Random Walk with Restart similar al algoritmo de Flow Propagation [202] y una propagación ponderada entre redes de diferente dominio.

Para permitir su aplicación a cualquier ámbito, se ha diseñado un meta modelo de descripción de redes heterogéneas de un número arbitrario de subredes. Se han incluido en la metodología paradigmas de validación como el test Leave-One-Out (LOO), y el test LOO con validación cruzada para poder evaluar la efectividad de

la priorización y elegir la configuración de subredes heterogéneas que proporcione los mejores resultados con ProphTools.

Tanto la propuesta de priorización de ProphTools como su implementación son altamente modulares, lo que se ha podido comprobar con la inclusión de RANKS [102], una nueva metodología para priorización basada en funciones kernel, como alternativa de propagación *intra-red* al método RWR propuesto. Además, se han incorporado al desarrollo de la herramienta técnicas de desarrollo orientado a tests y de integración continua para garantizar la robustez del método, que adicionalmente ha sido publicado en forma de paquete descargable desde un repositorio público (disponible en GitHub: `http://www.github.com/cnluzon/prophtools`) bajo la licencia de código abierto GPLv3.

Como aplicación de la metodología de ProphTools a un problema concreto, se propone **DiGSNP** [89], un método de priorización jerárquico en dos niveles (enfermedad-gen, gen-SNP) que constituye un caso de prueba tanto para la metodología de priorización en redes como para IntuitSNP, la metodología propuesta en el capítulo de análisis de secuencias reguladoras. En el nivel más general de priorización, DiGSNP obtiene mediante priorización en redes heterogéneas enfermedad-proteína-gen, una lista priorizada de genes candidatos a estar relacionados con una enfermedad, que posteriormente son analizados a nivel de SNP para inferir posibles hipótesis sobre los mecanismos reguladores que guían dichas relaciones.

La metodología es validada mediante casos de estudio en la literatura, que estiman prometedores los resultados, aunque no muy numerosos, debido a la dificultad para obtener datos validados en ambos niveles: tanto de SNPs reguladores como de su relación con enfermedades.

En segundo lugar, la metodología de priorización general propuesta también se aplica a un campo de gran interés en el marco de la Medicina Personalizada: el reposicionamiento de medicamentos.

Para este fin se propone **DrugNet** [90], una aplicación de la metodología de ProphTools para el reposicionamiento de medicamentos que permite sugerir nuevos fármacos para tratar enfermedades y nuevos usos para fármacos existentes.

La búsqueda de aplicaciones nuevas para fármacos ya comercializados puede suponer un ahorro en costes económicos y temporales en el desarrollo de nuevos medicamentos, así como ampliar las posibilidades de tratamiento para un mismo individuo.

DrugNet está disponible públicamente a través de un servidor web en http://genome.ugr.es:9000/drugnet. Con respecto al estudio de la efectividad de **DrugNet**, la conexión fármaco-enfermedad ha sido validada mediante los métodos de LOO y validación cruzada mencionados. Estos tests han demostrado la importancia de integrar múltiples fuentes de información, ya que la configuración de redes más efectiva involucraba tres dominios diferentes: fármacos, proteínas y enfermedades. El método ha sido validado adicionalmente mediante el uso de datos reales de ensayos clínicos, mostrando que DrugNet predice con consistencia un porcentaje mucho mayor de medicamentos en fases tardías de desarrollo. Por último, se ha comparado DrugNet con PREDICT, una conocida herramienta de reposicionamiento de medicamentos, obteniendo resultados prometedores.

Detección de elementos reguladores en secuencias de ADN

La segunda línea de trabajo propuesta en esta memoria abarca el análisis de secuencias de ADN en la búsqueda de elementos o conjuntos de elementos reguladores de los genes. Para ello se han tenido en cuenta, por un lado, las características de vaguedad e incertidumbre presentes en la búsqueda de motivos en secuencias de ADN, y, por otro lado, características biológicas conocidas de los procesos de regulación de los genes, explorando la influencia de SNPs, las mutaciones más frecuentes y estudiadas del genoma, en posibles lugares de unión de factores de transcripción.

Las metodologías propuestas se basan en modelos de Inteligencia Artificial de uso extendido, como la lógica difusa o la extracción de itemsets frecuentes de una base de datos transaccional, cuyos parámetros han sido refinados gracias a varios casos de estudio con conjuntos de datos reales.

Se han descrito dos metodologías: **IntuitSNP**, para el análisis del potencial regulador de SNPs o mutaciones de nucleótido simple, y **CisMiner** [258], una

herramienta de búsqueda de módulos reguladores en *cis* en las regiones no codificantes de un genoma completo.

La primera de las herramientas propuestas, **IntuitSNP**, es una metodología para el análisis de la influencia reguladora de un SNP que puede estar afectando la afinidad de unión de un factor de transcripción. Esta metodología, basada en una medida difusa de similitud secuencia-motivo, permite calcular un nivel de perturbación en los niveles de afinidad de la secuencia entre los diferentes alelos del SNP ubicado en ella, pudiendo estimar así el potencial regulador de dicha mutación.

Los resultados fueron validados utilizando la base de datos ORegAnno [151], que almacena información sobre variaciones reguladoras, así como mediante una búsqueda manual en la literatura. Además, la metodología IntuitSNP está incluida en una implementación en CUDA que permite realizar este estudio sobre gran cantidad de SNPs en un breve periodo de tiempo.

La segunda de las propuestas en esta línea de trabajo, **CisMiner** [258], analiza todo un genoma no codificante buscando *clusters* frecuentes de elementos reguladores. Esta tarea se realiza mediante un *clustering* difuso seguido de un algoritmo difuso de minería de itemsets frecuentes, que opera sobre una base de datos transaccional difusa.

CisMiner es una metodología única dentro de este ámbito en su concepción a genoma completo, ya que las herramientas propuestas siempre requieren conocimiento *a priori* sobre las características de los módulos reguladres objetivo. Otras metodologías exigen al usuario información sobre los módulos a buscar, su tamaño o elementos involucrados, genes coexpresados o algún tipo de información que permita reducir considerablemente el espacio de búsqueda, debido a la naturaleza combinatoria del problema a resolver.

En cuanto al alcance de CisMiner, la herramienta propuesta ha sido aplicada a los genomas no codificantes completos de dos organismos: *Saccharomyces cerevisiae* y *Drosophila melanogaster*, demostrando que la metodología es capaz de escalar a genomas relativamente grandes. Los resultados se han validado con datos reales extraídos de la literatura y de una base de datos de interacción proteína-proteína basada en múltiples fuentes de información. La metodología propuesta ha

sido aplicada, además, tanto a un conjunto de TFBSs validados experimentalmente obtenidos de Harbison *et al.* [290], obteniendo resultados de alta calidad, como a un conjunto de TFBSs predichos computacionalmente, demostrando la capacidad de la herramienta para manipular información donde pueda haber elementos espúreos. Por otro lado, aunque la metodología es robusta a este tipo de resultados, se ha podido concluir la influencia de la predicción de los TFBSs candidatos en la calidad de los resultados finales.

La aplicación a genomas no codificantes de gran tamaño, tales como el humano, supone todavía un reto de optimización que se pretende abordar en líneas de trabajo futuras. Adicionalmente, los resultados obtenidos con CisMiner han sido puestos a disposición pública, así como la herramienta desarrollada, en forma de web interactiva accesible en `http://genome.ugr.es:9000/cisminer/`.

ARNs largos no codificantes

Por último, se ha propuesto una línea de trabajo centrada en el análisis de una entidad biológica de gran potencial regulador: los ARN largos no codificantes o lncRNAs. Las características de esta entidad biológica son en la actualidad una incógnita que poco a poco va despejándose mediante esfuerzos conjuntos de anotación similares a los que ya se realizaran en el pasado con los genes codificantes. En este sentido, se han propuesto varias aproximaciones que pretenden explotar el conocimiento existente sobre las secuencias de ARN al estudio de estas nuevas entidades.

Se proponen en este apartado dos metodologías: **RNAIntuit** [316], una herramienta que busca lugares de unión de proteínas en secuencias de ARN largos no codificantes teniendo en cuenta la estructura secundaria de plegado del ARN; y **MoSI** (Motif Strand Imbalance), un pipeline de búsqueda de motivos específicos de ARNs largos no codificantes en un conjunto de transcritos de lncRNAs. MoSI emplea una propiedad específica de ARN para discriminar entre motivos de lncRNA y motivos de ADN: la direccionalidad de la transcripción.

La primera propuesta, **RNAIntuit** [316], tiene como meta poder realizar un estudio computacional de los posibles mecanismos de regulación de los lncRNAs

mediante su interacción con proteínas, a través de la cual se realizan diversas funciones reguladoras, como guía de factores de transcripción a sus lugares de unión, silenciador o andamiaje para estructuras moleculares más complejas. Esta meta se alcanza mediante la combinación de una predicción computacional de la estructura secundaria de los transcritos de ARN y una búsqueda de motivos de afinidad de unión de proteínas, las cuales son filtradas por zonas según la estructura del ARN. Este enfoque es validado después mediante el estudio de un lncRNA conocido por su relación con la carcinogénesis, Fendrr, y ha permitido generar una hipótesis sobre los mecanismos por los que Fendrr regula a un gen cercano, FoxF1. Estas hipótesis son prometedoras, ya que se alinean con el conocimiento existente sobre la regulación de FoxF1. No obstante, el conocimiento disponible sobre los mecanismos reguladores de los lncRNAs es todavía escaso, dado lo relativamente reciente de la investigación en este campo, y esperamos que en el futuro la validación y mejora de este enfoque mediante datos experimentales sea posible.

Por último, se realiza un análisis a nivel general de motivos específicos de ARN en lncRNAs, buscando motivos específicos en la función de los lncRNAs mediante **MoSI** (Motif Strand Imbalance pipeline). La metodología propuesta explota una característica propia del ARN para diferenciar motivos abundantes en las secuencias de análisis que provienen del ADN desde el que se transcribe de motivos específicos de ARN. Para la validación de esta metodología se ha recopilado un conjunto de motivos conocidos por su funcionalidad en la maquinaria del ARN. Este conjunto de motivos de prueba se ha estimado enriquecido en los resultados y se ha utilizado para definir un umbral de corte y seleccionar un conjunto de motivos candidatos a ser específicos de lncRNAs.

Adicionalmente, este conjunto de motivos se ha anotado en los transcritos de lncRNAs a nivel de transcrito y a nivel de genoma completo, y estas anotaciones son, junto con la metodología de extracción, los frutos de esta última línea de trabajo. Se espera que esta metodología pueda ser aplicada a versiones futuras de las bases de datos de anotaciones de lncRNAs, permitiendo mejorar y refinar los resultados. Las anotaciones de motivos putativos específicos de lncRNAs son

asimismo un recurso para futuros análisis que permitan arrojar luz sobre los mecanismos que rigen el funcionamiento de esta entidad de nuevo interés.

5.2. Trabajo Futuro

Las diversas propuestas llevadas a cabo en la presente memoria abren nuevos caminos que puedan permitir la mejora de las mismas en el futuro. A continuación se describen las líneas de investigación más prometedoras.

Priorización en redes biológicas heterogéneas

La integración de fuentes de datos biológicos heterogéneos también abre caminos para la mejora, entre los que se incluyen:

- **Aplicación de ProphTools a nuevos ámbitos de interés**. La metodología propuesta es independiente tanto del número de tipos de entidades biológicas utilizadas (esto es, el número de subredes entre las que priorizar) como de las fuentes de información utilizadas, por lo que se propone explotar esta generalidad para aplicar la herramienta a nuevos ámbitos, como puede ser la priorización de lncRNAs con respecto a enfermedades.
- **Integración de nuevas fuentes de información en DrugNet**. Como se ha mencionado, la metodología propuesta en DrugNet permite ampliar el número de fuentes de información. En este sentido, sería de interés añadir nuevas fuentes ya sea de forma implícita, como se mencionó en el caso de los efectos secundarios, como de forma explícita, buscando nuevas entidades clave que se relacionen con las ya existentes en DrugNet (proteínas, enfermedades, fármacos).
- **Mejora de la eficiencia de la priorización**. La priorización de Random Walk with Restart tiene ciertas limitaciones computacionales, que crecen conforme crece el tamaño de las redes, ya que requiere el cálculo de una inversión matricial. Este proceso se ha acelerado mediante el precálculo de dicha matriz para cada una de las subredes y su almacenamiento en el meta modelo de configuración de redes heterogéneas. Sin embargo, esto supone un coste de espacio elevado, que se podría reducir mediante técnicas más avanzadas de precálculo como la de Tong *et al.* [375]. También sería posible abordar una paralelización en GPU del algoritmo mediante CUDA.

- **Mejora de la eficiencia de los tests LOO**. Pese a que la priorización en la metodología propuesta es rápida para un conjunto de entrada, los métodos integrados en la herramienta para la validación de las configuraciones propuestas toman más tiempo debido a que es necesario probar la priorización eliminando uno por uno los arcos que unen ambas redes. Se propone pues mejorar este apartado mediante una paralelización de los tests.
- **Inclusión de otras modalidades de propagación *intra-red* y entre redes**. La naturaleza modular de ProphTools permite la adición de nuevas estrategias de propagación en ambas etapas, como se ha demostrado mediante la inclusión de RANKS [102] como metodología de propagación *intra-red*. También se podrían incluir otras modalidades de propagación entre redes como alternativas de la propagación ponderada actual.
- **Adaptación a ontologías de las modalidades de propagación *intra-red***. Las características propias de las ontologías podrían utilizarse a la hora de propagar por ellas la información.

Detección de elementos reguladores en secuencias biológicas

Con respecto a las propuestas sobre la búsqueda de elementos reguladores, se proponen las siguientes mejoras posibles:

- **IntuitSNP**. Entre las posibles mejoras al funcionamiento de IntuitSNP está la mejora de la medida de similitud SC_{intuit} utilizada para estimar afinidad de unión mediante el factor λ de ajuste de PWMs propuesto. Además, se ha observado que una acumulación de modificaciones en posiciones muy poco conservadas del motivo puede penalizar en exceso el resultado, por lo que serían de interés estudios en ese sentido.

  Por otra parte, la inclusión de información más detallada sobre los alelos del SNP, como puede ser la abundancia de cada alelo en relación con el nivel de incidencia de la enfermedad de estudio. Por otro lado, la medida de similitud difusa secuencia motivo utilizada se podría mejorar integrando conocimiento adicional sobre la secuencia donde se ubica el SNP, como

podría ser información epigenética tanto de accesibilidad de la cromatina, la cual se ha demostrado que tiene influencia en la unión o no de factores de transcripción a una secuencia [376] como de metilación, información que también está muy relacionada con el silenciado de los genes [377] y la cual empieza a estar más disponible a nivel de bases [378].

- **CisMiner**. La metodología propuesta para la búsqueda de módulos reguladores en *cis* admite varias líneas de mejora. La primera consiste en optimizar la minería de *itemsets* frecuentes difusa de forma escalable al tamaño del genoma humano, el cual supone actualmente un reto. Estas mejoras podrían ser algorítmicas, adaptando el paradigma a otros modelos, o de implementación, paralelizando la misma mediante CUDA. La paralelización en GPU ha mostrado ser de gran eficacia para problemas bioinformáticos.

ARNs largos no codificantes

La relativa novedad del ámbito de estudio de los ARNs largos no codificantes abre un amplio abanico de posibilidades en cuanto a su análisis:

- **RNAIntuit**. Una posible línea de mejora consiste, al igual que en IntuitSNP, en refinar la medida de similitud secuencia-motivo utilizada mediante los avances propuestos mediante el escalado de matrices PWM utilizando el factor λ.

  Por otra parte, la cantidad potencial de falsos positivos en la búsqueda de lugares de unión ARN-proteína, se puede también reducir mediante métodos de filtrado posterior para obtener un conjunto de candidatos más reducido. En este sentido, es posible añadir capas adicionales de información, como conservación entre especies y características epigenéticas como metilación o marcas de histonas.

  Con la finalidad de sugerir hipótesis más complejas con interacciones indirectas entre lncRNAs y complejos de proteínas, se propone integrar la metodología de RNAIntuit con redes de interacción proteína-proteína para

buscar conjuntos de proteínas que puedan colaborar con los transcritos del lncRNA en procesos de regulación.

- **MoSI**. El pipeline propuesto realiza un estudio de motivos específicos de lncRNAs tomando en cuenta la direccionalidad de la transcripción. Estos motivos han sido validados con un conjunto de datos reales de motivos de ARN extraídos de la literatura y de bases de datos de micro ARNs. El estudio principal se ha realizado con un conjunto de lncRNAs intergénicos, pero sería también posible realizar el mismo tipo de análisis en conjuntos de lncRNAs con relación funcional, por ejemplo, por conjuntos de lncRNAs específicos de tejido.

  Además, los motivos de lncRNA noveles descubiertos mediante este enfoque son coherentes con el conocimiento existente sobre ARN, pero no se ha podido validar su especificidad en lncRNAs. Estos candidatos a ser motivos específicos de lncRNAs son por ello de especial interés para estudios subsecuentes, ya sea de tipo experimental o computacional. Por otra parte, aunque los resultados utilizando esta aproximación son prometedores, sería posible incluir otras características propias de ARN para realizar una medida híbrida de especificidad de ARN.

Chapter

6

Conclusions and further work

6.1. Conclusions

This work makes several contributions to the study of biological elements presents in gene regulatory processes, with special focus on heterogeneous biological data integration and sequence analysis. More specifically, we have developed methodologies to address three different problems: i) biological entity prioritization in heterogeneous networks, and its aplication to regulatory SNP prediction in relation to diseases and drug repositioning; ii) DNA sequence analysis, in order to search regulatory elements and modules; and iii) long non-coding RNA analysis, biological entities currently drawing scientific interest due to their regulatory potential.

These are relevant contributions in the scope of Personalized Medicine, since Personalized Medicine can only be adressed from a better understanding of gene regulatory mechanisms. The proposed methodologies in this work include different key aspects to progress towards Personalized Medicine: i) the ability to integrate multiple and constantly growing amounts of data, ii) the ability to represent and analyse complex systems in which different biological entities interact, and iii) the ability to analyise individual variations and regulatory

elements in the genome in order to identify their role in gene regulation, and their potential influence in the individual's phenotype.

Prioritization on heterogeneous biological networks

The first line of work proposed in this dissertation encompasses knowledge inference in heterogeneous biological networks built from multiple sources of data. This field of application is of great relevance in a changing area such as biology, where the continous appearance of new experimental techniques and the massive production of data from these, generate even larger amounts of information, which usually are made available to the scientific community as large, diverse and heterogeneous databases.

In this line of work we have proposed **ProphTools** [182], a general methodology for the analysis of heterogeneous networks, that has been applied to two areas of interest: i) the search of regulatory SNPs related to diseases, developing **DiGSNP** [89], a methodology that combines gene-disease prioritization with the analysis of functional regulatory mutations, and ii) drug repositioning, resulting in **DrugNet** [90], an inference methodology in heterogeneous networks including diseases, drugs and proteins (therapeutic targets) as source of relation between both domains.

ProphTools [182] is a general and flexible heterogeneous network propagation methodology. Heterogeneous networks consist of several subnetworks that represent different biological domains of interest. The proposed methodology bases the information propagation within subnetworks in a version of the Random Walk with Restart (RWR) paradigm similar to the Flow Propagation algorithm in [202]. Secondly, a weighted across-network propagation, is performed. The final score is computed by a correlation between the propagation of each target node within the target subnetwork and the propagation result from the query network to the target network of the query nodes.

In order to allow the application of ProphTools to any other area of interest, a meta-model representing heterogeneous networks composed of an arbitrary number of subnetworks, was designed. This allows ProphTools to

perform prioritization in any heterogeneous network configuration. Furthermore, validation paradigms, such as Leave-One-Out (LOO) and LOO with Cross Validation, have been included to enable users to assess the quality of the prioritization results for any given heterogeneous network configuration.

Both ProphTools prioritization approach and its implementation are highly modular, as it has been proven by the inclusion of RANKS [102], a new prioritization methodology based in kernel functions, for within-network prioritization as an alternative to the proposed RWR method. Furthermore, ProphTools software has been developed applying Test-Driven Development techniques and Continuous Integration, in order to guarantee the robustness of the method. ProphTools has been additionally included as a downloadable package from a public repository (available on GitHub http://www.github.com/cnluzon/ProphTools) bajo la licencia de código abierto GPLv3.

As application of ProphTools methodology to a specific area, **DigSNP** [89] is described. DigSNP is a two-step hierarchical prioritization method (disease-gene, gene-SNP), which works as a case of study for both the network prioritization method proposed in this line of work and IntuitSNP, a methodology proposed in the DNA sequence analysis part of this work. In the more general prioritization level, DiGSNP obtains, by means of a heterogeneous disease-protein-gene network prioritization, a ranked list of candidate genes to be related to the query disease, which are analyised at SNP-level in the second step to obtain likely hypotheses about the regulatory mechanisms that guide the proposed disease-SNP relations.

The methodology was validated by cases of study obtained from the biomedical literature with promising results. However, the method was overall difficult to validate due to the lack of validated data in both levels: SNPs that have a proven regulatory functionality that are also known to be disease-related.

Furthermore, the general prioritization methodology proposed is also applied to drug repositioning, an interesting field in Personalized Medicine. The search for new application to already commercialized drugs can save large amounts of money and time in the development of new drugs. Additionally, it can widen the treatment possibilities for each individual, increasing the probability of success in any given patient.

To this end, DrugNet [90], was proposed. DrugNet is an application of ProphTools methodology that shows additional proof of the interest of a general heterogeneous network prioritization methodology. DrugNet is publicly available as a web application in http://genome.ugr.es:9000/drugnet. Regarding its throughput, drug-disease connection network has been validated by means of the LOO and cross-validated LOO methods. Results showed the relevance of being able of integrating multiple domains, since the three-network configuration (drugs, proteins and diseases) had better results than the network involving only drugs and diseases. On the other hand, DrugNet has been also validated using real clinical trials data, showing that it predicts consistently a higher percentage of drugs in later development phases. Furthermore, our method was compared to PREDICT, a state-of-the-art tool for drug repositioning, obtaining promising results.

Detection of regulatory elements on DNA sequences

The second line of work proposed in this thesis encompasses DNA sequence analysis in the search for regulatory elements or groups of regulatory elements. In order to perform this search, it was necessary to take into account on one hand, the vagueness and uncertainty present in DNA motif search, and on the other hand, the biological features known from the gene regulatory processes, exploring the influence of SNPs, the most frequent and studied mutations on the genome, in putative transcription factor binding sites.

The proposed methodologies are based on very consolidated Artificial Intelligence models such as fuzzy logic and frequent itemset extraction from a transactional database, whose parameters have been refined by several cases of study in real datasets.

Two approaches have been described: **IntuitSNP** for the analysis of regulatory potential of SNPs, and **CisMiner** [258], a tool for *cis*-regulatory module search through a whole non-coding genome.

The first of the proposed approaches, **IntuitSNP**, is a methodology to measure the regulatory influence of SNPs that could be located in transcription factor binding sites, changing their binding affinity. This methodology, based on a fuzzy

sequence-motif measure, allows to compute the difference on the affinity levels of the sequence for the different alleles of the SNP. This way, the regulatory potential of the SNP is calculated.

Results validation was performed by a manual review of the literature, and using ORegAnno [151], a database with information about regulatory variants. Additionally, IntuitSNP has a CUDA implementation that allows to perform this study over a large amounf of SNPs in a short time.

The second proposal in this line of work, **CisMiner**, is a methodology that analyses a whole non-coding genome looking for clusters of regulatory elements. This task is performed by a two step process. First, a fuzzy clustering is performed to locate closely located groups of regulatory elements. These fuzzy clusters are stored on a fuzzy transactional database. Secondly, a fuzzy itemset mining algorithm is applied to this fuzzy transactional database in order to obtain groups of frecuent regulatory elements.

CisMiner is a unique methodology in this area, due to its whole non-coding genome scope. Other approaches always require some prior knowledge about the searched modules: their size, the number of involved elements, coexpressed genes or any other type of information that can reduce considerably the amount of data input, due to the combinatorial nature of the frequent itemset mining problem.

Regarding its scope of application, CisMiner has been applied to two whole non-coding genomes of two organisms: *Saccharomyces cerevisiae* and *Drosophila melanogaster*, showing that the methodology is scalable to relatively large genomes. Results have been validated with real data from the scientific literature and from a protein-protein interaction database based on several sources of information. The proposed methodology has been applied, additionally, to both an experimentally validated TFBS dataset, obtained from Harbison *et al.* [290], obtaining high-quality results, and a set of computationally predicted TFBSs, showing the ability of our approach to handle information containing spurious results. On the other hand, although the methodology is robust to this type of information, the influence of the TFBS prediction quality on the quality of the final results, has been proven.

The application to non-coding genomes of large size, such as the human genome, is still an optimization challenge that we intend to undertake in future lines of work. Furthermore, the obtained results have been made public on the CisMiner web (http://genome.ugr.es:9000/cisminer), along with a web application that allows users to perform queries with their own annotation files.

Long non-coding RNAs

The last line of research described in this work is focused on the analysis of a biological entity of great regulatory potential: long non-coding RNAs, or lncRNAs. The features of these biological entities are gradually being discovered by collaborative annotation efforts similar to those performed in the past with coding genes. In this sense, this work proposes several approaches that intend to apply the existing knowledge about RNA sequence to the analysis of lncRNAs.

Two methodologies are proposed in this section: **RNAIntuit**, a tool that searches for protein binding sites in long non-coding RNA sequences taking into account RNA secondary structure; and **MoSI** (Motif Strand Imbalance), a pipeline that searches for lncRNA-specific motifs in a set of lncRNAs. MoSI uses a RNA-specific feature to differentiate between RNA-specific and DNA-specific motifs: transcription strand bias.

The first proposed method, **RNAIntuit**, aims to perform a computational study of the likely regulatory mechanisms of lncRNAs through their interaction with proteins. It is known that such interaction can be part of multiple regulatory functions, such as guiding transcription factor to their binding sites or the opposite, working as a decoy to prevent transcription factors from binding. LncRNAs can also work as scaffolding for more complex molecular structures, among other functionalities. The described goal is achieved by combining a computational prediction of RNA secondary structure and a motif search for protein binding affinity areas, which are filtered according to RNA secondary structure. This approach is validated afterwards by a mechanistic study of Fendrr, a lncRNA gene known for its relation with cancer. This case of study has generated a hypothesis about the mechanisms by which Fendrr regulates a nearby gene, FoxF1. Such

hypotheses are promising, since they are consistent with the current knowledge about FoxF1 regulation. However, existing knowledge about lncRNA's regulatory mechanisms is still scarce, since these entities have rencently drawn interest. Nonetheless, we expect that it will be possible to further validate this approach in the future, using experimental approaches to confirm the initial hypothesis.

Finally, a genome-wide analysis for RNA-specific motifs in lncRNAs is performed by the **MoSI** (Motif Strand Imbalance) pipeline. The proposed methodology exploits a RNA-specific feature to differentiate overrepresented motifs in the analysed sequences that come from the DNA sequence they were transcribed from, and the motifs that are specific to the RNA sequence. To validate this methodology, a set of known RNA-specific motifs has been gathered from the literature and RNA-specific databases. This set has been used as a labeled result set to assess the quality of MoSI results and to select a threshold cutoff for the selection of putative novel lncRNA-specific motifs.

Furthermore, the resulting putative interest lncRNA-specific motif set has been genome-wide annotated both relative to transcript start and also in absolute genomic coordinates. These annotations are, along with the analysis methodology proposed, the results of this last line of work. We expect that this methodology can be applied to future versions of the lncRNA annotation databases, allowing us to reproduce the procedure and improve the annotations in newer versions. The putative lncRNA-specific motif annotation results are themselves a resource that can be used as an input for further analysis. This analysis can, in turn, shed light on these new biological entities of interest.

6.2. Further work

The multiple approaches performed in this thesis open new paths for further research. The most promising lines of work are described below.

Prioritization on heterogeneous biological networks

Heterogeneous biological data source integration opens some new lines of work:

- **Application of ProphTools to new areas of interest**. The proposed methodology is independent of the number of different biological entities (*i.e.* the number of different subnetworks present in the model) involved in the proiritization process, as it also independent of the data sources that are being integrated. For this reason, we intend to explore other areas of application, such as lncRNA-disease prioritization.
- **Integration of new data sources in DrugNet**. As it has been said, the methodology proposed in DrugNet allows to extend the number of data sources. In this sense, it would be of interest to add new data sources, either implicitly, for instance, in the form of side-effects information, or explicitly, searching for new key entities that can be related to the entities in DrugNet heterogeneous network (diseases, proteins, drugs).
- **Throughput improvement in ProphTools prioritization**. Random Walk with Restart prioritization has some computational limitations, that increase with the size of the network, since it require the computing of a matrix inversion. This process has been precomputed for this proposal, but such precomputation is costful in memory requirements. This problem could be solved by more advanced techniques for precomputing, such as Tong *et al.* [375]. It would be also possible to perform a CUDA parallelization of the algorithm. GPU parallelization has proven useful for bioinformatics problems.

- **Throughput improvement in LOO tests**. Even though the prioritization methodology proposed is fast for a single query, the integrated methods for LOO and cross validation with LOO are more computationally demanding, since they perform prioritization for each node in the query network that is connected with the target network. In this sense, a paralellization of these tests is proposed.
- **Including further propagation modes within and across networks**. The modular nature of ProphTools allows to include alternative methods for prioritization. The proposed methodology for within-network propagation, a RWR modified by a laplacian transformation of the addjacency matrix, opens the door to other matrix transformations suchas the kernel functions proposed in RANKS [102]. Additionally, other prioritization approaches could be added to the propagation across networks, as an alternative to the current weighted across-network propagation.
- **Ontology-specific within-network propagation modes**. Ontology-specific features, such as its hierarchical nature, from more general to more specific concepts, could be used to improve the propagation in such networks.

Detection of regulatory elements on biological sequences

Regarding regulatory elements and modules in biological sequence, the following improvements are proposed:

- **IntuitSNP**. It would be possible to improve IntuitSNP by improving SC_{intuit} sequence-motif similarity measure, by using the λ factor of PWM adjustment described. Additionally, it has been observed that an accumulation of modifications in low conserved positions of a motif can significantly reduce the affinity score. In this sense, it would be of interest to further study this mechanism in order to increase IntuitSNP's sensitivity.

  On the other hand, the inclusion of more detailed information about each SNP's alleles, such as relative frequency in relation to the query disease's incidence level. Furthermore, the fuzzy similarity measure SC_{intuit} could be

integrated with additional knowledge about the sequence in which the SNP is located, such as epigenetics information. Along this epigenetics data that could be integrated in a new measure, chromatine accesibility data has been proven to have influence in transcription factor binding to a sequence [376]). Methylation data, which is starting to be available at single-nucleotide level [378], would be of interest as well, since it has been shown to be also related to phenomena such as gene silecing [377].

- **CisMiner**. The proposed methodology to find *cis*-regulatory modules allows improvement in several lines of work. First, the proposed fuzzy frequent itemset mining could be improved to scale to the human genome problem size. These improvements could be algorithmic, adapting the paradigm to other models. On the other hand, a CUDA parallelization could be performed.

Long non-coding RNAs

The novelty of this area of work, due to the recent interest drawn by lncRNAs opens a wide range of lines of work.

- **RNAIntuit**. A possible improvement to RNAIntuit is, as in IntuitSNP, refining the sequence-motif similarity measure by applying the scaling with λ factor proposed in the regulatory sequence analysis chapter.

  Additionally, the potential amount of false positives in the transcription factor binding site search for RNA-protein can be reduced by subsequent filtering steps to obtain a more reduced set of candidates. In this sense, it is possible to add additional data layers, such as evolutionary conservation and epigenetic features such as methylation or histone marks.

  On the other hand, in order to suggest more complex hypotheses with indirect interactions between lncRNAs and protein complexes, we propose to integrate RNAIntuit methodology with protein-protein interaction networks, to search for protein groups that cooperate with lncRNA transcripts in regulatory processes.

- **MoSI**. The proposed pipeline performs a lncRNA-specific study taking into account the RNA transcription strand bias. These motifs have been validated by a real dataset consisting of motifs extracted manually from the literature and micro RNA seed data extracted from micro RNA databases. The main study has been performed on a set of intergenic lncRNAs (lincRNAs), but it could be also possible to perform the same type of analysis on sets of lncRNAs functionally related, for instance, to sets of tissue-specific lncRNAs.

  Furthermore, the putative novel lncRNA motifs discovered by this approach are consistent with the existing knowledge on RNA, but it has not been possible to validate their lncRNA-specificity. These putative lncRNA-specific motifs are of interest for further research, either experimental or computational. On the other hand, although the results obtained by this approach are promising, it would be possible to include additional RNA-specific features to include in the RNA-specificity measure as a hybrid approach.

Parte IV

Publicaciones

Capítulo

7

Trabajos publicados

Publicaciones derivadas del trabajo expuesto en la memoria

Relacionados con el desarrollo de esta memoria, se han publicado los siguientes trabajos:

Publicaciones en revistas

- Navarro, C., Lopez, F. J., Cano, C., Garcia-Alcalde, F., Blanco, A. (2014). CisMiner: Genome-wide in-silico cis-regulatory module prediction by fuzzy itemset mining. PloS one, 9(9), e108065.
- Martínez, V., Navarro, C., Cano, C., Fajardo, W., Blanco, A. (2015). DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data. Artificial intelligence in medicine, 63(1), 41-49.
- Ma, X., Ezer, D., Navarro, C., Adryan, B. (2015). Reliable scaling of position weight matrices for binding strength comparisons between transcription factors. BMC bioinformatics, 16(1), 1.
- Navarro, Carmen, Martínez, Víctor, Cano, Carlos, and Blanco, Armando. ProphTools: General Prioritization Tools for Heterogeneous Biological Networks. GigaScience (2016) - Submitted

Publicaciones en actas de congresos

- Navarro, C., Cano, C., Blanco, A., García-Alcalde, F. (2012). DiGSNP: a web tool for Disease-Gene-SNP hierarchical prioritization. EMBnet. journal, 18(B), pp-74.
- Martınez, V., Navarro, C., Cano, C., Blanco, A. (2013). Network-based drug-disease relation prioritization using ProphNet. IWBBIO 2013
- Navarro, C., Cano, C., Cuadros, M., Blanco, A. (2015). Computationally assessing RNA-protein binding affinity in long non-coding RNAs. IWBBIO 2015, 63.
- Navarro, C., Cano, C., Cuadros, M., Herrera-Merchan, A., Molina, M., Blanco, A. (2016, April). A Mechanistic Study of lncRNA Fendrr Regulation of FoxF1 Lung Cancer Tumor Supressor. In International Conference on Bioinformatics and Biomedical Engineering (pp. 781-789). Springer International Publishing.

Herramientas disponibles

- **ProphTools**. Paquete python descargable desde GitHub: `http://www.github.com/cnluzon/prophtools`
- **DrugNet**. Web disponible en `http://genome.ugr.es:9000/drugnet`.
- **CisMiner**. Web disponible en `http://genome.ugr.es:9000/cisminer`.

Índice de figuras

Índice de tablas

Glosario

ADN Ácido Desoxirribonucleico.

ARN Ácido Ribonucleico.

bp Base-pair. Abreviación de par de bases. Medida de longitud de secuencia de cadena de doble hélice.

ChIP Chromatine Immuno-Precipitation. Inmunoprecipitación de cromatina..

CRM Cis-Regulatory Module. Módulo regulador en cis.

GWAS Genome-Wide Association Study. Estudio de asociación a genoma completo.

lncRNA Long non-coding RNA. ARN largo no codificante.

miRNA Micro-RNA. Micro ARN.

mRNA Messenger RNA. ARN mensajero.

NGS Next-Generation Sequencing. Secuenciación de siguiente generación.

nt Abreviación de nucleótido. Medida de longitud de secuencia de cadena única.

ORF Open Reading Frame. Secuencia codificante que comienza con un codón de inicio.

PWM Position-Weight Matrix. Matriz de pesos por posición en un motivo.

RBP RNA-binding protein. Proteína capaz de unirse al ARN.

RWR Random Walks with Restarts.

TF Transcription factor. Factor de transcripción.

TFBS Transcription Factor Binding Site. Lugar de unión de factores de transcripción.

tRNA Transfer RNA.

Bibliografía

[1] International Human Genome Sequencing Consortium et al. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[2] Amir Rahimi. The promising prospects of precision medicine. *Journal of Advanced Medical Sciences and Applied Technologies*, 2(3):244–246, 2016.

[3] Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, 6 edition, 2014.

[4] James D. Watson, Tania A. Baker, Stephen P. Bell, Alexander Gann, Michael Levine, and Richard Losick. *Molecular Biology of the Gene*. Pearson, 7 edition, 2013.

[5] Erwin Chargaff, Stephen Zamenhof, and Charlotte Green. Human desoxypentose nucleic acid: Composition of human desoxypentose nucleic acid. *Nature*, 1950.

[6] R. Franklin and R.G. Gosling. Molecular configuration in sodium thymonucleate. *Nature*, 171(4356):740–741, 1953.

[7] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[8] B Alberts, A. Johnson, J. Lewis, et al. *Molecular Biology of the Cell*. New York: Garland Science, 2002.

[9] Tim R Mercer and John S Mattick. Structure and function of long noncoding rnas in epigenetic regulation. *Nature structural & molecular biology*, 20(3):300–307, 2013.

[10] Sarah B Ng, Emily H Turner, Peggy D Robertson, Steven D Flygare, Abigail W Bigham, Choli Lee, Tristan Shaffer, Michelle Wong, Arindam Bhattacharjee, Evan E Eichler, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276, 2009.

[11] Matthew J Hangauer, Ian W Vaughn, and Michael T McManus. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding rnas. *PLoS Genet*, 9(6):e1003569, 2013.

[12] Francis S Collins and Monique K Mansoura. The human genome project. revealing the shared inheritance of all humankind. *Cancer*, 91(1 Suppl):221–225, 2001.

[13] Barkur S Shastry. Snps: impact on gene function and phenotype. *Single Nucleotide Polymorphisms: Methods and Protocols*, pages 3–22, 2009.

[14] Kim M Keeling, Ming Du, and David M Bedwell. Therapies of nonsense-associated diseases. 2000.

[15] Guangfu Jin, Jielin Sun, Sarah D Isaacs, Kathleen E Wiley, Seong-Tae Kim, Lisa W Chu, Zheng Zhang, Hui Zhao, Siqun Lilly Zheng, William B Isaacs, et al. Human polymorphisms at long non-coding rnas (lncrnas) and association with prostate cancer risk. *Carcinogenesis*, 32(11):1655–1659, 2011.

[16] Mihaela Pertea and Steven L Salzberg. Between a chicken and a grape: estimating the number of human genes. *Genome biology*, 11(5):1, 2010.

[17] Kevin V Morris and John S Mattick. The rise of regulatory rna. *Nature reviews. Genetics*, 15(6):423, 2014.

[18] Ewan A Gibb, Carolyn J Brown, and Wan L Lam. The functional role of long non-coding rna in human carcinomas. *Molecular cancer*, 10(1):1, 2011.

[19] John S Mattick and Igor V Makunin. Non-coding rna. *Human molecular genetics*, 15(suppl 1):R17–R29, 2006.

[20] Matthew K Iyer, Yashar S Niknafs, Rohit Malik, Udit Singhal, Anirban Sahu, Yasuyuki Hosono, Terrence R Barrette, John R Prensner, Joseph R Evans, Shuang Zhao, et al. The landscape of long noncoding rnas in the human transcriptome. *Nature genetics*, 47(3):199–208, 2015.

[21] Jeffrey J Quinn and Howard Y Chang. Unique features of long non-coding rna biogenesis and function. *Nature Reviews Genetics*, 17(1):47–62, 2016.

[22] Stefanie A Mortimer, Mary Anne Kidwell, and Jennifer A Doudna. Insights into rna structure and function from genome-wide studies. *Nature reviews Genetics*, 15(7):469–479, 2014.

[23] Moran N Cabili, Margaret C Dunagin, Patrick D McClanahan, Andrew Biaesch, Olivia Padovan-Merhar, Aviv Regev, John L Rinn, and Arjun Raj. Localization and abundance analysis of human lncrnas at single-cell and single-molecule resolution. *Genome biology*, 16(1):1, 2015.

[24] Kevin C Wang and Howard Y Chang. Molecular mechanisms of long noncoding rnas. *Molecular cell*, 43(6):904–914, 2011.

[25] Tim R Mercer, Marcel E Dinger, and John S Mattick. Long non-coding rnas: insights into functions. *Nature Reviews Genetics*, 10(3):155–159, 2009.

[26] Orly Wapinski and Howard Y Chang. Long noncoding rnas and human disease. *Trends in cell biology*, 21(6):354–361, 2011.

[27] Wenhu Pi, Xingguo Zhu, Min Wu, Yongchao Wang, Sadanand Fulzele, Ali Eroglu, Jianhua Ling, and Dorothy Tuan. Long-range function of an intergenic retrotransposon. *Proceedings of the National Academy of Sciences*, 107(29):12992–12997, 2010.

[28] Thomas Derrien, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, David Martin, Angelika Merkel, David G Knowles, et al. The gencode v7 catalog of human long noncoding rnas: analysis of their gene structure, evolution, and expression. *Genome research*, 22(9):1775–1789, 2012.

[29] Xiu Cheng Quek, Daniel W Thomson, Jesper LV Maag, Nenad Bartonicek, Bethany Signal, Michael B Clark, Brian S Gloss, and Marcel E Dinger. lncrnadb v2. 0: expanding the reference database for functional long noncoding rnas. *Nucleic acids research*, page gku988, 2014.

[30] Geng Chen, Ziyun Wang, Dongqing Wang, Chengxiang Qiu, Mingxi Liu, Xing Chen, Qipeng Zhang, Guiying Yan, and Qinghua Cui. Lncrnadisease: a database for long-non-coding rna-associated diseases. *Nucleic acids research*, 41(D1):D983–D986, 2013.

[31] Chaoyong Xie, Jiao Yuan, Hui Li, Ming Li, Guoguang Zhao, Dechao Bu, Weimin Zhu, Wei Wu, Runsheng Chen, and Yi Zhao. Noncodev4: exploring the world of long non-coding rna genes. *Nucleic acids research*, 42(D1):D98–D103, 2014.

[32] Yi Zhao, Hui Li, Shuangsang Fang, Yue Kang, Yajing Hao, Ziyang Li, Dechao Bu, Ninghui Sun, Michael Q Zhang, Runsheng Chen, et al. Noncode 2016: an informative and valuable data source of long non-coding rnas. *Nucleic acids research*, page gkv1252, 2015.

[33] Pieter-Jan Volders, Kenneth Verheggen, Gerben Menschaert, Klaas Vandepoele, Lennart Martens, Jo Vandesompele, and Pieter Mestdagh. An update on lncipedia: a database for annotated human lncrna sequences. *Nucleic acids research*, 43(D1):D174–D180, 2015.

[34] Shangwei Ning, Jizhou Zhang, Peng Wang, Hui Zhi, Jianjian Wang, Yue Liu, Yue Gao, Maoni Guo, Ming Yue, Lihua Wang, et al. Lnc2cancer: a

manually curated database of experimentally supported lncrnas associated with various human cancers. *Nucleic acids research*, page gkv1094, 2015.

[35] Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

[36] Erwin L van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends in genetics*, 30(9):418–426, 2014.

[37] Elaine R Mardis. Next-generation dna sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402, 2008.

[38] Elaine R Mardis. The impact of next-generation sequencing technology on genetics. *Trends in genetics*, 24(3):133–141, 2008.

[39] Jason A Reuter, Damek V Spacek, and Michael P Snyder. High-throughput sequencing technologies. *Molecular cell*, 58(4):586–597, 2015.

[40] Richard J Roberts, Mauricio O Carneiro, and Michael C Schatz. The advantages of smrt sequencing. *Genome biology*, 14(7):1, 2013.

[41] Miten Jain, Ian T Fiddes, Karen H Miga, Hugh E Olsen, Benedict Paten, and Mark Akeson. Improved data analysis for the minion nanopore sequencer. *Nature methods*, 12(4):351–356, 2015.

[42] Sara Goodwin, James Gurtowski, Scott Ethe-Sayers, Panchajanya Deshpande, Michael C Schatz, and W Richard McCombie. Oxford nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome research*, 25(11):1750–1756, 2015.

[43] Victor E Velculescu, Lin Zhang, Bert Vogelstein, and Kenneth W Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484, 1995.

[44] Karl V Voelkerding, Shale A Dames, and Jacob D Durtschi. Next-generation sequencing: from basic research to diagnostics. *Clinical chemistry*, 55(4):641–658, 2009.

[45] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[46] Mark J Solomon, Pamela L Larsen, and Alexander Varshavsky. Mapping proteindna interactions in vivo with formaldehyde: Evidence that histone h4 is retained on a highly transcribed gene. *Cell*, 53(6):937–947, 1988.

[47] Christian A Heid, Junko Stevens, Kenneth J Livak, and P Mickey Williams. Real time quantitative pcr. *Genome research*, 6(10):986–994, 1996.

[48] Bing Ren, François Robert, John J Wyrick, Oscar Aparicio, Ezra G Jennings, Itamar Simon, Julia Zeitlinger, Jörg Schreiber, Nancy Hannett, Elenita Kanin, et al. Genome-wide location and function of dna binding proteins. *Science*, 290(5500):2306–2309, 2000.

[49] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.

[50] Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England Journal of Medicine*, 372(9):793–795, 2015.

[51] David J Duffy. Problems, challenges and promises: perspectives on precision medicine. *Briefings in bioinformatics*, page bbv060, 2015.

[52] Nicholas J Schork. Personalized medicine: time for one-person trials. *Nature*, 520(7549):609–11, 2015.

[53] Jessica Xin Hu, Cecilia Engel Thomas, and Søren Brunak. Network biology concepts in complex disease comorbidities. *Nature Reviews Genetics*, 2016.

[54] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[55] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[56] David C Lay. Linear algebra and its applications, 2005.

[57] Thiago Christiano Silva and Liang Zhao. Complex networks. In *Machine Learning in Complex Networks*, pages 15–70. Springer, 2016.

[58] P ERDdS and A R&WI. On random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959.

[59] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[60] Jie Wu and Duncan J Watts. Small worlds: the dynamics of networks between order and randomness. *Acm Sigmod Record*, 31(4):74–75, 2002.

[61] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[62] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[63] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

[64] Reka Albert. Scale-free networks in cell biology. *Journal of cell science*, 118(21):4947–4957, 2005.

[65] Mark EJ Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.

[66] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[67] Shi Zhou and Raúl J Mondragón. The rich-club phenomenon in the internet topology. *IEEE Communications Letters*, 8(3):180–182, 2004.

[68] Tore Opsahl, Vittoria Colizza, Pietro Panzarasa, and Jose J Ramasco. Prominence and control: the weighted rich-club effect. *Physical review letters*, 101(16):168702, 2008.

[69] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[70] Hyun-Joo Kim and Jin Min Kim. Cyclic topology in complex networks. *Physical Review E*, 72(3):036109, 2005.

[71] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[72] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[73] Ulrik Brandes and Thomas Erlebach. *Network analysis: methodological foundations*, volume 3418. Springer Science & Business Media, 2005.

[74] Mark Newman. *Networks: an introduction*. Oxford university press, 2010.

[75] Ernesto Estrada and Naomichi Hatano. Communicability in complex networks. *Physical Review E*, 77(3):036111, 2008.

[76] Nicola Perra and Santo Fortunato. Spectral centrality measures in complex networks. *Physical Review E*, 78(3):036107, 2008.

[77] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[78] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Random walk with restart: fast solutions and applications. *Knowledge and Information Systems*, 14(3):327–346, 2008.

[79] Paul J Flory. *Principles of polymer chemistry*. Cornell University Press, 1953.

[80] Andrew Chatr-Aryamontri, Bobby-Joe Breitkreutz, Rose Oughtred, Lorrie Boucher, Sven Heinicke, Daici Chen, Chris Stark, Ashton Breitkreutz, Nadine Kolas, Lara O'Donnell, et al. The biogrid interaction database: 2015 update. *Nucleic acids research*, 43(D1):D470–D478, 2015.

[81] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, page gkw937, 2016.

[82] UniProt Consortium et al. Uniprot: a hub for protein information. *Nucleic acids research*, page gku989, 2014.

[83] TS Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, et al. Human protein reference database—2009 update. *Nucleic acids research*, 37(suppl 1):D767–D772, 2009.

[84] Stanley Dagley, Donald E Nicholson, et al. An introduction to metabolic pathways. *An introduction to metabolic pathways.*, 1970.

[85] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Kegg as a reference resource for gene and protein annotation. *Nucleic acids research*, page gkv1070, 2015.

[86] Victoria Petri, Pushkala Jayaraman, Marek Tutaj, G Thomas Hayman, Jennifer R Smith, Jeff De Pons, Stanley JF Laulederkind, Timothy F Lowry,

Rajni Nigam, Shur-Jen Wang, et al. The pathway ontology–updates and applications. *Journal of biomedical semantics*, 5(1):1, 2014.

[87] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. Pid: the pathway interaction database. *Nucleic acids research*, 37(suppl 1):D674–D679, 2009.

[88] Víctor Martínez, Carlos Cano, and Armando Blanco. Prophnet: A generic prioritization method through propagation of information. *BMC bioinformatics*, 15(1):1, 2014.

[89] Carmen Navarro, Carlos Cano, Armando Blanco, and Fernando García-Alcalde. Digsnp: a web tool for disease-gene-snp hierarchical prioritization. *EMBnet. journal*, 18(B):pp–74, 2012.

[90] Víctor Martínez, Carmen Navarro, Carlos Cano, Waldo Fajardo, and Armando Blanco. Drugnet: Network-based drug–disease prioritization by integrating heterogeneous data. *Artificial intelligence in medicine*, 63(1):41–49, 2015.

[91] Joanna S Amberger, Carol A Bocchini, François Schiettecatte, Alan F Scott, and Ada Hamosh. Omim. org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. *Nucleic acids research*, 43(D1):D789–D798, 2015.

[92] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946, 2012.

[93] Sune Pletscher-Frankild, Albert Pallejà, Kalliopi Tsafou, Janos X Binder, and Lars Juhl Jensen. Diseases: Text mining and data integration of disease–gene associations. *Methods*, 74:83–89, 2015.

[94] Vivian Law, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski, David Arndt, Michael Wilson, Vanessa Neveu, et al. Drugbank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, 42(D1):D1091–D1097, 2014.

[95] Yanli Wang, Tugba Suzek, Jian Zhang, Jiyao Wang, Siqian He, Tiejun Cheng, Benjamin A Shoemaker, Asta Gindulyte, and Stephen H Bryant. Pubchem bioassay: 2014 update. *Nucleic acids research*, page gkt978, 2013.

[96] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

[97] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology*, 6(1):343, 2010.

[98] Ana Kozomara and Sam Griffiths-Jones. mirbase: annotating high confidence micrornas using deep sequencing data. *Nucleic acids research*, 42(D1):D68–D73, 2014.

[99] Nathan Wong and Xiaowei Wang. mirdb: an online resource for microrna target prediction and functional annotations. *Nucleic acids research*, page gku1104, 2014.

[100] Jun-Hao Li, Shun Liu, Hui Zhou, Liang-Hu Qu, and Jian-Hua Yang. starbase v2. 0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data. *Nucleic acids research*, page gkt1248, 2013.

[101] Simona Panni and Simona E Rombo. Searching for repetitions in biological networks: methods, resources and tools. *Briefings in bioinformatics*, 16(1):118–136, 2015.

[102] Giorgio Valentini, Giuliano Armano, Marco Frasca, Jianyi Lin, Marco Mesiti, and Matteo Re. Ranks: a flexible tool for node label ranking and classification in biological networks. *Bioinformatics*, page btw235, 2016.

[103] Guomin Ren and Zhihua Liu. Netcad: a network analysis tool for coronary artery disease-associated ppi network. *Bioinformatics*, 29(2):279–280, 2013.

[104] Andrea Franceschini, Jianyi Lin, Christian von Mering, and Lars Juhl Jensen. Svd-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. *Bioinformatics*, page btv696, 2015.

[105] Charles Blatti and Saurabh Sinha. Characterizing gene sets using discriminative random walks with restart on heterogeneous biological networks. *Bioinformatics*, page btw151, 2016.

[106] Mohashin Pathan, Shivakumar Keerthikumar, Ching-Seng Ang, Lahiru Gangoda, Camelia YJ Quek, Nicholas A Williamson, Dmitri Mouradov, Oliver M Sieber, Richard J Simpson, Agus Salim, et al. Funrich: An open access standalone functional enrichment and interaction network analysis tool. *Proteomics*, 15(15):2597–2601, 2015.

[107] TaeHyun Hwang, Wei Zhang, Maoqiang Xie, Jinfeng Liu, and Rui Kuang. Inferring disease and gene set associations with rank coherence in networks. *Bioinformatics*, 27(19):2692–2699, 2011.

[108] Wangshu Zhang, Yong Chen, Fengzhu Sun, and Rui Jiang. Domainrbf: a bayesian regression approach to the prioritization of candidate domains for complex diseases. *BMC systems biology*, 5(1):1, 2011.

[109] Guanghui Hu and Pankaj Agarwal. Human disease-drug network based on genomic expression profiles. *PloS one*, 4(8):e6536, 2009.

[110] Feixiong Cheng, Chuang Liu, Jing Jiang, Weiqiang Lu, Weihua Li, Guixia Liu, Weixing Zhou, Jin Huang, and Yun Tang. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol*, 8(5):e1002503, 2012.

[111] Xing Chen, Ming-Xi Liu, and Gui-Ying Yan. Drug–target interaction prediction by random walk on the heterogeneous network. *Molecular BioSystems*, 8(7):1970–1978, 2012.

[112] Lotfi A Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.

[113] Krassimir T Atanassov. Intuitionistic fuzzy sets. *Fuzzy sets and Systems*, 20(1):87–96, 1986.

[114] Krassimir Atanassov and Christo Georgiev. Intuitionistic fuzzy prolog. *Fuzzy Sets and Systems*, 53(2):121–128, 1993.

[115] Supriya Kumar De, Ranjit Biswas, and Akhil Ranjan Roy. An application of intuitionistic fuzzy sets in medical diagnosis. *Fuzzy sets and Systems*, 117(2):209–213, 2001.

[116] Vahid Khatibi and Gholam Ali Montazer. Intuitionistic fuzzy set vs. fuzzy set application in medical pattern recognition. *Artificial Intelligence in Medicine*, 47(1):43–52, 2009.

[117] E Szmidt and J Kacprzyk. Intuitionistic fuzzy sets in group decision making. *Notes on IFS*, 2(1):11–14, 1996.

[118] Wen-Liang Hung and Miin-Shen Yang. Similarity measures of intuitionistic fuzzy sets based on l p metric. *International Journal of Approximate Reasoning*, 46(1):120–136, 2007.

[119] Michael J Buck and Jason D Lieb. Chip-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–360, 2004.

[120] Valentina Boeva, Didier Surdez, Noëlle Guillon, Franck Tirode, Anthony P Fejes, Olivier Delattre, and Emmanuel Barillot. De novo motif identification improves the accuracy of predicting transcription factor binding sites in chip-seq data analysis. *Nucleic acids research*, 38(11):e126–e126, 2010.

[121] Anthony P Fejes, Gordon Robertson, Mikhail Bilenky, Richard Varhol, Matthew Bainbridge, and Steven JM Jones. Findpeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15):1729–1730, 2008.

[122] Ming Hu, Jindan Yu, Jeremy MG Taylor, Arul M Chinnaiyan, and Zhaohui S Qin. On the detection and refinement of transcription factor binding sites using chip-seq data. *Nucleic acids research*, 38(7):2154–2167, 2010.

[123] Kun Liang and Sündüz Keleş. Detecting differential binding of transcription factors with chip-seq. *Bioinformatics*, 28(1):121–122, 2012.

[124] Morgane Thomas-Chollier, Carl Herrmann, Matthieu Defrance, Olivier Sand, Denis Thieffry, and Jacques van Helden. Rsat peak-motifs: motif analysis in full-size chip-seq datasets. *Nucleic acids research*, 40(4):e31–e31, 2012.

[125] Jason Ernst, Heather L Plasterer, Itamar Simon, and Ziv Bar-Joseph. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome research*, 20(4):526–536, 2010.

[126] Shane Neph and Martin Tompa. Microfootprinter: a tool for phylogenetic footprinting in prokaryotic genomes. *Nucleic acids research*, 34(suppl 2):W366–W368, 2006.

[127] Mathieu Blanchette and Martin Tompa. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome research*, 12(5):739–748, 2002.

[128] Alexander Kawrykow, Gary Roumanis, Alfred Kam, Daniel Kwak, Clarence Leung, Chu Wu, Eleyine Zarour, Luis Sarmenta, Mathieu Blanchette, Jérôme Waldispühl, et al. Phylo: a citizen science approach for improving multiple sequence alignment. *PloS one*, 7(3):e31362, 2012.

[129] Anthony Mathelier, Oriol Fornes, David J Arenillas, Chih-yu Chen, Grégoire Denay, Jessica Lee, Wenqiang Shi, Casper Shyr, Ge Tan, Rebecca Worsley-Hunt, et al. Jaspar 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic acids research*, page gkv1176, 2015.

[130] Vea Matys, Ellen Fricke, R Geffers, Ellen Gößling, Martin Haubrock, R Hehl, Klaus Hornischer, Dagmar Karas, Alexander E Kel, Olga V Kel-Margoulis, et al. Transfac®: transcriptional regulation, from patterns to profiles. *Nucleic acids research*, 31(1):374–378, 2003.

[131] Fernando Garcia, Francisco J Lopez, Carlos Cano, and Armando Blanco. Fisim: a new similarity measure between transcription factor binding sites based on the fuzzy integral. *BMC bioinformatics*, 10(1):1, 2009.

[132] Dustin E Schones, Pavel Sumazin, and Michael Q Zhang. Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, 21(3):307–313, 2005.

[133] In-Geol Choi, Jaimyoung Kwon, and Sung-Hou Kim. Local feature frequency profile: a method to measure structural similarity in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3797–3802, 2004.

[134] Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome biology*, 8(2):1, 2007.

[135] Thomas D Schneider and R Michael Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, 18(20):6097–6100, 1990.

[136] Gavin E Crooks, Gary Hon, John-Marc Chandonia, and Steven E Brenner. Weblogo: a sequence logo generator. *Genome research*, 14(6):1188–1190, 2004.

[137] Xiaohui Xie, Jun Lu, EJ Kulbokas, Todd R Golub, Vamsi Mootha, Kerstin Lindblad-Toh, Eric S Lander, and Manolis Kellis. Systematic discovery of regulatory motifs in human promoters and 3' utrs by comparison of several mammals. *Nature*, 434(7031):338–345, 2005.

[138] Xiaohui Xie, Tarjei S Mikkelsen, Andreas Gnirke, Kerstin Lindblad-Toh, Manolis Kellis, and Eric S Lander. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of ctcf insulator sites. *Proceedings of the National Academy of Sciences*, 104(17):7145–7150, 2007.

[139] Daniel E Newburger and Martha L Bulyk. Uniprobe: an online database of protein binding microarray data on protein–dna interactions. *Nucleic acids research*, 37(suppl 1):D77–D82, 2009.

[140] Michael F Berger, Gwenael Badis, Andrew R Gehrke, Shaheynoor Talukder, Anthony A Philippakis, Lourdes Pena-Castillo, Trevis M Alleyne, Sanie Mnaimneh, Olga B Botvinnik, Esther T Chan, et al. Variation in homeodomain dna binding revealed by high-resolution analysis of sequence preferences. *Cell*, 133(7):1266–1276, 2008.

[141] Gerald Z Hertz, George W Hartzell, and Gary D Stormo. Identification of consensus patterns in unaligned dna sequences known to be functionally related. *Computer applications in the biosciences: CABIOS*, 6(2):81–92, 1990.

[142] Albin Sandelin, Wyeth W Wasserman, and Boris Lenhard. Consite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic acids research*, 32(suppl 2):W249–W252, 2004.

[143] Panayiotis V Benos, Alan S Lapedes, and Gary D Stormo. Probabilistic code for dna recognition by proteins of the egr family. *Journal of molecular biology*, 323(4):701–727, 2002.

[144] Martha L Bulyk, Philip LF Johnson, and George M Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic acids research*, 30(5):1255–1261, 2002.

[145] Andrija Tomovic and Edward J Oakeley. Position dependencies in transcription factor binding sites. *Bioinformatics*, 23(8):933–941, 2007.

[146] Fatemeh Zare-Mirakabad, Hayedeh Ahrabian, Mehdei Sadeghi, Abbas Nowzari-Dalini, and Bahram Goliaei. New scoring schema for finding motifs in dna sequences. *BMC bioinformatics*, 10(1):1, 2009.

[147] Fernando Garcia-Alcalde, Armando Blanco, and Adrian J Shepherd. An intuitionistic approach to scoring dna sequences against transcription factor binding site motifs. *BMC bioinformatics*, 11(1):551, 2010.

[148] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.

[149] Paul Flicek, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, et al. Ensembl 2012. *Nucleic acids research*, page gkr991, 2011.

[150] Donna Karolchik, Robert Baertsch, Mark Diekhans, Terrence S Furey, Angie Hinrichs, YT Lu, Krishna M Roskin, Matthias Schwartz, Charles W Sugnet, Daryl J Thomas, et al. The ucsc genome browser database. *Nucleic acids research*, 31(1):51–54, 2003.

[151] SB Montgomery, Obi L Griffith, Monica C Sleumer, Casey M Bergman, Misha Bilenky, ED Pleasance, Y Prychyna, X Zhang, and Steven JM Jones. Oreganno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics*, 22(5):637–640, 2006.

[152] Geoff Macintyre, James Bailey, Izhak Haviv, and Adam Kowalczyk. is-rsnp: a novel technique for in silico regulatory snp detection. *Bioinformatics*, 26(18):i524–i530, 2010.

[153] Thomas Manke, Matthias Heinig, and Martin Vingron. Quantifying the effect of sequence variation on regulatory interactions. *Human mutation*, 31(4):477–483, 2010.

[154] Helge G Roider, Aditi Kanhere, Thomas Manke, and Martin Vingron. Predicting transcription factor affinities to dna from a biophysical model. *Bioinformatics*, 23(2):134–141, 2007.

[155] Adam Ameur, Alvaro Rada-Iglesias, Jan Komorowski, and Claes Wadelius. Identification of candidate regulatory snps by combination of transcription-factor-binding site prediction, snp genotyping and haplochip. *Nucleic acids research*, 37(12):e85–e85, 2009.

[156] Lucía Conde, Juan M Vaquerizas, Javier Santoyo, Fátima Al-Shahrour, Sergio Ruiz-Llorente, Mercedes Robledo, and Joaquín Dopazo. Pupasnp finder: a web tool for finding snps with putative effect at transcriptional level. *Nucleic acids research*, 32(suppl 2):W242–W248, 2004.

[157] Dongwon Lee, David U Gorkin, Maggie Baker, Benjamin J Strober, Alessandro L Asoni, Andrew S McCallion, and Michael A Beer. A method to predict the impact of regulatory variants from dna sequence. *Nature genetics*, 47(8):955–961, 2015.

[158] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods*, 12(10):931–934, 2015.

[159] François Spitz and Eileen EM Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626, 2012.

[160] Maria I Arnone and Eric H Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124(10):1851–1864, 1997.

[161] Eric H Davidson. *Genomic regulatory systems: in development and evolution*. Academic Press, 2001.

[162] Peter Van Loo and Peter Marynen. Computational methods for the detection of cis-regulatory modules. *Briefings in Bioinformatics*, page bbp025, 2009.

[163] Martin C Frith, Ulla Hansen, and Zhiping Weng. Detection of cis-element clusters in higher eukaryotic dna. *Bioinformatics*, 17(10):878–889, 2001.

[164] Martin C Frith, Michael C Li, and Zhiping Weng. Cluster-buster: Finding dense clusters of motifs in dna sequences. *Nucleic acids research*, 31(13):3666–3668, 2003.

[165] Saurabh Sinha, Erik Van Nimwegen, and Eric D Siggia. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19(suppl 1):i292–i301, 2003.

[166] Carl Herrmann, Bram Van de Sande, Delphine Potier, and Stein Aerts. i-cistarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic acids research*, 40(15):e114–e114, 2012.

[167] Soumyadeep Nandi, Alexandre Blais, and Ilya Ioshikhes. Identification of cis-regulatory modules in promoters of human genes exploiting mutual

positioning of transcription factors. *Nucleic acids research*, page gkt578, 2013.

[168] Mathieu Blanchette, Alain R Bataille, Xiaoyu Chen, Christian Poitras, Josée Laganière, Céline Lefèbvre, Geneviève Deblois, Vincent Giguère, Vincent Ferretti, Dominique Bergeron, et al. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome research*, 16(5):656–668, 2006.

[169] Outi Hallikas, Kimmo Palin, Natalia Sinjushina, Reetta Rautiainen, Juha Partanen, Esko Ukkonen, and Jussi Taipale. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, 124(1):47–59, 2006.

[170] Cristian O Rohr, R Gonzalo Parra, Patricio Yankilevich, and Carolina Perez-Castro. Insect: In-silico search for co-occurring transcription factors. *Bioinformatics*, 29(22):2852–2858, 2013.

[171] Anna A Nikulova, Alexander V Favorov, Roman A Sutormin, Vsevolod J Makeev, and Andrey A Mironov. Coreclust: identification of the conserved crm grammar together with prediction of gene regulation. *Nucleic acids research*, 40(12):e93–e93, 2012.

[172] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[173] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *ACM Sigmod Record*, volume 29, pages 1–12. ACM, 2000.

[174] Haoyuan Li, Yi Wang, Dong Zhang, Ming Zhang, and Edward Y Chang. Pfp: parallel fp-growth for query recommendation. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 107–114. ACM, 2008.

[175] Le Zhou, Zhiyong Zhong, Jin Chang, Junjie Li, Joshua Zhexue Huang, and Shengzhong Feng. Balanced parallel fp-growth with mapreduce. In *Information Computing and Telecommunications (YC-ICT), 2010 IEEE Youth Conference on*, pages 243–246. IEEE, 2010.

[176] Lingling Deng and Yuansheng Lou. Improvement and research of fp-growth algorithm based on distributed spark. In *2015 International Conference on Cloud Computing and Big Data (CCBD)*, pages 105–108. IEEE, 2015.

[177] Yves Moreau and Léon-Charles Tranchevent. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics*, 13(8):523–536, 2012.

[178] Daniela Börnigen, Léon-Charles Tranchevent, Francisco Bonachela-Capdevila, Koenraad Devriendt, Bart De Moor, Patrick De Causmaecker, and Yves Moreau. An unbiased evaluation of gene prioritization tools. *Bioinformatics*, 28(23):3081–3088, 2012.

[179] Zhichao Liu, Hong Fang, Kelly Reagan, Xiaowei Xu, Donna L Mendrick, William Slikker, and Weida Tong. In silico drug repositioning–what we need to know. *Drug discovery today*, 18(3):110–115, 2013.

[180] Jiao Li, Si Zheng, Bin Chen, Atul J Butte, S Joshua Swamidass, and Zhiyong Lu. A survey of current trends in computational drug repositioning. *Briefings in bioinformatics*, 17(1):2–12, 2016.

[181] Joel T Dudley, Tarangini Deshpande, and Atul J Butte. Exploiting drug–disease relationships for computational drug repositioning. *Briefings in bioinformatics*, page bbr013, 2011.

[182] Carmen Navarro, Victor Martínez, Carlos Cano, and Armando Blanco. Prophtools: General prioritization tools for heterogeneous biological networks. *GigaScience*, Submitted, 2017.

[183] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo D Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(Suppl 1):S7, 2006.

[184] Alexander Lachmann, Federico M Giorgi, Gonzalo Lopez, and Andrea Califano. Aracne-ap: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics*, page btw216, 2016.

[185] Patrick E Meyer, Frederic Lafitte, and Gianluca Bontempi. minet: Ar/bioconductor package for inferring large transcriptional networks using mutual information. *BMC bioinformatics*, 9(1):1, 2008.

[186] Alexandre Irrthum, Louis Wehenkel, Pierre Geurts, et al. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776, 2010.

[187] Gökmen Altay and Frank Emmert-Streib. Inferring the conservative causal core of gene regulatory networks. *BMC Systems Biology*, 4(1):132, 2010.

[188] Fei Xiao, Lin Gao, Yusen Ye, Yuxuan Hu, and Ruijie He. Inferring gene regulatory networks using conditional regulation pattern to guide candidate genes. *PloS one*, 11(5):e0154953, 2016.

[189] Fabio Rinnone, Giovanni Micale, Vincenzo Bonnici, Gary D Bader, Dennis Shasha,
Alfredo Ferro, Alfredo Pulvirenti, and Rosalba Giugno. Netmatchstar: an enhanced cytoscape network querying app. *F1000Research*, 4, 2015.

[190] Fazle E Faisal, Lei Meng, Joseph Crawford, and Tijana Milenković. The post-genomic era of biological network alignment. *EURASIP Journal on Bioinformatics and Systems Biology*, 2015(1):1, 2015.

[191] Marianna Milano, Olga Tymofiyeva, Duan Xu, Christopher Hess, Mario Cannataro, and Pietro H Guzzi. Using network alignment for analysis of connectomes: Experiences from a clinical dataset. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 649–656. ACM, 2016.

[192] Lei Meng, Aaron Striegel, and Tijana Milenkovic. Igloo: Integrating global and local biological network alignment. *arXiv preprint arXiv:1604.06111*, 2016.

[193] Maximilian Malek, Rashid Ibragimov, Mario Albrecht, and Jan Baumbach. Cytogedevo—global alignment of biological networks with cytoscape. *Bioinformatics*, 32(8):1259–1261, 2016.

[194] Lei Meng, Aaron Striegel, and Tijana Milenkovic. Local versus global biological network alignment. *arXiv preprint arXiv:1509.08524*, 2015.

[195] Wooyoung Kim, Martin Diko, and Keith Rawson. Network motif detection: Algorithms, parallel and cloud computing, and related tools. *Tsinghua Science and Technology*, 18(5):469–489, 2013.

[196] Junzhong Ji, Aidong Zhang, Chunnian Liu, Xiaomei Quan, and Zhijun Liu. Survey: Functional module detection from protein-protein interaction networks. *IEEE Transactions on Knowledge and Data Engineering*, 26(2):261–277, 2014.

[197] Jing Chen, Eric E Bardes, Bruce J Aronow, and Anil G Jegga. Toppgene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*, 37(suppl 2):W305–W311, 2009.

[198] Jung Hoon Woo, Yishai Shimoni, Wan Seok Yang, Prem Subramaniam, Archana Iyer, Paola Nicoletti, María Rodríguez Martínez, Gonzalo López, Michela Mattioli, Ronald Realubit, et al. Elucidating compound mechanism of action by network perturbation analysis. *Cell*, 162(2):441–451, 2015.

[199] Xuebing Wu, Rui Jiang, Michael Q Zhang, and Shao Li. Network-based global inference of human disease genes. *Molecular systems biology*, 4(1):189, 2008.

[200] Xiujuan Wang, Natali Gulbahce, and Haiyuan Yu. Network-based methods for human disease gene prediction. *Briefings in functional genomics*, 10(5):280–293, 2011.

[201] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N Robinson. Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, 82(4):949–958, 2008.

[202] Oron Vanunu and Roded Sharan. A propagation-based algorithm for inferring gene-disease assocations. In *German Conference on Bioinformatics*, pages 54–52, 2008.

[203] Saket Navlakha and Carl Kingsford. The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, 26(8):1057–1063, 2010.

[204] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16(16):321–328, 2004.

[205] Matteo Re, Marco Mesiti, and Giorgio Valentini. A fast ranking algorithm for predicting gene functions in biomolecular networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(6):1812–1818, 2012.

[206] Lipika R Pal and John Moult. Genetic basis of common human disease: Insight into the role of missense snps from genome-wide association studies. *Journal of molecular biology*, 427(13):2271–2289, 2015.

[207] Feng Zhang and James R Lupski. Non-coding genetic variants in human disease. *Human molecular genetics*, 24(R1):R102–R110, 2015.

[208] Vladimir Makarov, Tina O'Grady, Guiqing Cai, Jayon Lihm, Joseph D Buxbaum, and Seungtai Yoon. Anntools: a comprehensive and versatile annotation toolkit for genomic variants. *Bioinformatics*, 28(5):724–725, 2012.

[209] Erkhmbayar Jadamba and Miyoung Shin. A snp prioritization method using linkage disequilibrium network for disease association study. *A a*, 2:2, 2012.

[210] George K Acquaah-Mensah, Ronald C Taylor, and Sanjiv V Bhave. Pacap interactions in the mouse brain: implications for behavioral and other disorders. *Gene*, 491(2):224–231, 2012.

[211] Carlos Cruchaga, Sumitra Chakraverty, Kevin Mayo, Francesco LM Vallania, Robi D Mitra, Kelley Faber, Jennifer Williamson, Tom Bird, Ramon Diaz-Arrastia, Tatiana M Foroud, et al. Rare variants in app, psen1 and psen2 increase risk for ad in late-onset alzheimer's disease families. *PloS one*, 7(2):e31039, 2012.

[212] Aris Persidis. The benefits of drug repositioning. *Drug Discov World*, 12:9–12, 2011.

[213] Hermann AM Mucke. Drug repositioning: extracting added value from prior r&d investments. *Insight Pharma Reports*, 2010.

[214] Divya Sardana, Cheng Zhu, Minlu Zhang, Ranga C Gudivada, Lun Yang, and Anil G Jegga. Drug repositioning for orphan diseases. *Briefings in bioinformatics*, 12(4):346–356, 2011.

[215] Ann I Graul, Lisa Sorbera, Patricia Pina, Montse Tell, Elisabet Cruces, Esmeralda Rosa, Mark Stringer, Rosa Castaner, and Laura Revel. The year's new drugs & biologics-2009. *Drug News Perspect*, 23(1):7–36, 2010.

[216] Francesco Iorio, Roberta Bosotti, Emanuela Scacheri, Vincenzo Belcastro, Pratibha Mithbaokar, Rosa Ferriero, Loredana Murino, Roberto Tagliaferri, Nicola Brunetti-Pierri, Antonella Isacchi, et al. Discovery of drug mode of

action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences*, 107(33):14621–14626, 2010.

[217] Dorothea Emig, Alexander Ivliev, Olga Pustovalova, Lee Lancashire, Svetlana Bureeva, Yuri Nikolsky, and Marina Bessarabova. Drug target prediction and repositioning using an integrated network-based approach. *PLoS One*, 8(4):e60618, 2013.

[218] Francesco Iorio, Timothy Rittman, Hong Ge, Michael Menden, and Julio Saez-Rodriguez. Transcriptional data: a new gateway to drug repositioning? *Drug discovery today*, 18(7):350–357, 2013.

[219] Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *science*, 313(5795):1929–1935, 2006.

[220] Dik-Lung Ma, Daniel Shiu-Hin Chan, and Chung-Hang Leung. Drug repositioning by structure-based virtual screening. *Chemical Society Reviews*, 42(5):2130–2141, 2013.

[221] A Srinivas Reddy and Shuxing Zhang. Polypharmacology: drug discovery for the future. *Expert review of clinical pharmacology*, 6(1):41–47, 2013.

[222] Joachim Von Eichborn, Manuela S Murgueitio, Mathias Dunkel, Soeren Koerner, Philip E Bourne, and Robert Preissner. Promiscuous: a database for network-based drug-repositioning. *Nucleic acids research*, 39(suppl 1):D1060–D1066, 2011.

[223] Lun Yang and Pankaj Agarwal. Systematic drug repositioning based on clinical side-effects. *PloS one*, 6(12):e28025, 2011.

[224] Assaf Gottlieb, Gideon Y Stein, Eytan Ruppin, and Roded Sharan. Predict: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology*, 7(1):496, 2011.

[225] Annie P Chiang and Atul J Butte. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clinical pharmacology and therapeutics*, 86(5):507, 2009.

[226] Michael G Walker, Wayne Volkmuth, Tod M Klingler, et al. Pharmaceutical target discovery using guilt-by-association: schizophrenia and parkinson's disease genes. In *ISMB*, pages 282–286, 1999.

[227] John Quackenbush. Microarrays–guilt by association. *Science*, 302(5643):240–241, 2003.

[228] L Aravind. Guilt by association: contextual information in genome analysis. *Genome Research*, 10(8):1074–1077, 2000.

[229] David C Swinney and Jason Anthony. How were new medicines discovered? *Nature reviews Drug discovery*, 10(7):507–519, 2011.

[230] Samiul Hasan, Bhushan K Bonde, Natalie S Buchan, and Matthew D Hall. Network analysis has diverse roles in drug discovery. *Drug discovery today*, 17(15):869–874, 2012.

[231] Sachin Mathur and Deendayal Dinakarpandian. Drug repositioning using disease associated biological processes and network analysis of drug targets. In *AMIA Annu Symp Proc*, volume 2011, pages 305–311, 2011.

[232] José L Medina-Franco, Marc A Giulianotti, Gregory S Welmaker, and Richard A Houghten. Shifting from the single to the multitarget paradigm in drug discovery. *Drug discovery today*, 18(9):495–501, 2013.

[233] Leticia Diaz-Beltran, Carlos Cano, Dennis P Wall, and Francisco J Esteban. Systems biology as a comparative approach to understand complex gene expression in neurological diseases. *Behavioral Sciences*, 3(2):253–272, 2013.

[234] Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, et al. Drugbank

3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, 39(suppl 1):D1035–D1041, 2011.

[235] Dekang Lin. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304. Citeseer, 1998.

[236] Andrew Chatr-Aryamontri, Bobby-Joe Breitkreutz, Sven Heinicke, Lorrie Boucher, Andrew Winter, Chris Stark, Julie Nixon, Lindsay Ramage, Nadine Kolas, Lara O'Donnell, et al. The biogrid interaction database: 2013 update. *Nucleic acids research*, 41(D1):D816–D823, 2013.

[237] Kai Peng, Wei Xu, Jianyong Zheng, Kegui Huang, Huisong Wang, Jiansong Tong, Zhifeng Lin, Jun Liu, Wenqing Cheng, Dong Fu, et al. The disease and gene annotations (dga): an annotation resource for human disease. *Nucleic acids research*, page gks1244, 2012.

[238] Mark H Zweig and Gregory Campbell. Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561–577, 1993.

[239] Allen Krantz. Diversification of the drug discovery process. *Nature biotechnology*, 16(13):1294–1294, 1998.

[240] Brian G Feagan, James Rochon, Richard N Fedorak, E Jan Irvine, Gary Wild, Lloyd Sutherland, A Hillary Steinhart, Gordon R Greenberg, Richard Gillies, Marybeth Hopkins, et al. Methotrexate for the treatment of crohn's disease. *New England Journal of Medicine*, 332(5):292–297, 1995.

[241] Vivian A Fonseca, Yehuda Handelsman, and Bart Staels. Colesevelam lowers glucose and lipid levels in type 2 diabetes: the clinical evidence. *Diabetes, Obesity and Metabolism*, 12(5):384–392, 2010.

[242] Mark H Pollack, John Matthews, and Erin L Scott. Gabapentin as a potential treatment for anxiety disorders. *American Journal of Psychiatry*, 155(7):992–993, 1998.

[243] Daniel P Silver, Andrea L Richardson, Aron C Eklund, Zhigang C Wang, Zoltan Szallasi, Qiyuan Li, Nicolai Juul, Chee-Onn Leong, Diana Calogrias, Ayodele Buraimoh, et al. Efficacy of neoadjuvant cisplatin in triple-negative breast cancer. *Journal of Clinical Oncology*, 28(7):1145–1153, 2010.

[244] D Aarsland, K Laake, JP Larsen, and C Janvin. Donepezil for cognitive impairment in parkinson's disease: a randomised controlled study. *Journal of Neurology, Neurosurgery & Psychiatry*, 72(6):708–712, 2002.

[245] Frederick M Jacobsen. Risperidone in the treatment of affective illness and obsessive-compulsive disorder. *Journal of Clinical Psychiatry*, 1995.

[246] HP Van Bever and WJ Stevens. Nedocromil sodium cream in the treatment of atopic dermatitis. *European journal of pediatrics*, 149(1):74–74, 1989.

[247] HN Williams. Multi-centre clinical trial of nedocromil sodium in reversible obstructive airways disease in adults: a general practitioner collaborative study. *Current medical research and opinion*, 11(7):417–426, 1989.

[248] Richard L Mabry. Topical pharmacotherapy for allergic rhinitis: nedocromil. *American journal of otolaryngology*, 14(6):379–381, 1993.

[249] CH Chesnut. Tiludronate: development as an osteoporosis therapy. *Bone*, 17(5):S517–S519, 1995.

[250] MR McClung, CKP Tou, NH Goldstein, and C Picot. Tiludronate therapy for paget's disease of bone. *Bone*, 17(5):S493–S496, 1995.

[251] Maxim Moreau, Pascale Rialland, Jean-Pierre Pelletier, Johanne Martel-Pelletier, Daniel Lajeunesse, Christielle Boileau, Judith Caron, Diane Frank, Bertrand Lussier, Jerome RE del Castillo, et al. Tiludronate treatment improves structural changes and symptoms of osteoarthritis in the canine anterior cruciate ligament model. *Arthritis research & therapy*, 13(3):1, 2011.

[252] Jean Claude Dumon, A Magritte, and Jean-Jacques Body. Efficacy and safety of the bisphosphonate tiludronate for the treatment of tumor-associated hypercalcemia. *Bone and mineral*, 15(3):257–266, 1991.

[253] ARMIN E Heufelder, BJOERN E Wenzel, and Rebecca Sue Bahn. Methimazole and propylthiouracil inhibit the oxygen free radical-induced expression of a 72 kilodalton heat shock protein in graves' retroocular fibroblasts. *The Journal of Clinical Endocrinology & Metabolism*, 74(4):737–742, 1992.

[254] SJ Mandel, GA Brent, and PR Larsen. Levothyroxine therapy in patients with thyroid disease. *The Endocrinologist*, 4(2):152, 1994.

[255] Eva Marie T Erfurth, Ulla-Brltt C Ericsson, Karln Egervall, and Stefan R Lethagen. Effect of acute desmopressin and of long-term thyroxine replacement on haemostasis in hypothyroidism. *Clinical endocrinology*, 42(4):373–378, 1995.

[256] W Ronald Skowsky and Thomas A Kikuchi. The role of vasopressin in the impaired water excretion of myxedema. *The American journal of medicine*, 64(4):613–621, 1978.

[257] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263–266, 2008.

[258] Carmen Navarro, Francisco J Lopez, Carlos Cano, Fernando Garcia-Alcalde, and Armando Blanco. Cisminer: Genome-wide in-silico cis-regulatory module prediction by fuzzy itemset mining. *PloS one*, 9(9):e108065, 2014.

[259] Hélène Touzet and Jean-Stéphane Varré. Efficient and accurate p-value computation for position weight matrices. *Algorithms for Molecular Biology*, 2(1):1, 2007.

[260] Thomas Manke, Helge G Roider, and Martin Vingron. Statistical modeling of transcription factor binding affinities predicts regulatory interactions. *PLoS Comput Biol*, 4(3):e1000039, 2008.

[261] Alexander E Kel, Ellen Gößling, Ingmar Reuter, Evgeny Cheremushkin, Olga V Kel-Margoulis, and Edgar Wingender. Matchtm: a tool for searching transcription factor binding sites in dna sequences. *Nucleic acids research*, 31(13):3576–3579, 2003.

[262] Maria Charalambous, Pedro Trancoso, and Alexandros Stamatakis. Initial experiences porting a bioinformatics application to a graphics processor. In *Panhellenic Conference on Informatics*, pages 415–425. Springer, 2005.

[263] Michael C Schatz, Cole Trapnell, Arthur L Delcher, and Amitabh Varshney. High-throughput sequence alignment using graphics processing units. *BMC bioinformatics*, 8(1):474, 2007.

[264] Cole Trapnell and Michael C Schatz. Optimizing data intensive gpgpu computations for dna sequence alignment. *Parallel computing*, 35(8):429–440, 2009.

[265] Weiguo Liu, Bertil Schmidt, Gerrit Voss, and Wolfgang Müller-Wittig. Gpu-clustalw: Using graphics hardware to accelerate multiple sequence alignment. In *International Conference on High-Performance Computing*, pages 363–374. Springer, 2006.

[266] Svetlin A Manavski and Giorgio Valle. Cuda compatible gpu cards as efficient hardware accelerators for smith-waterman sequence alignment. *BMC bioinformatics*, 9(2):1, 2008.

[267] John Nickolls, Ian Buck, Michael Garland, and Kevin Skadron. Scalable parallel programming with cuda. *Queue*, 6(2):40–53, 2008.

[268] Otto G Berg and Peter H von Hippel. Selection of dna binding sites by regulatory proteins. *Trends in biochemical sciences*, 13(6):207–211, 1988.

[269] Xiaoyan Ma, Daphne Ezer, Carmen Navarro, and Boris Adryan. Reliable scaling of position weight matrices for binding strength comparisons between transcription factors. *BMC bioinformatics*, 16(1):1, 2015.

[270] Nicolae Radu Zabet and Boris Adryan. Estimating binding properties of transcription factors from genome-wide binding profiles. *Nucleic acids research*, page gku1269, 2014.

[271] Zeba Wunderlich and Leonid A Mirny. Different gene regulation strategies revealed by analysis of binding motifs. *Trends in genetics*, 25(10):434–440, 2009.

[272] Jason S Leith, Anahita Tafvizi, Fang Huang, William E Uspal, Patrick S Doyle, Alan R Fersht, Leonid A Mirny, and Antoine M van Oijen. Sequence-dependent sliding kinetics of p53. *Proceedings of the National Academy of Sciences*, 109(41):16552–16557, 2012.

[273] Sebastian J Maerkl and Stephen R Quake. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809):233–237, 2007.

[274] Adam Woolfe, Martin Goodson, Debbie K Goode, Phil Snell, Gayle K McEwen, Tanya Vavouri, Sarah F Smith, Phil North, Heather Callaway, Krys Kelly, et al. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol*, 3(1):e7, 2004.

[275] Tanya Vavouri and Greg Elgar. Prediction of cis-regulatory elements using binding site matrices—the successes, the failures and the reasons for both. *Current opinion in genetics & development*, 15(4):395–402, 2005.

[276] Emmanouil T Dermitzakis and Andrew G Clark. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Molecular biology and evolution*, 19(7):1114–1121, 2002.

[277] Eldon Emberly, Nikolaus Rajewsky, and Eric D Siggia. Conservation of regulatory elements between two species of drosophila. *BMC bioinformatics*, 4(1):1, 2003.

[278] Josef Laimer, Clemens J Zuzan, Tobias Ehrenberger, Monika Freudenberger, Simone Gschwandtner, Carina Lebherz, and Peter Lackner. D-light on promoters: a client-server system for the analysis and visualization of cis-regulatory elements. *BMC bioinformatics*, 14(1):1, 2013.

[279] Nati Ha, Maria Polychronidou, and Ingrid Lohmann. Cops: detecting co-occurrence and spatial arrangement of transcription factor binding motifs in genome-wide datasets. *PloS one*, 7(12):e52055, 2012.

[280] Anil G Jegga, Shawn P Sherwood, James W Carman, Andrew T Pinski, Jerry L Phillips, John P Pestian, and Bruce J Aronow. Detection and visualization of compositionally similar cis-regulatory element clusters in orthologous and coordinately controlled genes. *Genome research*, 12(9):1408–1417, 2002.

[281] Mauro Leoncini, Manuela Montangero, Marco Pellegrini, and Karina Panucia Tillán. Cmf: a combinatorial tool to find composite motifs. In *International Conference on Learning and Intelligent Optimization*, pages 196–208. Springer, 2013.

[282] Igor V Deyneko, Alexander E Kel, Olga V Kel-Margoulis, Elena V Deineko, Edgar Wingender, and Siegfried Weiss. Matrixcatch-a novel tool for the recognition of composite regulatory elements in promoters. *BMC bioinformatics*, 14(1):1, 2013.

[283] Francisco J Lopez, Armando Blanco, Fernando Garcia, Carlos Cano, and Antonio Marin. Fuzzy association rules for biological data analysis: a case study on yeast. *BMC bioinformatics*, 9(1):1, 2008.

[284] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.

[285] Aaron Ceglar and John F Roddick. Association mining. *ACM Computing Surveys (CSUR)*, 38(2):5, 2006.

[286] Stefan Naulaerts, Pieter Meysman, Wout Bittremieux, Trung Nghia Vu, Wim Vanden Berghe, Bart Goethals, and Kris Laukens. A primer to frequent itemset mining for bioinformatics. *Briefings in bioinformatics*, 16(2):216–231, 2015.

[287] Miguel Delgado, Nicolás Marín, Daniel Sánchez, and M-A Vila. Fuzzy association rules: general model and applications. *IEEE Transactions on fuzzy systems*, 11(2):214–225, 2003.

[288] Arianna Gallo, Tijl De Bie, and Nello Cristianini. Mini: Mining informative non-redundant itemsets. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 438–445. Springer, 2007.

[289] Xochitl C Morgan, Shulin Ni, Daniel P Miranker, and Vishwanath R Iyer. Predicting combinatorial binding of transcription factors to regulatory elements in the human genome by association rule mining. *BMC bioinformatics*, 8(1):1, 2007.

[290] Christopher T Harbison, D Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D Macisaac, Timothy W Danford, Nancy M Hannett, Jean-Bosco Tagne, David B Reynolds, Jane Yoo, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.

[291] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian Von Mering, et al. String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(D1):D808–D815, 2013.

[292] Christian Koch, Thomas Moll, Manfred Neuberg, Horst Ahorn, and Kim Nasmyth. A role for the transcription factors mbp1 and swi4 in progression from g1 to s phase. *Science*, 261(5128):1551–1557, 1993.

[293] Julia van der Felden, Sarah Weisser, Stefan Brückner, Peter Lenz, and Hans-Ulrich Mösch. The transcription factors tec1 and ste12 interact with coregulators msa1 and msa2 to activate adhesion and multicellular development. *Molecular and cellular biology*, 34(12):2283–2293, 2014.

[294] Stephan B Schawalder, Mehdi Kabani, Isabelle Howald, Urmila Choudhury, Michel Werner, and David Shore. Growth-regulated recruitment of the essential yeast ribosomal protein gene activator ifh1. *Nature*, 432(7020):1058–1061, 2004.

[295] J Michael Cherry, Eurie L Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T Chan, Karen R Christie, Maria C Costanzo, Selina S Dwight, Stacia R Engel, et al. Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic acids research*, page gkr1029, 2011.

[296] Anthony Mathelier, Xiaobei Zhao, Allen W Zhang, François Parcy, Rebecca Worsley-Hunt, David J Arenillas, Sorana Buchman, Chih-yu Chen, Alice Chou, Hans Ienasescu, et al. Jaspar 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic acids research*, page gkt997, 2013.

[297] Gerald Z Hertz and Gary D. Stormo. Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7):563–577, 1999.

[298] Heladia Salgado, Socorro Gama-Castro, Agustino Martínez-Antonio, Edgar Díaz-Peredo, Fabiola Sánchez-Solano, Martín Peralta-Gil, Delfino Garcia-Alonso, Verónica Jiménez-Jacinto, Alberto Santos-Zavaleta, César Bonavides-Martínez, et al. Regulondb (version 4.0): transcriptional

regulation, operon organization and growth conditions in escherichia coli k-12. *Nucleic Acids Research*, 32(suppl 1):D303–D306, 2004.

[299] Elizabeth A Jones and Richard A Flavell. Distal enhancer elements transcribe intergenic rna in the il-10 family gene cluster. *The Journal of Immunology*, 175(11):7437–7446, 2005.

[300] Jean-Valery Turatsinze, Morgane Thomas-Chollier, Matthieu Defrance, and Jacques van Helden. Using rsat to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature protocols*, 3(10):1578–1588, 2008.

[301] Taewoo Ryu, Younghoon Kim, Dae-Won Kim, Doheon Lee, et al. Computational identification of combinatorial regulation and transcription factor binding sites. *Biotechnology and bioengineering*, 97(6):1594–1602, 2007.

[302] Song Chou, Shelley Lane, and Haoping Liu. Regulation of mating and filamentation genes by two distinct ste12 complexes in saccharomyces cerevisiae. *Molecular and Cellular Biology*, 26(13):4794–4805, 2006.

[303] Wan-Sheng Lo and Anne M Dranginis. The cell surface flocculin flo11 is required for pseudohyphae formation and invasion by saccharomyces cerevisiae. *Molecular biology of the cell*, 9(1):161–171, 1998.

[304] Tae Soo Kim, Sung Bae Lee, and Hyen Sam Kang. Glucose repression of sta1 expression is mediated by the nrg1 and sfl1 repressors and the srb8-11 complex. *Molecular and cellular biology*, 24(17):7695–7706, 2004.

[305] Qun Yu, Runxiang Qiu, Travis B Foland, Dan Griesen, Carl S Galloway, Ya-Hui Chiu, Joseph Sandmeier, James R Broach, and Xin Bi. Rap1p and other transcriptional regulators can function in defining distinct domains of gene expression. *Nucleic acids research*, 31(4):1224–1233, 2003.

[306] Michal Ronen and David Botstein. Transcriptional response of steady-state yeast cultures to transient perturbations in carbon source. *Proceedings of the National Academy of Sciences of the United States of America*, 103(2):389–394, 2006.

[307] Hans-Joachim Schüller. Transcriptional control of nonfermentative metabolism in the yeast saccharomyces cerevisiae. *Current genetics*, 43(3):139–160, 2003.

[308] Tony R Hazbun and Stanley Fields. A genome-wide screen for site-specific dna-binding proteins. *Molecular & Cellular Proteomics*, 1(7):538–543, 2002.

[309] Pedro M Santos, Tânia Simões, and Isabel Sá-Correia. Insights into yeast adaptive response to the agricultural fungicide mancozeb: a toxicoproteomics approach. *Proteomics*, 9(3):657–670, 2009.

[310] Miguel Cacho Teixeira, Alexandra Ramos Fernandes, Nuno Pereira Mira, Jörg Dieter Becker, and Isabel Sá-Correia. Early transcriptional response of saccharomyces cerevisiae to stress imposed by the herbicide 2, 4-dichlorophenoxyacetic acid. *FEMS yeast research*, 6(2):230–248, 2006.

[311] Elisabetta Cameroni, Nicolas Hulo, Johnny Roosen, Joris Winderickx, and Claudio De Virgilio. The novel yeast pas kinase rim15 orchestrates g0-associated antioxidant defense mechanisms. *Cell cycle*, 3(4):460–466, 2004.

[312] Linda L Lutfiyya, Vishwanath R Iyer, Joe DeRisi, Michael J DeVit, Patrick O Brown, and Mark Johnston. Characterization of three related glucose repressors and genes they regulate in saccharomyces cerevisiae. *Genetics*, 150(4):1377–1391, 1998.

[313] Susan Tweedie, Michael Ashburner, Kathleen Falls, Paul Leyland, Peter McQuilton, Steven Marygold, Gillian Millburn, David Osumi-Sutherland, Andrew Schroeder, Ruth Seal, et al. Flybase: enhancing drosophila gene ontology annotations. *Nucleic acids research*, 37(suppl 1):D555–D559, 2009.

[314] Miguel Delgado, N Manín, MJ Martin-Bautista, D Sanchez, and M-A Vila. Mining fuzzy association rules: an overview. In *Soft Computing for Information Processing and Analysis*, pages 351–373. Springer, 2005.

[315] Chris P Ponting, Peter L Oliver, and Wolf Reik. Evolution and functions of long noncoding rnas. *Cell*, 136(4):629–641, 2009.

[316] Carmen Navarro, Carlos Cano, Marta Cuadros, and Armando Blanco. Computationally assessing rna-protein binding affinity in long non-coding rnas. *IWBBIO 2015*, page 63, 2015.

[317] Carmen Navarro, Carlos Cano, Marta Cuadros, Antonio Herrera-Merchan, Miguel Molina, and Armando Blanco. A mechanistic study of lncrna fendrr regulation of foxf1 lung cancer tumor supressor. In *International Conference on Bioinformatics and Biomedical Engineering*, pages 781–789. Springer, 2016.

[318] Philipp Kapranov, Jill Cheng, Sujit Dike, David A Nix, Radharani Duttagupta, Aarron T Willingham, Peter F Stadler, Jana Hertel, Jörg Hackermüller, Ivo L Hofacker, et al. Rna maps reveal new rna classes and a possible function for pervasive transcription. *Science*, 316(5830):1484–1488, 2007.

[319] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

[320] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774, 2012.

[321] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. Tophat2: accurate alignment of transcriptomes in the

presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):1, 2013.

[322] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[323] Szymon M Kiełbasa, Raymond Wan, Kengo Sato, Paul Horton, and Martin C Frith. Adaptive seeds tame genomic sequence comparison. *Genome research*, 21(3):487–493, 2011.

[324] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3):562–578, 2012.

[325] Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nature biotechnology*, 28(5):503–510, 2010.

[326] Michael F Lin, Irwin Jungreis, and Manolis Kellis. Phylocsf: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, 27(13):i275–i282, 2011.

[327] Stefan Washietl, Sven Findeiß, Stephan A Müller, Stefan Kalkhof, Martin von Bergen, Ivo L Hofacker, Peter F Stadler, and Nick Goldman. Rnacode: robust discrimination of coding and noncoding regions in comparative sequence data. *Rna*, 17(4):578–594, 2011.

[328] Liguo Wang, Hyun Jung Park, Surendra Dasari, Shengqin Wang, Jean-Pierre Kocher, and Wei Li. Cpat: Coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic acids research*, 41(6):e74–e74, 2013.

[329] Lei Kong, Yong Zhang, Zhi-Qiang Ye, Xiao-Qiao Liu, Shu-Qi Zhao, Liping Wei, and Ge Gao. Cpc: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic acids research*, 35(suppl 2):W345–W349, 2007.

[330] Liang Sun, Haitao Luo, Dechao Bu, Guoguang Zhao, Kuntao Yu, Changhai Zhang, Yuanning Liu, Runsheng Chen, and Yi Zhao. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic acids research*, page gkt646, 2013.

[331] John L Rinn and Jernej Ule. 'oming in on rna–protein interactions. *Genome biology*, 15(1):1, 2014.

[332] Julian Konig, Kathi Zarnack, Gregor Rot, Tomaz Curk, Melis Kayikci, Blaz Zupan, Daniel J Turner, Nicholas M Luscombe, and Jernej Ule. iclip-transcriptome-wide mapping of protein-rna interactions with individual nucleotide resolution. *Journal of visualized experiments: JoVE*, (50), 2011.

[333] Grzegorz Kudla, Sander Granneman, Daniela Hahn, Jean D Beggs, and David Tollervey. Cross-linking, ligation, and sequencing of hybrids reveals rna–rna interactions in yeast. *Proceedings of the National Academy of Sciences*, 108(24):10010–10015, 2011.

[334] Robert Darnell. Clip (cross-linking and immunoprecipitation) identification of rnas bound by a specific protein. *Cold Spring Harbor Protocols*, 2012(11):pdb–prot072132, 2012.

[335] Charles Danan, Sudhir Manickavel, and Markus Hafner. Par-clip: A method for transcriptome-wide identification of rna binding protein interaction sites. *Post-Transcriptional Gene Regulation*, pages 153–173, 2016.

[336] Matthew D Simon, Charlotte I Wang, Peter V Kharchenko, Jason A West, Brad A Chapman, Artyom A Alekseyenko, Mark L Borowsky, Mitzi I Kuroda, and Robert E Kingston. The genomic binding sites of a noncoding rna.

Proceedings of the National Academy of Sciences, 108(51):20497–20502, 2011.

[337] Debashish Ray, Hilal Kazan, Kate B Cook, Matthew T Weirauch, Hamed S Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, et al. A compendium of rna-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, 2013.

[338] Avinash Achar and Pål Sætrom. Rna motif discovery: a computational overview. *Biology direct*, 10(1):1, 2015.

[339] Michael Hiller, Rainer Pudimat, Anke Busch, and Rolf Backofen. Using rna secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic acids research*, 34(17):e117–e117, 2006.

[340] Hilal Kazan, Debashish Ray, Esther T Chan, Timothy R Hughes, and Quaid Morris. Rnacontext: a new method for learning the sequence and structure binding preferences of rna-binding proteins. *PLoS Comput Biol*, 6(7):e1000832, 2010.

[341] Daniel Maticzka, Sita J Lange, Fabrizio Costa, and Rolf Backofen. Graphprot: modeling binding preferences of rna-binding proteins. *Genome biology*, 15(1):1, 2014.

[342] Ronny Lorenz, Stephan H Bernhart, Christian Hoener Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for Molecular Biology*, 6(1):1, 2011.

[343] Peter Kerpedjiev, Stefan Hammer, and Ivo L Hofacker. Forna (force-directed rna): simple and effective online rna secondary structure diagrams. *Bioinformatics*, page btv372, 2015.

[344] Rajnish A Gupta, Nilay Shah, Kevin C Wang, Jeewon Kim, Hugo M Horlings, David J Wong, Miao-Chih Tsai, Tiffany Hung, Pedram Argani, John L Rinn, et al. Long non-coding rna hotair reprograms chromatin state to promote cancer metastasis. *Nature*, 464(7291):1071–1076, 2010.

[345] Eric Pasmant, Ingrid Laurendeau, Delphine Héron, Michel Vidaud, Dominique Vidaud, and Ivan Bièche. Characterization of a germ-line deletion, including the entire ink4/arf locus, in a melanoma-neural system tumor family: identification of anril, an antisense noncoding rna whose expression coclusters with arf. *Cancer research*, 67(8):3963–3969, 2007.

[346] Xiao Li, Hilal Kazan, Howard D Lipshitz, and Quaid D Morris. Finding the target sites of rna-binding proteins. *Wiley Interdisciplinary Reviews: RNA*, 5(1):111–130, 2014.

[347] Michael Zuker and Patrick Stiegler. Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic acids research*, 9(1):133–148, 1981.

[348] David H Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of rna secondary structure. *Journal of molecular biology*, 288(5):911–940, 1999.

[349] Przemyslaw Szafranski, Avinash V Dharmadhikari, Erwin Brosens, Priyatansh Gurha, Katarzyna E Kołodziejska, Ou Zhishuo, Piotr Dittwald, Tadeusz Majewski, K Naga Mohan, Bo Chen, et al. Small noncoding differentially methylated copy-number variants, including lncrna genes, cause a lethal lung developmental disorder. *Genome research*, 23(1):23–33, 2013.

[350] Phillip Grote and Bernhard G Herrmann. The long non-coding rna fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis. *RNA biology*, 10(10):1579–1585, 2013.

[351] Sonja Hanzelmann, Chao-Chung Kuo, Marie Kalwa, Wolfgang Wagner, and Ivan G Costa. Triplex domain finder: Detection of triple helix binding domains in long non-coding rnas. *bioRxiv*, page 020297, 2016.

[352] Nara Lee and Joan A Steitz. Noncoding rna-guided recruitment of transcription factors: A prevalent but undocumented mechanism? *BioEssays*, 37(9):936–941, 2015.

[353] Ivan V Kulakovskiy, Yulia A Medvedeva, Ulf Schaefer, Artem S Kasianov, Ilya E Vorontsov, Vladimir B Bajic, and Vsevolod J Makeev. Hocomoco: a comprehensive collection of human transcription factor binding sites models. *Nucleic acids research*, 41(D1):D195–D202, 2013.

[354] Fabian A Buske, John S Mattick, and Timothy L Bailey. Potential in vivo roles of nucleic acid triple-helices. *RNA biology*, 8(3):427–439, 2011.

[355] Fabian A Buske, Denis C Bauer, John S Mattick, and Timothy L Bailey. Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. *Genome research*, 22(7):1372–1381, 2012.

[356] Rory Johnson and Roderic Guigó. The ridl hypothesis: transposable elements as functional domains of long noncoding rnas. *Rna*, 20(7):959–976, 2014.

[357] Charles E Grant, Timothy L Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[358] Kate B Cook, Hilal Kazan, Khalid Zuberi, Quaid Morris, and Timothy R Hughes. Rbpdb: a database of rna-binding specificities. *Nucleic acids research*, 39(suppl 1):D301–D308, 2011.

[359] Moritz Bauer, Johanna Trupke, and Leonie Ringrose. The quest for mammalian polycomb response elements: are we there yet? *Chromosoma*, pages 1–26, 2015.

[360] Kate R Rosenbloom, Joel Armstrong, Galt P Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, Maximilian Haeussler, et al. The ucsc genome browser database: 2015 update. *Nucleic acids research*, 43(D1):D670–D681, 2015.

[361] Sascha Steinbiss, Gordon Gremme, Christin Schärfer, Malte Mader, and Stefan Kurtz. Annotationsketch: a genome annotation drawing library. *Bioinformatics*, 25(4):533–534, 2009.

[362] Bradley M Lunde, Claire Moore, and Gabriele Varani. Rna-binding proteins: modular design for efficient function. *Nature reviews Molecular cell biology*, 8(6):479–490, 2007.

[363] Tina Glisovic, Jennifer L Bachorik, Jeongsik Yong, and Gideon Dreyfuss. Rna-binding proteins and post-transcriptional gene regulation. *FEBS letters*, 582(14):1977–1986, 2008.

[364] Seyed Yahya Anvar, Lusine Khachatryan, Martijn Vermaat, Michiel van Galen, Irina Pulyakhina, Yavuz Ariyurek, Ken Kraaijeveld, Johan T den Dunnen, Peter de Knijff, Peter AC't Hoen, et al. Determining the quality and complexity of next-generation sequencing data without a reference genome. *Genome biology*, 15(12):1, 2014.

[365] Florian Plaza Onate, Jean-Michel Batto, Catherine Juste, Jehane Fadlallah, Cyrielle Fougeroux, Doriane Gouas, Nicolas Pons, Sean Kennedy, Florence Levenez, Joel Dore, et al. Quality control of microbiota metagenomics by k-mer analysis. *BMC genomics*, 16(1):1, 2015.

[366] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[367] Sandrine Dudoit, Yee Hwa Yang, Matthew J Callow, and Terence P Speed. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica sinica*, pages 111–139, 2002.

[368] Aaron R Quinlan. Bedtools: the swiss-army tool for genome feature analysis. *Current protocols in bioinformatics*, pages 11–12, 2014.

[369] Jayita Guhaniyogi and Gary Brewer. Regulation of mrna stability in mammalian cells. *Gene*, 265(1):11–23, 2001.

[370] Harpreet Kaur Saini, Sam Griffiths-Jones, and Anton James Enright. Genomic analysis of human microrna transcripts. *Proceedings of the National Academy of Sciences*, 104(45):17719–17724, 2007.

[371] Emmanuel Beaudoing, Susan Freier, Jacqueline R Wyatt, Jean-Michel Claverie, and Daniel Gautheret. Patterns of variant polyadenylation signal usage in human genes. *Genome research*, 10(7):1001–1010, 2000.

[372] William G Fairbrother, Gene W Yeo, Rufang Yeh, Paul Goldstein, Matthew Mawson, Phillip A Sharp, and Christopher B Burge. Rescue-ese identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic acids research*, 32(suppl 2):W187–W190, 2004.

[373] Zefeng Wang and Christopher B Burge. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *Rna*, 14(5):802–813, 2008.

[374] Benjamin P Lewis, I-hung Shih, Matthew W Jones-Rhoades, David P Bartel, and Christopher B Burge. Prediction of mammalian microrna targets. *Cell*, 115(7):787–798, 2003.

[375] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Fast random walk with restart and its applications. 2006.

[376] Xiao-Yong Li, Sean Thomas, Peter J Sabo, Michael B Eisen, John A Stamatoyannopoulos, and Mark D Biggin. The role of chromatin accessibility in directing the widespread, overlapping patterns of drosophila transcription factor binding. *Genome biology*, 12(4):1, 2011.

[377] Stephen B Baylin. Dna methylation and gene silencing in cancer. *Nature clinical practice Oncology*, 2:S4–S11, 2005.

[378] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, et al. Human dna methylomes at base resolution show widespread epigenomic differences. *nature*, 462(7271):315–322, 2009.