Universidad de Granada

E. T. S. de Ingenierías Informática y de Telecomunicación

Departamento de Ciencias de la Computación e I. A.

Integración y análisis de datos genómicos mediante patrones difusos: Aplicaciones

Tesis Doctoral

Francisco Javier López Domingo

Directores: Dr. Armando Blanco Morón y
Dr. Antonio Marín Rodríguez

Editor: Editorial de la Universidad de Granada
Autor: Francisco Javier López Domingo
D.L.: GR-2382-2010
ISBN: 978-84-693-1310-7

Universidad de Granada

E. T. S. de Ingenierías Informática y de Telecomunicación

Departamento de Ciencias de la Computación e I. A.

Fuzzy pattern mining for the integration and analysis of genomic information. Applications

PhD Dissertation

Francisco Javier López Domingo

Supervisors: Armando Blanco Morón &
Antonio Marín Rodríguez

La memoria “Integración y análisis de datos genómicos mediante patrones difusos: Aplicaciones”, que presenta D. Francisco Javier López Domingo para optar al grado de Doctor en Informática, ha sido realizada en el departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada, bajo la dirección del Dr. D. Armando Blanco Morón, Profesor Titular del departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada y del Dr. D. Antonio Marín Rodríguez, Catedrático del departamento de Genética de la Universidad de Sevilla.

Granada, Enero de 2010.

Armando Blanco Morón

Antonio Marín Rodríguez

Francisco Javier López Domingo

Índice general

Índice de figuras

Índice de tablas

Resumen

La secuenciación de los genomas de diversas especies, así como el desarrollo de nuevas tecnologías genómicas, han dado lugar a una enorme cantidad de datos biológicos que se encuentran dispersos en muchas bases de datos. La integración y el análisis de estos datos es necesaria para alcanzar un mayor entendimiento del funcionamiento celular. Así, las reglas de asociación son una herramienta muy útil en este campo, dada su eficiencia al manejar grandes conjuntos de datos, su capacidad de tratar información heterogénea y la fácil interpretación de los resultados obtenidos con esta técnica. Además, los datos biológicos tienden a ser imprecisos y ruidosos. Existen técnicas computacionales, como las técnicas difusas, que han demostrado ser especialmente apropiadas para modelar este tipo de datos. En este trabajo se propone una metodología basada en un algoritmo de extracción de reglas de asociación difusas para extraer conocimiento de datos biológicos. Dicha metodología se aplica sobre una base de datos en la que se integró información estructural y funcional del genoma de la levadura. Los buenos resultados obtenidos de este estudio permitieron abordar un trabajo más ambicioso: analizar las características genómicas del cáncer de mama. Se integró en un conjunto de datos información de los principales factores de pronóstico en el cáncer de mama con valores de expresión del genoma completo. El descubrimiento de vínculos entre estos dos tipos de datos puede dar lugar a nuevos marcadores del cáncer de mama, lo que a su vez ayudará a mejorar los tratamientos que se aplican a pacientes con un pronóstico poco claro.

La última parte de esta memoria se dedica al estudio de los mecanismos de regulación genética. En las células eucariotas, las regiones de control de

los genes están formadas por su promotor y por una serie de elementos reguladores que pueden encontrarse lejos del gen. Combinaciones de proteínas reguladoras (los factores de transcripción), se unen de forma coordinada a dichas secuencias (sitios de unión de factores de transcripción ó TFBSs) y producen los patrones de expresión adecuados. Aprovechando la capacidad de las técnicas difusas para manejar la imprecisión, inherente a la información acerca de secuencias reguladoras, se presenta un nuevo enfoque para estudiar las coocurrencias significativas de TFBSs cercanos en el genoma de la levadura. La metodología se basa en el uso de un algoritmo de extracción de itemsets frecuentes y difusos, y solventa algunas de las limitaciones de propuestas previas. Los resultados confirman su buen funcionamiento y permiten plantear su aplicación a genomas más complejos en trabajos futuros.

Abstract

Last years' mapping of diverse genomes has generated huge amounts of biological data which are currently dispersed through many databases. Integration of the information available in the various databaes is required in order to achieve higher understanding levels. In this context, association rules appear as a powerful tool to analyze biological data, due to their ability to manage large datasets, their capacity to treat heterogeneous information and the intuitive interpretation of the results obtained with this technique. Likewise, biological data are often imprecise and noisy. Advanced computational methodologies, such as fuzzy techniques, have been developed and have shown to be specially suitable to model this type of data. Hence, in this work it is proposed a novel fuzzy methodology based on a fuzzy association rule mining method for biological knowledge extraction. This methodology is applied over a yeast genome dataset containing heterogeneous information regarding structural and functional genome features. The good results obtained from this study enabled us to carry out a more ambitious work: analyzing the genomic peculiarities of breast cancer. Thus, information from the main prognostic factors in breast cancer is integrated with whole-genome microarray data to study the potential associations between these two types of data. Unveiling links between prognostic factors and gene expression values may result in new biomarkers and will help to select the most suitable treatment for patients with an unclear prognostic.

The last part of this work is devoted to the analysis of gene regulatory mechanisms. Eucaryotic gene control regions consist of a promoter plus regulatory DNA sequences which may appear distant from the gene promoter. Regulatory proteins (called transcription factors, TFs), coordinately bind to

these regions (TF binding sites, TFBSs) and produce the correct gene expression patterns. Taking advantage of the ability of fuzzy techniques to handle imprecision, inherent to TFBSs and regulatory-regions location data, a novel fuzzy approach is developed to study significant co-occurrences of closely located TFBSs in the yeast whole-genome. The methodology is based on a fuzzy frequent itemset mining algorithm and overcomes some of the limitations of previous approaches. The results obtained from the yeast genome enable us to propose the application of the procedure over more complex genomes in future works.

Agradecimientos

Me gustaría agradecer, en primer lugar, a mi director de tesis, Armando Blanco, por sus útiles consejos y su continuo apoyo.

Como no a mis compañeros de trabajo, Carlos y Fernando, con quienes he compartido fatigas desde el principio. A Luis Adarve, que nos dejó últiles consejos antes de huir a tierras más lejanas. También tengo que agradecer, como no podía ser de otra forma, su ayuda a mis compañeros más recientes, Marta y Alberto.

Tengo que aprovechar la ocasión para expresar mi agradecimiento a todos mis compañeros del despacho 16, por aguantar mis “palizas” de la tesis un día tras otro.

Desde luego, no habría podido llevar a cabo este trabajo sin el apoyo de mi familia y mi novia, Ángeles. Gracias por aguantarme estos años y, sobre todo, estos últimos meses.

CAPÍTULO 1

Introducción

1.1 Antecedentes

La secuenciación de los genomas de distintas especies, así como el desarrollo de tecnologías genómicas de alto rendimiento (como por ejemplo los microarrays), han generado una ingente cantidad de información génica estructural y funcional. Nunca ha habido disponible tal cantidad de información que permita estudiar sistemas biológicos, tales como células, órganos ó pacientes. Sin embargo, transformar todos estos datos en conocimiento útil no es una tarea sencilla. No se trata sólo del volumen de información creciente, sino de los diferentes tipos de datos recopilados, las relaciones entre ellos y otras peculiaridades de los mismos, tales como la imprecisión, valores perdidos, etc.

En la última década se hecho bastante hincapié en aspectos relacionados con herramientas orientadas a la automatización para la obtención de conocimiento biológico, prestándose poca atención a la integración de información de diversas fuentes. La razón de ello pudiera estar en la relativa simplicidad de los datos de los que se disponía. Sin embargo, en la era post-genómica, es necesario avanzar hacia la integración de los datos [22]. No es suficiente ya con conocer qué es un genoma: se necesita comprender lo que significan sus componentes, cómo funcionan y cómo se relacionan entre sí y con el todo. A pesar de la necesidad de la integración, en el momento del desarrollo de

este trabajo, la mayoría de los estudios previamente publicados sólo tenían en cuenta una única fuente de datos (por ejemplo, la matriz de expresión).

Así pues, no se trata sólo de un problema relacionado con el volumen de información, sino también de los diferentes tipos de datos generados, en otras palabras, de la heterogeneidad de los mismos. Estos datos pueden venir dados en forma de ontologías, secuencias, medidas, etc. Aunque en la actualidad están surgiendo algunos trabajos que afrontan este problema, aún no existen suficientes propuestas capaces de manejar la heterogeneidad que presenta este tipo de datos.

Por otra parte, es bien conocido que la información biológica tiende a ser imprecisa y a presentar un cierto grado de incertidumbre. Como norma general se suele hacer uso de técnicas clásicas para el análisis de datos biológicos. No obstante, existen otras metodologías (por ejemplo las metodologías difusas), que han demostrado ser más apropiadas para el tratamiento de este tipo de información y cuya aplicación no es habitual [290, 75].

Se propone pues, el desarrollo de una metodología capaz de integrar y analizar grandes cantidades de datos genómicos heterogéneos, y su aplicación a un organismo ampliamente estudiado que permita validar los resultados obtenidos. Éste podría ser el caso de la levadura *Saccharomyces cerevisiae*. Las propiedades únicas de este organismo, junto con sus muchas aplicaciones industriales, lo han convertido en uno de los organismos favoritos para la investigación biológica. Así, el genoma de la levadura fue el primer genoma eucariota en secuenciarse [82]. Desde entonces, el trabajo con este organismo ha ido abriendo paso al resto de estudios en genómica estructural y funcional, estableciendo un estándar en biología celular y molecular, y facilitando de este modo estudios similares en otros organismos [82, 102, 55, 279]. Todas estas investigaciones han generado conjuntos de datos de alta calidad, así como bibliografía abundante en la que se describen las tendencias y patrones en la organización genética.

Lo habitual para estudiar las propiedades del genoma de la levadura, ha sido la aplicación de técnicas tradicionales estadísticas [168, 169, 67, 136]. Sin embargo, la naturaleza de las técnicas estadísticas hace difícil la inte-

gración de datos heterogéneos en el análisis. Además, este tipo de técnicas permiten estudiar tan sólo unas pocas variables simultáneamente.

El análisis satisfactorio del genoma de la levadura permitiría llevar a cabo un estudio más ambicioso, como es trasladar la metodología a la extracción de conocimiento en el genoma humano. Más concretamente, se propone el estudio de las características genómicas del cáncer de mama. Esta enfermedad es un problema de salud pública de gran importancia; se trata del segundo cáncer más común en el mundo (una de cada ocho mujeres lo padece), y de la quinta causa más común de muerte por cáncer. Su alta incidencia y mortalidad lo han convertido en el centro de atención de gran cantidad de investigaciones. Sin embargo, no se conoce aún la causa por la qué algunas personas (principalmente mujeres) desarrollan el cáncer de mama. La etiología del cáncer de mama es aún hoy desconocida, aunque se sabe que las hormonas, factores genéticos y condiciones ambientales tienen un papel principal en su desarrollo [133].

El procedimiento tradicional de clasificación de tumores, permite determinar de forma clara el tratamiento para aquellos casos en los que el riesgo es elevado o bajo. Sin embargo, la mayoría de los tumores se encuentran en un grupo “intermedio”. En estas últimas situaciones, la opción “segura” consiste en el sobretratamiento del paciente, beneficiando a una relativamente baja proporción de los pacientes, y exponiendo innecesariamente al resto a los efectos secundarios derivados del tratamiento. Por el contrario, una estrategia más conservadora evitaría un tratamiento sin garantías y reduciría costes, provocando en este caso que ciertos pacientes que se beneficiarían del tratamiento no lo recibieran. Trabajos recientes [227] instan a la comunidad científica a investigar los vínculos existentes entre la expresión de los genes y el estado de los factores de pronóstico conocidos. Los avances en este sentido, podrían mejorar los tratamientos que se aplican a pacientes cuyo tumor presenta un pronóstico poco claro [227].

Por otra parte, el ADN de un organismo codifica todas las moléculas de ARN y de proteína necesarias para la construcción de sus células. Sin embargo, la descripción completa de la secuencia de ADN de un organismo no permitiría reconstruirlo. El problema radica en saber cómo se utilizan los

elementos de la secuencia de ADN, es decir, las reglas y mecanismos que determinan que en cada célula sólo se exprese específicamente una selección de los genes.

La vía que conduce del ADN hasta las proteínas está formada por muchas etapas (transcripción, maduración, transporte al citosol, etc.), y en principio todas ellas se pueden regular. Sin embargo, para la mayoría de los genes, los controles transcripcionales son los más importantes. Esto tiene sentido ya que, de todos los posibles puntos de control solamente el control transcripcional asegura que no se sinteticen intermediarios superfluos [11].

En las células eucariotas, la transcripción de un gen está controlada por combinaciones de proteínas (los factores de transcripción), que se unen de forma coordinada a las secuencias reguladoras de los genes. Estas secuencias reguladoras se encuentran dispersas por todo el ADN y se organizan en los llamados módulos de regulación (CRMs). La unión de los complejos de factores de transcripción a dichos módulos de regulación da lugar a los patrones de expresión adecuados.

Existen muchas propuestas computacionales previas para el estudio de los módulos de regulación [266, 147, 267]. En ellas, se pueden distinguir dos estrategias fundamentales para reducir el espacio de búsqueda [267]: *1)* estudiar tan sólo las regiones promotoras de los genes y *2)* centrar la búsqueda en regiones del genoma conservadas entre especies. El primer enfoque no puede, obviamente, capturar los elementos reguladores que aparecen en los intrones o en zonas alejadas del gen. Por otra parte, el estudio de las regiones conservadas del genoma requiere que las correspondientes secuencias sean lo suficientemente parecidas para poder ser alineadas, algo que no es frecuente en especies no cercanas [267]. Es más, experimentos con especies de mamíferos y *Drosophila* han demostrado que, incluso entre especies muy cercanas, entre uno y dos tercios de los TFBSs identificados no aparecen conservados [267, 78, 68, 87].

Por otra parte, según el objetivo pretendido por los diferentes métodos, se pueden distinguir tres categorías [266]: *1)* métodos que escanean secuencias (o genomas completos) buscando CRMs que siguen un modelo predefinido, *2)* métodos que buscan CRMs similares en un conjunto de genes relaciona-

dos (por ejemplo co-expresados o co-regulados), y *3)* métodos que escanean secuencias (o genomas completos) buscando grupos de TFBSs de cualquier combinación de factores de transcripción. Los primeros, requieren especificar el modelo de CRM que se busca, lo que es en muchos casos difícil o incluso imposible de proporcionar. Los segundos se centran en un conjunto determinado de genes, lo que significa que se restringen al análisis de sus secuencias promotoras, lo que a su vez conlleva las limitaciones antes descritas. Finalmente, los métodos de la tercera clase no necesitan realizar ningún tipo de asunción acerca del conjunto de TFs que cooperan. Se trata, por ello, de métodos más generales que los anteriores y, por tanto, no requieren ningún tipo de conocimiento previo. En este trabajo se presentará una metodología que puede enmarcarse en esta última clase de métodos.

1.2 Objetivos

El propósito general de este trabajo, consiste en desarrollar nuevas metodologías computacionales que solventen algunas de las limitaciones enunciadas. Las aplicaciones de dichas metodologías serán fundamentalmente dos:

- La integración y el análisis de datos biológicos heterogéneos, lo que permitirá llevar a cabo un estudio sobre ciertas características estructurales y funcionales del genoma de la levadura. Posteriormente, el estudio se trasladará al genoma humano, más concretamente, al análisis de las relaciones entre los niveles de expresión y los factores de pronóstico en el cáncer de mama.
- El estudio de los módulos de regulación, lo que permitirá avanzar en el conocimiento de las redes de regulación génica. La metodología desarrollada se aplicará en una primera fase sobre el genoma de la levadura, para posteriormente ampliar el estudio a otros genomas más complejos.

Contribuciones

Con el propósito de alcanzar estos objetivos, la atención se ha centrado inicialmente en el desarrollo de una metodología difusa-integrativa, que ha per-

mitido analizar conjuntos de datos heterogéneos de grandes dimensiones. El principal aspecto de esta metodología consiste en la implementación de un algoritmo eficiente de extracción de reglas de asociación difusas. Los conjuntos difusos son especialmente apropiados para modelar datos imprecisos y con ruido, mientras que las reglas de asociación manejan con cierta facilidad datos heterogéneos. Por tanto, las técnicas de extracción de reglas de asociación difusas son particularmente apropiadas para los propósitos que aquí se perseguían.

A continuación, se ha llevado a cabo un estudio sobre el genoma de la levadura. Para esto, se ha aplicado la metodología difusa desarrollada, con el objetivo de analizar las relaciones existentes entre diversas características estructurales y funcionales del genoma de la levadura. Los resultados obtenidos de este análisis han permitido:

- *Validar la metodología*, comparando los resultados obtenidos con las tendencias descritas previamente entre las variables estudiadas.
- *Descubrir nuevas relaciones*, ya que se obtendrían asociaciones nuevas que podrían contribuir al entendimiento de las relaciones genómicas estructurales-funcionales.

La experiencia obtenida de este estudio ha permitido abordar un nuevo trabajo, como es el análisis de las características genómicas del cáncer de mama. Este nuevo estudio se ha desarrollado en colaboración con el departamento de Anatomía Patológica del Hospital Universitario Virgen de las Nieves de Granada. El trabajo se ha llevado a cabo en dos fases:

- *Obtención de medidas de expresión génicas a partir de arrays de Affymetrix GeneChip U133 plus 2.0*. Lo que ha requerido el estudio de las diferentes técnicas de preprocesamiento disponibles para este modelo de arrays. Una vez confirmada la validez de los datos preprocesados, éstos se integraron en una base de datos junto con información clínica de los pacientes correspondientes.
- *Análisis del conjunto de datos resultante*. Dicho conjunto de datos presenta ciertas características comunes con el conjunto de datos anterior (grandes dimensiones, heterogeneidad, imprecisión e incertidumbre),

por lo que la metodología basada en reglas de asociación difusas es apropiada para su análisis. El conocimiento proporcionado por las reglas de asociación puede ayudar a los expertos a detectar nuevos marcadores en el cáncer de mama.

Con respecto al segundo objetivo, se ha centrado la atención en el estudio de los mecanismos que permiten a la célula regular la expresión de los genes. En este sentido, se ha desarrollado una nueva metodología para el estudio de los módulos de regulación que solventa algunas de las limitaciones presentadas anteriormente. La metodología se basa en el uso de un algoritmo de cluster para la localización en el genoma de sitios de unión de factores de transcripción cercanos, y el posterior análisis de los clusters obtenidos mediante la extracción de itemsets difusos. Algunas de las propiedades interesantes que presenta dicha metodología son:

- Requiere poca información del usuario.
- Permite analizar el genoma completo.
- Uso de una metodología difusa, lo que permite modelar de forma más realista los módulos de regulación.
- No impone restricciones *a priori* en el tipo de elementos reguladores que se buscan.
- Los resultados obtenidos son fácilmente interpretables.

El método se ha aplicado sobre el genoma de la levadura, lo que ha permitido confirmar su buen funcionamiento y plantear su aplicación a genomas más complejos en trabajos futuros.

1.3 Estructura de la memoria

Los capítulos de esta memoria se organizan como sigue:

- En el Capítulo 2 se exponen algunos conceptos básicos que pueden facilitar la comprensión del resto de la memoria. En primer lugar, se incluyen ciertas nociones básicas acerca del almacenamiento y uso de la información genética en las células. A continuación, se proporciona una visión global de aplicaciones de técnicas computacionales en

Bioinformática. Finalmente, se centra la atención en técnicas computacionales de especial interés en este trabajo, tales como algoritmos de clústering, técnicas de extracción de reglas de asociación ó métodos de detección de elementos reguladores en cadenas de ADN.

- El Capítulo 3 expone la metodología basada en reglas de asociación difusas, así como su aplicación al genoma de la levadura. En primer lugar, se incluye una introducción que explica las motivaciones que dieron lugar a este trabajo. Seguidamente, se describe la información incluida en el análisis y su preprocesamiento. A continuación, se explica detalladamente el algoritmo de extracción de reglas de asociación junto con la aplicación web que permite su utilización. Los resultados obtenidos se validan y se comentan en la Sección 3.5. Finalmente, se presenta una sección que incluye las conclusiones extraídas de este trabajo.
- El Capítulo 4 contiene el análisis de la información genómica en el cáncer de mama. Nuevamente, la primera sección expone las motivaciones que dieron lugar al trabajo. A continuación, la Sección 4.2 introduce la problemática del cáncer de mama desde un punto de vista biomédico. Seguidamente, se describe extensivamente el conjunto de factores de pronóstico incluidos en el estudio, así como el procedimiento seguido para obtener las medidas de expresión de los genes. A continuación, se comenta la metodología usada para extraer las reglas de asociación difusas, la cual coincide básicamente con la expuesta en la Sección 3.3.1. Seguidamente, las relaciones obtenidas son comentadas y comparadas con los resultados previamente publicados. Para finalizar, se incluye una sección con las conclusiones obtenidas del estudio.
- El Capítulo 5 describe la nueva metodología propuesta para estudiar los elementos reguladores de un genoma. Como ya ocurriera en los dos capítulos anteriores, la primera sección del Capítulo introduce el trabajo. A continuación, se dedica una sección completa a la descripción de la metodología, la cual comprende tres pasos fundamentales: *1)* obtención de grupos de sitios de unión de factores de transcripción cercanos, *2)* análisis de los grupos obtenidos mediante la extracción de

itemsets frecuentes difusos y *3)* postprocesado de los itemsets resultantes. En la última parte del capítulo se comentan tres experimentos diferentes llevados a cabo sobre el genoma de la levadura. La última Sección del Capítulo incluye las conclusiones de este trabajo.

- A continuación, se incluye un último apartado que contiene las conclusiones globales del trabajo, las nuevas líneas de investigación surgidas durante el desarrollo de esta memoria y las publicaciones derivadas de esta investigación.
- Finalmente, se presenta la bibliografía reseñada en esta memoria.

CHAPTER 1

Introduction

1.1 Background

The availability of the complete genome from diverse species and the advent of high-throughput genomic technologies have generated a great amount of structural and functional genomic information. There has never been more potentially available information to study biological systems like cells, organs, or patients. However, it is a non-trivial task to transform the vast amount of biomedical data into useful information supporting scientific progress and/or patient management. It is not only the volume of information growing, but also the types of data being collected, the relationships between them and certain peculiarities they have, such as imprecision, noise or missing values.

Until recently biological data analysis had put emphasis on the automation aspects of tools, and relatively little attention had been paid to the integration of information and models. This is probably due to the relative simplicity of pre-genomic data. However, in the post-genomic era integrative analysis are required in order to achieve higher understanding levels. Now, it is no longer enough just to know what a genome is: it is necessary understand what its components mean, how they work and how they relate to the whole [22]. Nevertheless, it is noteworthy that at the time of developing this

work, most of previous studies focussed on the analysis of a single-source dataset (e.g. a gene expression matrix).

As already stated, it is not only the volume of information, but also the diversity of data types, in other words, the heterogeneity of biological data. These data can be found in the form of ontologies, sequences, measures etc. Although some approaches that carry out analysis of heterogeneous information are emerging, there is still a lack of integrative approaches capable of handling a broad variety of data types. Another key point is the imprecision and noisy nature of biological data. Classical crisp techniques are usually applied to analyze biological data. Nevertheless, other methods which are known to perform better when dealing with imprecise and noisy data (e.g. fuzzy techniques) are barely used.

Thus, we here propose the development of a methodology for the integration and analysis of heterogeneous and high-dimensional genomic datasets. The yeast *S. cerevisiae* genome was choosen as a benchmark, since intensive work on this model organism has provided high quality datasets and also abundant literature exploring the trends and patterns in genomic organization and function. The yeast *Saccharomyces cerevisiae* was the first eukaryote to have its genome sequenced [82]. Since then, work with this organism has led the way in structural and functional genomics, setting the standard for the global analysis of cellular and molecular biology and paving the way for similar approaches in other organisms [82, 102, 55, 279].

Traditional statistical techniques have been typically used to analyze the yeast genome properties [168, 169, 67, 136]. Nevertheless, the nature of statistical techniques makes hard the integration of diverse heterogeneous data into the analysis. Furthermore, these techniques allow to study only few potential relations between the biological variables they consider.

A satisfactory application of the methodology over the yeast genome would allow to move the study to more complex genomes, such as the human genome. In particular, it is proposed the analysis of the genomic peculiarities of breast cancer. Breast cancer is a great public health problem. In fact, it is the second most common cancer worldwide (one in eight women) and the fifth most common cause of cancer death. Its high incidence and death rate

have made it to be the focus of a huge research effort. However, it is still unknown why some people (mainly women) get breast cancer. The specific causes of breast cancer remain unknown, although many epidemiological risk factors, such as hormone status, genetic factors, and environmental conditions, have been identified [133].

Traditional classifications of tumors may provide clear-cut treatment options in high-risk and low-risk cases. Nevertheless, often tumors fall into an "intermediate" group; it is in these borderline cases where improvements are most urgently required. In these cases the "safe" option is to overtreat, benefiting a relatively small minority of cases and exposing the rest to undesirable side effects. Conversely, a more conservative approach may avoid unwarranted treatment and additionally reduce costs, but some patients who would benefit may be undertreated. Recent works encouraged the scientific community to study the links between gene expression and known prognostic factors, arguing that this type of studies may be beneficial for this intermediate group [227].

An organism's DNA encodes all of the RNA and protein molecules required to construct its cells. Yet a complete description of the DNA sequence of an organism does not enable us to reconstruct it. The problem is to know how the elements in the DNA sequence are used. Under what conditions is each gene product made, and, once made, what it does.

There are many steps in the pathway leading from DNA to protein, and all of them can in principle be regulated. However, for most genes the initiation of RNA transcription is the most important point of control, only transcriptional control ensures that the cell will not synthesize superfluous intermediates [11].

Eucaryotic gene control regions consist of a promoter plus regulatory DNA sequences which may appear distant from the gene promoter. These regulatory regions are, in turn, organized in the so-called *cis*-regulatory modules (CRMs). Regulatory proteins (called transcription factors, TFs), coordinately bind to these regulatory modules and produce the correct gene expression patterns.

There are many works which study the regulatory modules *in silico* [266, 147, 267]. Two main strategies are followed in order to reduce the search

space, which consist of either limiting the study to the gene promoter regions or focussing on genome conserved regions [267]. On one hand, the former approach clearly misses regulatory elements not located in the upstream of known genes. On the other hand, the latter approaches are based on the evolutionary conservation of regulatory sequences. The drawback of this approach is that regulatory sequences have to be similar enough to be aligned, which is often not the case when the compared sequences are not very closely related. Moreover, experiments on mammalian and *Drosophila* species have shown that between one and two-thirds of identified regulatory sequences are not conserved even between relatively closely related species [267, 78, 68, 87].

In addition, three conceptually different classes of methods can be identified according to their specific aims: CRM scanners, CRM builders and CRM genome screeners [266]. The first ones scan for sequences that satisfy a strictly defined CRM model, which is often difficult or even impossible to provide. CRM builders look for similar CRMs in a set of co-regulated or co-expressed genes. This implies to analyze the upstream region of these genes, which is subjected to the above described limitations. Finally, the third type of methods screen sequences or complete genomes for CRMs. These last methods make only few assumptions regarding the CRMs they aim to detect and are the most generally applicable methods. The methodology presented in this work can be framed into this last type of approaches.

1.2 Objectives

The overall objective of this work consists of developing new computational techniques that overcome some of the above presented limitations. These methodologies will be oriented to their application in two different fields:

- The integration and analysis of heterogeneous biological data. This will allow to carry out a study on the relationships between a diversity of structural and functional variables of the yeast genome. The study will then be moved to the analysis of the links between gene expression levels and prognostic factors in breast cancer.

- The study of *cis*-regulatory modules. The developed methodology will contribute to achieve a better understanding of gene regulatory mechanisms. The yeast genome will be analyzed by applying such methodology.

Contributions

In order to achieve these objectives, the attention has been firstly focussed on the development of a fuzzy-integrative methodology capable of analyzing high-dimensional and heterogeneous biological data. The millstone of this methodology is the implementation of an efficient fuzzy association rule mining algorithm. Fuzzy set theory is specially suitable to model imprecise data, while association rules are very appropriate to carry out an integrative analysis of heterogeneous data. Thus, a fuzzy association rule mining algorithm is a suitable method for the purposes of this work.

Then, a study has been carried out over the yeast *S. cerevisiae* genome. The developed fuzzy methodology has been applied to find relationships between a variety of genomic characteristics comprising both structural and functional features. This study has been useful in two different aspects:

- *Validating the methodology*, by comparing the results with the previously reported trends.
- *Getting new insights*, since new associations have been obtained which may contribute to the framing of genomic structural and functional relationships.

The experience gained from this study has been helpful to face up the thorough investigation of genomic peculiarities in breast cancer. This study has been carried out in collaboration with the department of Pathology of the Hospital Universitario Virgen de las Nieves (Granada). This work comprised two stages:

- *Obtaining gene expression measures from Affymetrix arrays GeneChip U133 plus 2.0*. This required a comprehensive study of the available preprocessing techniques for this type of arrays. Once the validity of the preprocessed data was confirmed, they were integrated in a

database together with clinical information of the corresponding patients.

- *Analysis of the resultant dataset*. Such dataset shares some features with the previous one: high-dimensionality, heterogeneity, imprecision and uncertainty. Thus, the methodology is appropriate for its analysis. The knowledge derived from the discovered rules may help experts to unveil new biomarkers in breast cancer.

Regarding the second objective, the attention has been focussed on studying the mechanisms involved in gene expression regulation. A new methodology has been developed which overcomes some of the limitations of previous approaches. The procedure is based on the use of a clustering algorithm to find groups of closely-located transcription factor binding sites, and the subsequent analysis of the obtained groups by fuzzy frequent itemset mining. Some of the main properties of the procedure are:

- Requires little information from the user.
- Allows to scan a complete genome.
- The fuzzy aspect of the methodology allows to better model regulatory modules.
- Does not impose *a priori* constraints on the type of recovered regulatory elements.
- Results are presented in an intuitive way.

The methodology has been applied over the yeast genome in order to validate its functioning. Results derived from this analysis enable us to propose the application of the procedure over more complex genomes in future works.

1.3 Structure of the manuscript

Chapters are organized as follows:

- Chapter 2 includes some basic concepts which may be helpful for the easy-understanding of the manuscript. Firstly, some basic notions regarding the storage and use of genetic information in cells are shown.

Then, an overview on applications of computer science techniques in Bioinformatics is given. The state-of-the-art of those Bioinformatics fields closely related with this work is provided at the end of this Chapter.

- Chapter 3 describes the fuzzy association rule mining methodology and its application to the *S. cerevisiae* genome. First, an introductory section motivating the work is presented. Then, the information gathered from the yeast genome and its preprocessing is described. Next, the fuzzy association rule mining algorithm is explained together with a freely-accessible web application. The results are validated and discussed in Section 3.5. Finally, the conclusions extracted from this study are included in the last section.
- Chapter 4 contains the analysis of genomic information in breast cancer. An introductory section motivates this study. Next, Section 4.2 introduces the breast cancer problem from a biomedical point of view. Then, the set of prognostic factors included in the study is described in detail, as well as the preprocessing procedure used to obtain a measure of the gene expression levels. Next, the methodology used to obtain the fuzzy association rules is commented. This bassically coincides with the exposed in Section 4.4. Then, the obtained relations are compared with previously reported results. Finally, the conclusions obtained from this work are presented.
- Chapter 5 describes the new methodology proposed for the study of regulatory elements in a genome. As in the previous chapters, the work is firstly motivated in an introductory section. Then, the developed methodology is described, which comprises: *1)* the procedure followed for finding groups of closely located transcription factor binding sites, *2)* the methodology based on fuzzy frequent itemset mining for analyzing co-ocurrences in the obtained groups and *3)* the post-processing of the resulting itemsets. The last part of the chapter discusses three different experiments carried out over the yeast genome. The conclusions are included at the end of the Chapter.

- The last Chapter contains the overall conclusions, the future tasks arised during the development of this project and the publications derived from this work.
- Finally, the complete list of publications referenced throughout the manuscript is provided.

CAPÍTULO 2

Preliminares

En este primer capítulo se exponen algunos conceptos básicos necesarios para una mejor comprensión de la memoria. En primer lugar, se incluyen ciertas nociones básicas acerca del almacenamiento y uso de la información genética en las células. Seguidamente se describen los microarrays: su importancia y funcionamiento. A continuación, se proporciona una visión global de aplicaciones de técnicas computacionales en Bioinformática. Finalmente, se centra la atención en técnicas computacionales de especial interés en este trabajo, tales como algoritmos de clústering, técnicas de extracción de reglas de asociación y métodos de detección de elementos reguladores en cadenas de ADN.

2.1 Algunos conceptos biológicos básicos

Todos los seres vivos están formados por células. Estas unidades de materia viva comparten una maquinaria común para sus funciones más básicas. Muchos organismos son enormes conglomerados pluricelulares en los que grupos de células realizan funciones específicas y se relacionan mediante elaborados sistemas de comunicación. No obstante, en todos los casos cada organismo se ha originado por división celular a partir de una sola célula. Así, la célula es el vehículo a través del cual se transmite la información hereditaria (*información genética*) que define cada especie. Además, determinada

por esta misma información, la célula contiene la maquinaria necesaria para obtener materiales del ambiente y generar una nueva célula a su imagen, que contendrá una nueva copia de su información genética.

2.1.1 El flujo de la información genética: dogma central

Todas las células guardan su información genética mediante un mismo código en moléculas de *ácido desoxirribonucléico* (ADN). A principio de la década de los años cuarenta, se descubrió que la información almacenada en el ADN consta esencialmente de instrucciones para producir otras moléculas diferentes denominadas *proteínas*. Se sabe que fragmentos del ADN llamados *genes* contienen la información necesaria para sintetizar estas proteínas. Cada gen contiene información para la producción de una proteína única, la cual realizará una función especializada en la célula. El genoma humano, por ejemplo, contiene más de 25000 genes.

Las propiedades y funciones de una célula vienen determinadas por las proteínas que sintetiza, ya que estas macromoléculas son las responsables de la mayor parte de las funciones celulares: actúan como bloques de construcción para las estructuras celulares, como enzimas que catalizan todas las reacciones químicas, permiten el movimiento celular y la comunicación intercelular, etc. En la síntesis protéica a partir de la información contenida en el ADN intervienen otras moléculas de estructura química similar a la del ADN, las moléculas de *ácido ribonucléico* (ARN). Éstas se encargan de transportar una copia de la información contenida en las moléculas de ADN al lugar de síntesis de las proteínas en la célula.

Las relaciones de información entre ADN, ARN y proteínas son circulares: el ADN contiene instrucciones para la síntesis de todas las proteínas que la célula necesita, el ARN es una molécula que sirve de intermediaria entre las instrucciones del ADN y la formación de proteínas, y proteínas específicas participan en la síntesis y el metabolismo del ADN y el ARN. A este flujo de información se le denomina *dogma central* de la biología molecular (Figura 2.1).

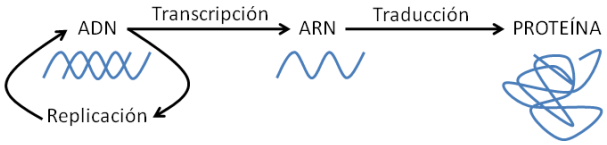
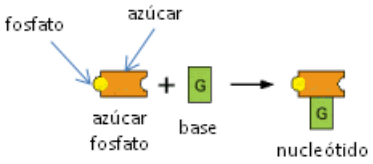
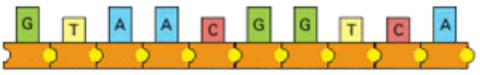


Figura 2.1: Flujo de información genética.



(a) Cada monómero de la cadena de ADN está formado por un grupo azúcar-fosfato y una base.



(b) Polímero formado por una cadena de nucleótidos.

Figura 2.2: Estructura del ADN (Figuras obtenidas de la referencia [11])

2.1.2 Almacenamiento de la información genética

Tal y como se acaba de comentar, todas las células guardan su información genética mediante un mismo código en moléculas de ácido desoxirribonucleico (ADN). Las moléculas de ADN están formadas por dos largas cadenas de *nucleótidos*. Estos *nucleótidos* están unidos entre sí formando una larga secuencia lineal que codifica la información genética de la célula. En la actualidad y gracias a diversos métodos químicos, los científicos pueden obtener y leer la secuencia de *nucleótidos* presente en cualquier molécula de ADN y así descifrar la información genética contenida en cada organismo.

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) (Figura 2.2a). Cada azúcar está unido al siguiente azúcar de la cadena por el grupo fosfato mediante un enlace fosfodiéster, formando un polímero cuyo eje central está formado por los azúcares fosfato y al cual se unen las bases (Figura 2.2b). Los extremos de la cadena de ADN se nombran habitualmente por la numeración de los carbonos de la desoxirribosa: $5'$ y $3'$ respectivamente (Figura 2.3). La lectura de la cadena de ADN se lleva a cabo siempre en dirección $3' \rightarrow 5'$.

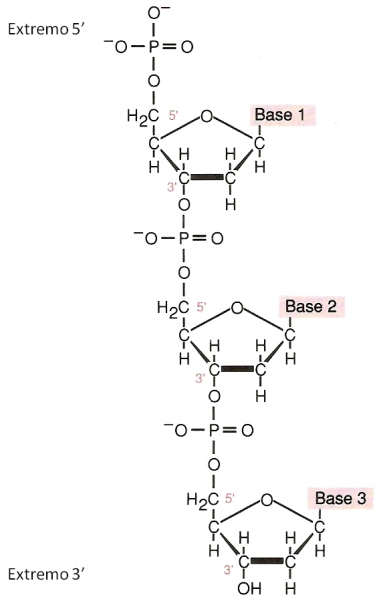
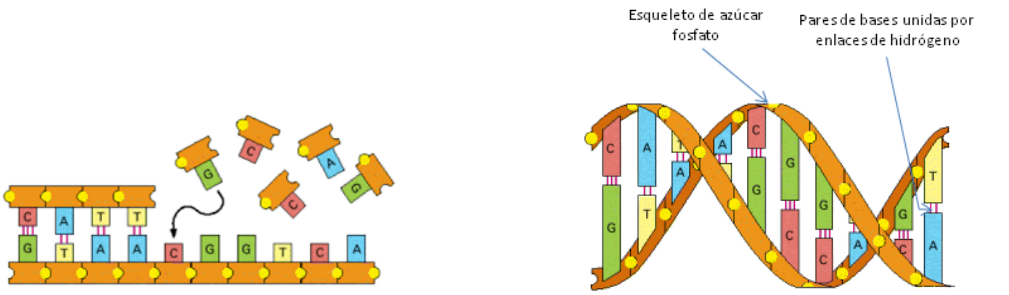


Figura 2.3: Detalle de una cadena de ADN que muestra los enlaces de fosfodiéster $3' - 5'$ que unen nucleótidos adyacentes (Figura obtenida de [255]).

El polímero de ADN puede crecer por la unión de monómeros a uno de sus extremos. En el caso de una cadena sencilla de ADN, los monómeros pueden incorporarse al polímero de forma aleatoria, sin un orden preestablecido, ya que todos los nucleótidos pueden unirse entre sí del mismo modo en el sentido del crecimiento del polímero del ADN. Por el contrario, en la célula viva existe una limitación, ya que el ADN no se sintetiza como una cadena libre aislada, sino sobre un patrón o molde de ADN de otra cadena preexistente. Las bases contenidas en la cadena patrón se unen a las bases de la nueva cadena siguiendo una estricta norma de complementariedad: A se une a T y C se une a G. Este apareamiento une los nuevos monómeros de la cadena y además controla la selección del monómero que se añade a la cadena (Figura 2.4a). De esta forma, una estructura de doble cadena consiste en dos secuencias complementarias de A, C, T y G. Además, las dos cadenas de nucleótidos se enrollan una sobre la otra generando una doble hélice (Figura 2.4b).

Los enlaces establecidos entre las bases son débiles si se comparan con



(a) La nueva cadena de ADN se sintetiza a partir de un patrón.

(b) Las dos cadenas se enrollan formando una doble hélice.

Figura 2.4: Estructura del ADN (Figuras obtenidas de la referencia [11]).

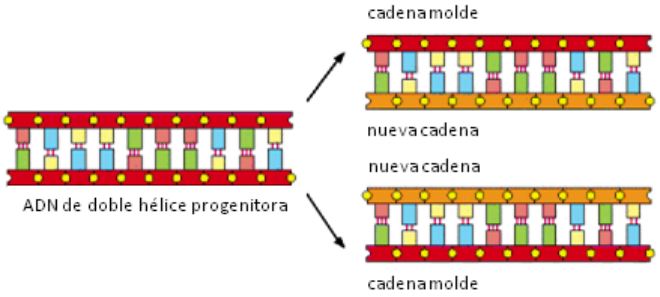


Figura 2.5: La doble hélice se abre permitiendo obtener una copia de la misma (Figura obtenida de [11]).

las uniones azúcar fosfato del resto del esqueleto. Esta debilidad permite separar las dos cadenas de ADN sin forzar la rotura de su esqueleto. Cada una de las cadenas puede comportarse como un molde o patrón en la síntesis de un nuevo ADN, una nueva copia de información genética (Figura 2.5). En diferentes tipos celulares, el proceso de *replicación de ADN* se produce a velocidades diferentes, con controles de inicio y final diferentes y con moléculas reguladoras auxiliares distintas. Pero la base del proceso es universal: el ADN constituye el almacén de la información y la polimerización sobre un patrón o molde es el modo en el que esta información hereditaria se copia y se transmite.

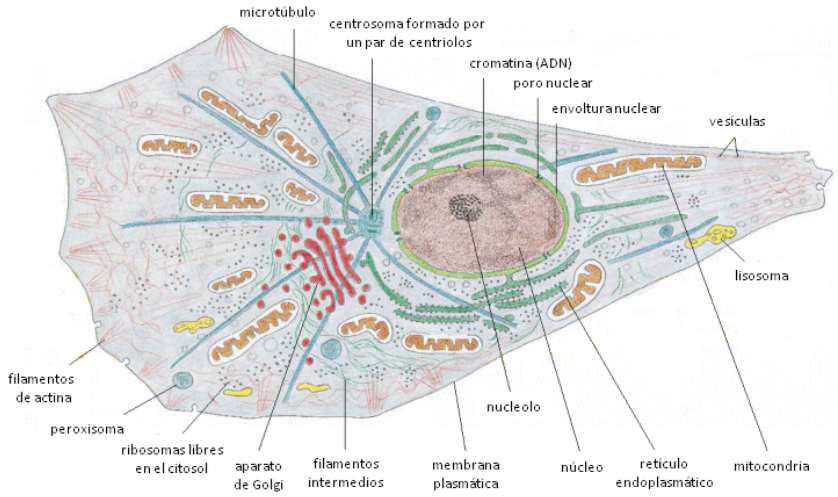


Figura 2.6: Esquema de una típica célula animal (Figura obtenida de la referencia [11]).

Organización del ADN en eucariotas

Los organismos vivos pueden clasificarse en dos grupos atendiendo a su estructura: organismos *eucariotas* y *procariotas*. Los eucariotas guardan su ADN en un compartimiento intracelular denominado núcleo (Figura 2.6). Los procariotas no presentan un compartimiento nuclear diferenciado para almacenar su ADN. Las plantas, hongos y los animales son eucariotas; las bacterias son procariotas. Este trabajo se centra en el estudio de genomas eucariotas.

El ADN se encuentra compactado en el núcleo. Téngase en cuenta por ejemplo, que cada célula humana contiene aproximadamente 2 metros de ADN. Si se uniera el ADN de todas las células humanas se obtendrían unos 40 km de un hilo extremadamente fino. Esta compleja tarea de empaquetar el ADN la realizan unas proteínas determinadas que se asocian al ADN. Dichas proteínas se encargan de plegarlo y facilitar la formación de enrollamientos y bucles, contribuyendo así al establecimiento de niveles de organización superiores del ADN. De este modo, se evita que la cadena se convierta en un enmarañado e inmanejable ovillo con el que no se pueda trabajar. Sorprendentemente, a pesar de que el ADN se empaqueta de una forma muy

compacta, lo hace de tal forma que permanece accesible a las enzimas que intervienen en su replicación, en su reparación y en la utilización de los genes para producir proteínas.

En las células eucariotas, el ADN se encuentra fragmentado en varias porciones lineales denominadas *cromosomas*. Por ejemplo, el genoma humano -de aproximadamente unos $3{,}2 \cdot 10^9$ nucleótidos- está distribuido en 24 pares de cromosomas diferentes. Cada cromosoma está formado por una sola molécula de ADN, muy larga, asociada a proteínas que pliegan y empaquetan la final hebra de ADN formando una estructura más compacta. Tradicionalmente las proteínas que se unen al ADN de los cromosomas eucariotas se clasifican en dos grupos: las *histonas* y las *proteínas cromosómicas no histonas*. El complejo formado por el ADN cromosómico y las dos clases de proteínas se denomina *cromatina*. Las histonas están presentes en grandes cantidades (alrededor de 60 millones de moléculas de cada tipo por célula humana), de forma que en la cromatina la masa total de histonas es aproximadamente igual a la de ADN.

Las histonas son responsables del primer y más básico nivel de organización del cromosoma, el *nucleosoma*. Si la cromatina se somete a tratamientos que la descondensen es posible su observación al microscopio electrónico como una serie de "cuentas de collar" (Figura 2.7). El collar es el ADN y cada una de las cuentas es el "núcleo de una partícula nucleosómica" que está formado por ADN que envuelve este núcleo de proteínas. El collar de cuentas constituye el primer nivel de empaquetamiento del ADN. El posicionamiento exacto de los nucleosomas a lo largo de la hebra de ADN depende de factores entre los que se incluye la secuencia del ADN y la presencia y naturaleza de otras proteínas unidas al ADN. Además, la distribución de los nucleosomas es un proceso muy dinámico, cambiando continuamente en relación a las necesidades de la célula.

Aunque la mayor parte del ADN cromosómico forma largas cadenas de nucleosomas, es poco probable que en una célula viva la cromatina se disponga en la forma de "collar de cuentas". En lugar de ello, los nucleosomas se empaquetan uno sobre otro formando una fibra de cromatina compacta (Figura 2.8a). Así, se han propuesto varios modelos para explicar cómo los

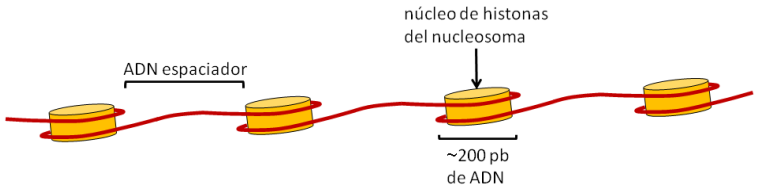


Figura 2.7: Estructural del ADN den forma de collar de cuentas.

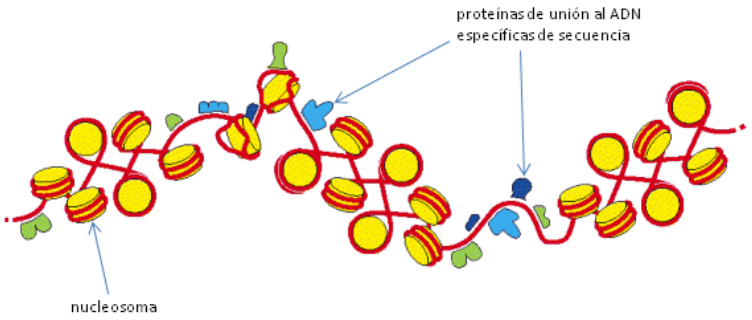
nucleosomas se empaquetan en esta fibra de cromatina; el que está más en concordancia con los datos de que se dispone es el que hace referencia a series de variaciones estructurales, denominado modelo en zigzag. En realidad, la estructura encontrada en los cromosomas es probablemente una secuencia de diferentes variaciones del zigzag. Tal y como se ha comentado, la longitud del ADN espaciador que conecta los nucleosomas vecinos puede ser variable; probablemente, estas variaciones del espaciador introducen perturbaciones locales en la estructura de zigzag. Finalmente, la presencia de otras proteínas que se unen al ADN dificulta el plegamiento entre los nucleosomas e introduce interrupciones irregulares en la fibra (Figura 2.8b).

Los genes

Un gen se puede definir como un segmento de ADN que contiene las instrucciones para fabricar una determinada proteína (o un grupo de proteínas que están muy estrechamente relacionadas entre sí). A pesar de que esta definición la cumplen la mayoría de genes, existe un porcentaje de genes cuyo producto final es una molécula de ARN en lugar de una proteína. Al igual que las proteínas, estas moléculas de ARN ejecutan, en células diversas, funciones estructurales y catalíticas.

Por lo general, un gen comprende tan sólo unos cuantos miles de nucleótidos. Ciertas señales codificadas en la cadena de ADN indican los puntos en los que comienza y acaba el gen. Sin embargo, estas señales presentan secuencias de nucleótidos heterogéneas, siendo enormemente difícil localizarlas en genomas eucariotas. A menudo, es necesario disponer de información adicional, en parte procedente de la experimentación directa, para localizar con precisión estas cortas señales de ADN contenidas en el genoma. Así, la

(a) Al microscopio electrónico, la cromatina aparece como una fibra compacta.



(b) Empaquetamiento de los nucleosomas.

Figura 2.8: Estructura de las fibras de cromatina (Figuras obtenidas de la referencia [11])

identificación de las secuencias codificadoras de los genes es un campo de investigación en continuo desarrollo.

Otro aspecto importante de los genes eucariotas es que las secuencias codificadoras de los mismos suelen estar interrumpidas por secuencias no codificantes. Así, los genes eucariotas están divididos en pequeños fragmentos de secuencia codificadora (*exones*), separadas por otras secuencias mucho más largas, los *intrones* (Figura 2.9). De nuevo, se sabe que la propia secuencia de nucleótidos indica los límites de los intrones. Sin embargo, resulta muy difícil determinar los límites precisos de un intrón.

Estructura del ARN

Como el ADN, el ARN es un polímero lineal de cuatro tipos diferentes de nucleótidos unidos por enlaces fosfodiéster. Desde el punto de vista químico se diferencia del ADN en dos aspectos: (1) los nucleótidos del ARN son *ribonucleótidos*, es decir, contienen el azúcar ribosa (de ahí el nombre de ácido

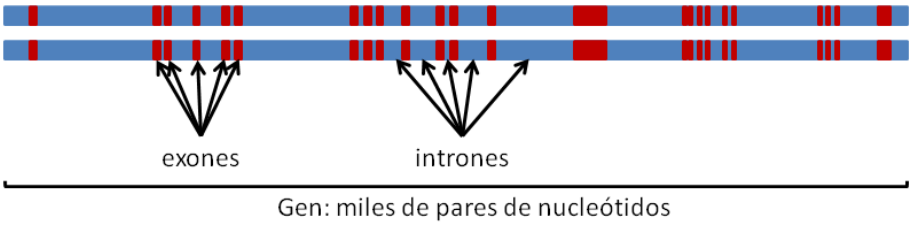


Figura 2.9: Los genes se encuentran divididos en intrones y exones.

*ribo*nucleico) en vez de desoxirribosa; (2) como el ADN, el ARN contiene las bases de adenina (A), guanina (G) y citosina (C), pero en lugar de timina (T) contiene uracilo (U). El U, al igual que la T, se aparea mediante puentes de hidrógeno con la A.

El ADN y el ARN se diferencian ampliamente en su estructural global. Así, el ADN siempre aparece como una doble hélice mientras que el ARN es de cadena sencilla. Las cadenas de ARN se pueden plegar en una amplia variedad de formas, lo que permite que algunas moléculas de ARN desarrollen funciones estructurales y catalíticas. Además, dado que sólo se copia una pequeña región del ADN, las moléculas de ARN son mucho más cortas. Una molécula de ADN de un cromosoma humano puede tener hasta 250 millones de pares de bases de longitud, mientras que la mayoría del ARN no tiene más de unos cuantos miles de nucleótidos y muchas de ellas son incluso menores.

2.1.3 El ciclo celular

Una célula se reproduce cuando lleva a cabo una secuencia ordenada de acontecimientos en los cuales duplica su contenido y luego se divide en dos. Este ciclo de duplicación y división, conocido como *ciclo celular*, es el mecanismo esencial mediante el cual todos los seres vivos se reproducen.

Durante el ciclo celular, el ADN se duplica y se reparte en dos células hijas idénticas. Estos procesos definen las dos *fases* principales del ciclo: la fase *S* y la fase *M*. La duplicación del ADN sucede en la fase *S* (de síntesis), la cual ocupa alrededor de la mitad del tiempo que dura el ciclo celular en una célula de mamífero típica. Después de la fase S tiene lugar la fase *M* (de

mitosis), la cual agrupa la mitosis (reparto de material genético nuclear) y la citocinesis (división del citoplasma), ocupando mucho menos tiempo del ciclo celular que la fase S.

Sin embargo, la mayoría de las células tardan mucho más tiempo en crecer y duplicar su masa de proteínas y orgánulos que el que necesitan para replicar su ADN y dividirse. Esto se debe a que en la mayoría de ciclos celulares, se intercalan *fases de descanso* que permiten a las células disponer de más tiempo para crecer -la *fase* G_1 entre la fase M y la fase S y la *fase* G_2 entre la fase S y la mitosis-. De este modo, el ciclo celular eucariota se divide tradicionalmente en cuatro fases secuenciales: G_1, S, G_2 y M (Figura 2.10).

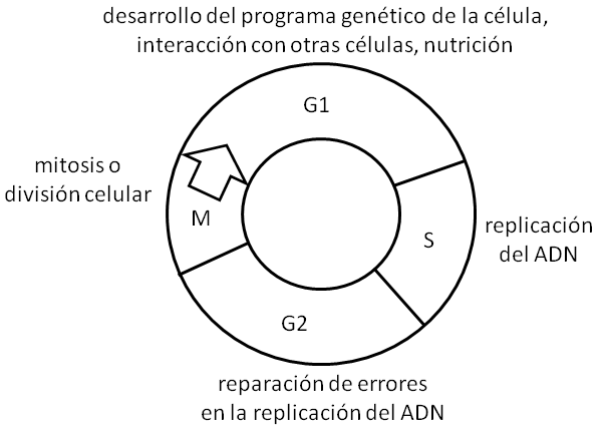


Figura 2.10: Fases del ciclo celular.

2.1.4 La expresión de los genes: del ADN a la proteína

Para llevar a cabo satisfactoriamente la función de almacén de la información, el ADN tiene que hacer algo más que duplicarse antes de cada división celular: debe expresar la información que contiene, utilizándola para dirigir la síntesis de otras moléculas en la célula. El mecanismo responsable de este proceso es el mismo en todos los organismos vivos y se divide en dos etapas: *transcripción* y *traducción*. El proceso comienza con la polimerización sobre un patrón, denominada *transcripción*, proceso en el que diferentes segmentos de la secuencia de ADN se utilizan como molde para la síntesis del *ARN*.

Posteriormente, en un proceso complejo denominado *traducción*, muchas de estas moléculas de ARN se utilizan para dirigir la síntesis de polímeros de una clase química radicalmente diferente: las *proteínas*.

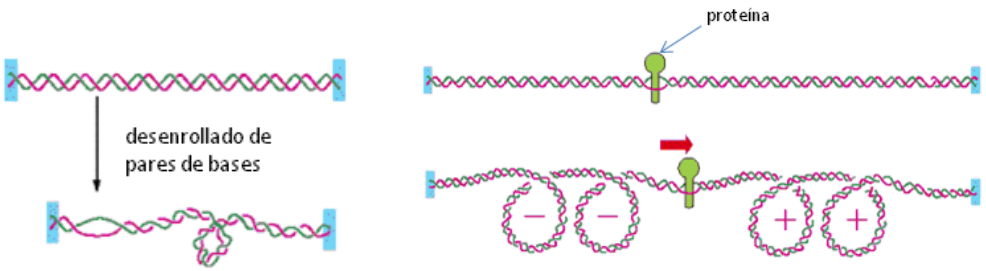
La transcripción

Tal y como se ha comentado, el primer paso que lleva a cabo una célula para leer la parte que necesita de las instrucciones genéticas, es copiar un fragmento particular de su secuencia de nucleótidos de ADN -un *gen*- a una secuencia de ARN. La información del ARN, aunque esté copiada en una forma química diferente, continúa estando escrita esencialmente con el mismo lenguaje que la del ADN.

La transcripción se inicia con la abertura y el desenrollamiento de una pequeña zona de la doble hélice de ADN, dejando al descubierto las bases de cada una de las dos hebras del ADN. Una de ellas actúa como molde para la síntesis de una molécula de ARN. La secuencia de nucleótidos de la cadena del ARN viene determinada por la complementariedad de bases entre los nucleótidos que se van incorporando y el molde del ADN. Cuando se produce un apareamiento correcto, el ribonucleótido recién incorporado se une covalentemente a la cadena de ARN que se está sintetizando, mediante una reacción catalizada enzimáticamente. La cadena de ARN originada durante la transcripción -el transcrito- se va elongando, nucleótido a nucleótido, produciéndose así una secuencia exactamente complementaria a la de la hebra de ADN utilizada como molde.

Las enzimas que llevan a cabo la transcripción reciben el nombre de *ARN polimerasas*. Estas ARN polimerasas catalizan la formación de los enlaces fosfodiéster que unen los nucleótidos formando una cadena lineal. La ARN polimerasa se desplaza, paso a paso, a lo largo del ADN, desenrollando la hélice de ADN ligeramente por delante del centro activo de polimerización, exponiendo una nueva región de la hebra molde para el apareamiento complementario de bases.

Es importante destacar que en células eucariotas, las ARN polimerasas necesitan proteínas adicionales para poder iniciar la transcripción. A dichas



(a) Al desenrollar la doble hélice se produce un lazo superenrollado que compensa la tensión generada.

(b) El desplazamiento de las ARN polimerasas produce una tensión superhelicoidal.

Figura 2.11: Tensiones generadas en el ADN durante la transcripción.(Figuras obtenidas de la referencia [11])

proteínas se les llama *factores generales de transcripción*. Además, las polimerasas deben salvar otra dificultad durante la elongación: el *superenrollamiento del ADN*. El desenrollamiento de una secuencia de nucleótidos generará un gran lazo de ADN superenrollado para compensar la tensión generada (Figura 2.11a). Además, cualquier proteína que se impulsa a sí misma a lo largo de un fragmento de la doble hélice de ADN tiende a generar una tensión superhelicoidal (Figura 2.11b). En las células eucariotas, las topoisomerasas de ADN eliminan esta tensión.

Finalmente, tal y como se comentó anteriormente, las secuencias codificadoras de los genes suelen estar interrumpidas por secuencias no codificadoras. Tanto los intrones como los exones son transcritos a ARN, por lo que los intrones tienen que ser eliminados de la molécula de ARN recién sintetizada mediante maduración por *corte y empalme del ARN* ó *ayuste del ARN*.

La traducción

La mayoría de los genes especifican la secuencia de aminoácidos de las proteínas. Las moléculas de ARN copiadas a partir de estos genes (que son las responsables, en último término, de la síntesis de proteínas) reciben el nombre de moléculas de *ARN mensajero (mARN)*. Sin embargo, en una pequeña

proporción de genes el producto final es el propio ARN. Hay otros tipos de ARN, tales como los *ARN pequeños nucleares* (snRNA, de small nuclear RNA), el *ARN ribosómico* y el *ARN de transferencia*. Los primeros dirigen la maduración del pre-mARN formando el mARN. El ARN ribosómico forma el núcleo de los ribosomas. Finalmente, las moléculas de ARN de transferencia son los adaptadores que seleccionan los aminoácidos y los retienen en el lugar adecuado del ribosoma para que se puedan incorporar las proteínas.

La traducción del mARN a proteína se lleva a cabo en los ribosomas. Por tanto, una vez terminada la traducción y maduración del mARN correspondiente, éste es trasladado al citoplasma. Una vez en el ribosoma, la información contenida en la secuencia del mARN se lee en grupos de tres nucleótidos: cada triplete de nucleótidos o *codón* codifica un aminoácido de la proteína. La secuencia de nucleótidos del mARN es leída por moléculas de ARN de transferencia. Cada molécula de tARN une en uno de sus extremos un aminoácido y en su otro extremo tiene una secuencia específica de tres nucleótidos -un *anticodón*- que le permite reconocer por apareamiento de bases un codón o un subgrupo de codones del mARN (Figura 2.12). Así, para la síntesis de la proteína, un conjunto de moléculas de tARN cargadas con sus aminoácidos respectivos se unen al mARN por apareamiento de sus anticodones con cada uno de los codones sucesivos del mARN. Posteriormente, los aminoácidos se van uniendo de forma que la proteína naciente va creciendo y cada tARN, relegado de su carga, se libera.

La clave para la traducción es el *código genético*, que relaciona los aminoácidos específicos con combinaciones de tres bases adyacentes a lo largo del mARN. Dado que en cada posición existen cuatro posibilidades (A, T, C ó G), hay $4^3 = 64$ combinaciones de tripletes posibles. Estos 64 codones constituyen el código genético (Tabla 3.1). Puesto que sólo existen 20 aminoácidos, necesariamente hay muchos casos en los que varios codones corresponden a un mismo aminoácido. Esta redundancia es una característica importante del código genético ya que, por ejemplo, un error en el código genético puede que sólo cause una mutación silenciosa ó un error que no afectará a la proteína correspondiente.

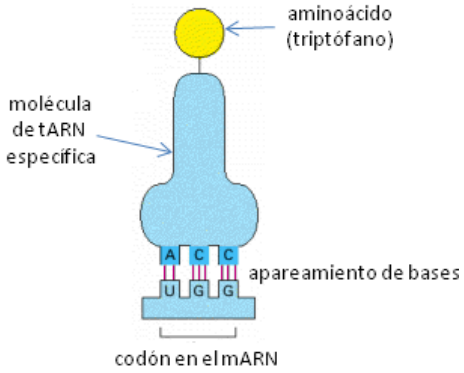


Figura 2.12: Una molécula de tARN específica para el aminoácido triptófano. Un extremo de la molécula se une al triptófano, mientras que el otro muestra la secuencia de nucleótidos CCA que reconoce el codón presente en el mARN (Figura obtenida de la referencia [11]).

Tabla 2.1: Código genético universal

AA	Codón	AA	Codón	AA	Codón	AA	Codón
Phe	UUU	Ser	UCU	Tyr	UAU	Cys	UGU
	UUC		UUC		UAC		UGC
Leu	UUA		UCA	TER	UUA	TER	UGA
	UUG		UCG		UAG	Trp	UGG
	CUU	Pro	CCU	His	CAU	Arg	CGU
	CUC		CCC		CAC		CGC
	CUA		CCA	Gln	CAA		CGA
	CUG		CCG		CAG		CGG
Ile	AUU	Thr	ACU	Asn	AAU	Ser	AGU
	AUC		ACC		AAC		AGC
	AUA		ACA	Lys	AAA	Arg	AGA
Met	AUG		ACG		AAG		AGG
Val	GUU	Ala	GCU	Asp	GAU	Gly	GGU
	GUC		GCC		GAC		GGC
	GUA		GCA	Glu	GAA		GGA
	GUG		GCG		GAG		GGG

Control de la expresión génica

El ADN de un organismo codifica todas las moléculas de ARN y de proteína necesarias para la construcción de sus células. Sin embargo, la descripción completa de la secuencia de ADN de un organismo no permitiría reconstruirlo. El problema radica en saber cómo se utilizan los elementos de la secuencia de ADN, es decir, las reglas y mecanismos que determinan que en cada célula sólo se exprese específicamente una selección de los genes.

Tal y como ya se ha comentado, la vía que conduce del ADN hasta las proteínas está formada por muchas etapas (transcripción, maduración, transporte al citosol, etc.), y en principio todas ellas se pueden regular. Sin embargo, para la mayoría de los genes, los controles transcripcionales son los más importantes. Esto tiene sentido ya que, de todos los posibles puntos de control solamente el control transcripcional asegura que no se sinteticen intermediarios superfluos.

La transcripción de cada uno de los genes de las células es activada o desactivada por proteínas de regulación génica denominadas *factores de transcripción* (TFs, de Transcription Factors en inglés). Es importante no confundir estas proteínas con los factores generales de transcripción antes mencionados. El término "general" se refiere al hecho de que éstos últimos se ensamblan en todos los promotores transcritos por la ARN polimerasa II, en lo cual difieren las proteínas de regulación génica ó *factores de transcripción*, que sólo actúan en determinados genes. Estas proteínas se unen a secuencias reguladoras del ADN, producen alteraciones locales en la estructura de la cromatina y facilitan/entorpecen el ensamblaje de la ARN polimerasa y de los factores generales de transcripción en la posición de inicio de la transcripción. Sería lógico preguntarse si estas uniones DNA-proteína vienen determinadas por un código de apareamiento aminoácido-par de bases sencillo: por ejemplo ¿es siempre el par de bases G-C el que se une a un cierto aminoácido? La respuesta parece ser que no, aunque algunos tipos de interacción aminoácido-base aparecen con mucha más frecuencia que otros.

Una característica particular de la regulación de la transcripción en células eucariotas, es que muchas proteínas reguladoras pueden actuar incluso cuando se encuentran unidas al ADN a miles de nucleótidos de distancia del

promotor que regulan. Esto significa que un determinado promotor puede estar controlado por un número casi ilimitado de secuencias reguladoras dispersas por el ADN. Esta acción a distancia es muy común. Aquí se utiliza el término *región de control de un gen* para indicar la región de ADN implicada en la regulación de la transcripción de un gen, incluido el *promotor*, en que se ensamblan los factores generales de transcripción y la polimerasa, más todas las *secuencias reguladoras* a las que se unen las proteínas reguladoras controlando la velocidad de este proceso de ensamblaje sobre el promotor. En eucariotas superiores no es inusual encontrar las secuencias reguladoras de un gen dispersas a distancias tan largas como 50000 pares de nucleótidos. Aunque la mayor parte de este ADN actúa como "espaciador" y no es reconocido por proteínas reguladoras, este ADN espaciador puede facilitar la transcripción aportando la flexibilidad necesaria para la comunicación entre las proteínas unidas al ADN. De hecho, el modelo más correcto que explica este control "a distancia" describe que el ADN existente entre la secuencia reguladora y el promotor se doblaría, permitiendo que las proteínas unidas a la secuencia reguladora interactúen con proteínas (ARN polimerasa, factores generales de transcripción, u otras proteínas) unidas al promotor.

Finalmente, es importante destacar que las proteínas reguladoras de genes eucariotas no suelen actuar como polipéptidos individuales. En general, estas proteínas reguladoras se unen de forma coordinada a los llamados *módulos de regulación* (CRMs, del inglés *Cis-Regulatory Modules*). Así, las regiones reguladoras de los genes se encuentran organizadas en una serie de módulos de regulación. Se piensa que estos módulos abarcan varios cientos de pares de bases [17]. Los resultados obtenidos con diversas técnicas computacionales indican que estos módulos tenderían a estar formados por $2-5$ sitios de unión de factores de transcripción (TFBSs, del inglés *Transcription Factor Binding Sites*) [212, 128].

2.1.5 Estudio de la expresión de los genes: los microarrays

Los *microarrays* fueron desarrollados en los años noventa y han revolucionado la forma en la que se estudia la expresión génica, al permitir el estudio de la expresión de miles de genes simultáneamente. De hecho, los microarrays se han utilizado hasta el momento para estudiar una gran variedad de procesos, que incluyen desde los cambios en la expresión génica responsables de la maduración de las fresas, hasta las "firmas" de expresión génica de diferentes células cancerígenas humanas.

Los arrays de expresión contienen miles de secuencias de ADN complementario (ADNc) inmovilizadas en una pequeña superficie del tamaño de un portaobjetos de vidrio usado para microscopía. En cada cristal se pueden imprimir, en miles de *spots* o puntos diferentes insertos de clones de ADNc. Cada uno de los puntos sirve para determinar en qué medida se está expresando el gen al que representa.

Algunos microarrays se generan a partir de fragmentos de ADN que un robot se encarga de disponer sobre el portaobjetos. Otros contienen oligonucleótidos cortos que se sintetizan sobre la superficie de la lámina de vidrio con técnicas similares a las utilizadas para grabar los circuitos de los chips de los ordenadores. En cualquier caso, debido a que se conocen la secuencia y la posición exactas de cada sonda en el microarray, cualquier fragmento de nucleótidos que se hibride con una de las sondas podrá ser identificado como un gen concreto simplemente detectando la posición en la que se encuentre unido en el microchip.

Para utilizar un microarray de expresión, se extrae en primer lugar el mARN de las células que se van a estudiar y se transforma en ADNc. Éste se marca con un producto fluorescente, el microchip se incuba con esta muestra de ADNc marcado y se deja que tenga lugar la reacción de hibridación. El microarray se lava posteriormente para eliminar el ADNc que no se haya unido, y se identifican las posiciones del microarray a las que se han unido los fragmentos de ADN marcado con un microscopio que tiene un escáner láser automatizado. Las posiciones detectadas por el escáner se hacen corresponder con el gen concreto que se había dispuesto inicialmente en esa posición.

En el caso de los arrays de oligonucleótidos el proceso es algo diferente, ya que el ADNc se transforma en ARNc mediante transcripción *in vitro*, siendo este ARNc fragmentado el que se utiliza en la reacción de hibridación. En cualquier caso, el principio básico de funcionamiento del microarray es el mismo. La Figura 2.13 muestra esquemáticamente todo el proceso.

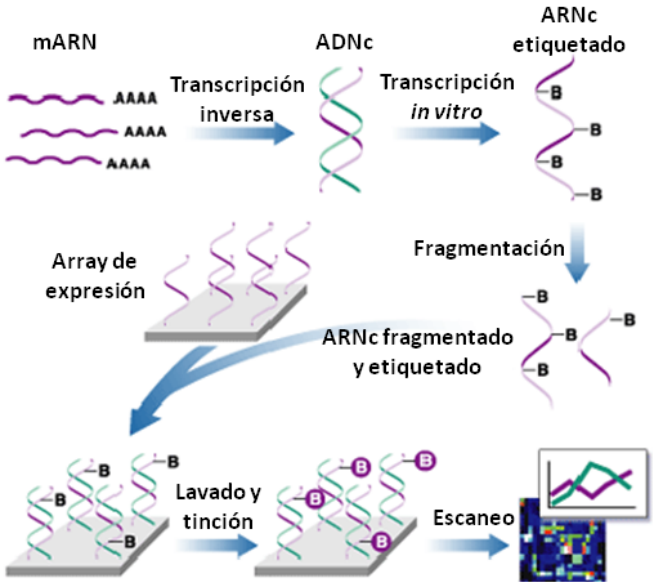


Figura 2.13: Uso de un microarray de oligonucleótidos

2.2 Técnicas computacionales para el análisis de información genética

La secuenciación de los genomas de distintas especies, así como el desarrollo de nuevas tecnologías como los microarrays, han generado una ingente cantidad de información genómica estructural y funcional, impulsando la investigación en Bioinformática hacia el desarrollo de técnicas computacionales capaces de analizar tal cantidad de datos [143, 141]. De esta forma, las técnicas computacionales de minería de datos se han aplicado a diferentes áreas de la Bioinformática. Un ejemplo típico de aplicación es el estudio

de cadenas (bien sean de nucleótidos o de aminoácidos): búsqueda de patrones y alineamiento de múltiples secuencias. En este campo, la aplicación de algoritmos de programación dinámica ha proporcionado excelentes resultados, dando lugar a métodos como BLAST o FASTA que han revolucionado las aplicaciones bionformáticas en Biología Molecular [19, 166]. Otro campo de aplicación fundamental en el análisis de secuencias es la identificación de "elementos" de interés en secuencias biológicas, ya sean sitios de unión de proteínas, regiones de ADN sujetas a más presión selectiva, secuencias codificadoras etc. En este área, la herramienta computacional predominante ha sido, desde su inicio, los modelos ocultos de Markov [65, 226, 185]. Esta misma técnica ha sido utilizada asimismo en otros ámbitos como por ejemplo la predicción de estructuras de proteínas [45]. En este mismo campo se han aplicado además otras técnicas computacionales como las redes neuronales ó las máquinas de vector soporte (SVMs) [211, 124]. Estrechamente relacionado con las estructuras de proteínas se encuentra el análisis de las redes de interacción proteína-proteína, campo en el que también se han aplicado técnicas computacionales tales como el clustering basado en grafos o las redes bayesianas [236, 137]. Las redes de regulación genética han sido objeto también de gran cantidad de análisis *in silico*, desarrollándose para su estudio una ingente cantidad de estrategias computacionales [269].

Tal y como se puede ver, existe una amplia variedad tanto de campos de aplicación, como de técnicas computacionales útiles en cada campo. Merece una especial mención el uso de métodos de minería de datos para el análisis de microarrays. De hecho, muchos de los trabajos anteriormente mencionados requieren el análisis de datos de expresión para sus propósitos. Las herramientas computacionales por excelencia en este campo son los métodos de clasificación, clustering y biclustering, siendo los algoritmos de clústering los más populares en entornos no estrechamente relacionados con las ciencias de la computación [139]. Asimismo, las técnicas de extracción de reglas de asociación se han utilizado con éxito en prácticamente todos los campos mencionados, pudiendo así encontrarlas en una amplia variedad de trabajos que comprenden desde aplicaciones puras de minería de datos a la inferencia de rutas de señalización, predicción de interacciones proteína-proteína

o el estudio de módulos de regulación [163, 52, 27, 176]. Por todo esto, y dado que tanto el análisis de microarrays como la aplicación de técnicas de extracción de reglas de asociación son de especial interés en este trabajo, a continuación se dedican un par de secciones a estos dos campos. Además, y dado que a lo largo de todo el trabajo se hará uso de técnicas difusas, se proporciona seguidamente una breve introducción a los conceptos básicos de teoría de conjuntos difusos. Finalmente, se muestran las principales técnicas computacionales descritas hasta el momento para la detección de TFBSs en secuencias de ADN, así como para el estudio de los módulos de regulación.

2.2.1 Conjuntos difusos

La información con la que se trabaja diariamente no siempre presenta el grado de perfección que caracteriza a los modelos matemáticos que se utilizan para su tratamiento automático. En muchas ocasiones, y aún más frecuentemente en el caso de información biológica, ésta tiende a presentar un alto grado de imprecisión e incertidumbre. En general, la imprecisión de una variable se asocia con la incapacidad de asignarle un valor preciso a la misma. Es decir, el valor de la variable se encuentra en un rango de valores pero no se puede determinar cuál es, debido por ejemplo a la presencia de errores en la medida. Por otra parte, la incertidumbre viene dada por el hecho de que no exista certeza de que la información sea verdadera, algo también frecuente en datos biológicos. En algunas ocasiones la incertidumbre puede venir dada por la propia imprecisión.

Todo esto ha hecho que a lo largo de los años se hayan desarrollado diversos modelos matemáticos para modelar la información imperfecta, uno de los cuales es la teoría de conjuntos difusos [290]. Dicha teoría fue propuesta por Zadeh en 1965, y desde entonces ha sido objeto de estudio por una gran cantidad de investigadores. De este modo, los conjuntos difusos han resultado ser especialmente adecuados para modelar conceptos e información imprecisa.

Definición de conjunto difuso

De manera informal, la teoría de conjuntos difusos permite que un objeto pertenezca a un conjunto con un grado de pertenencia entre 0 y 1. Así, la teoría de conjuntos clásica es un caso especial de la difusa en el que los grados de pertenencia se restringen a los valores 0 y 1.

Formalmente, sea X un conjunto de objetos no vacío considerado el *universo* de estudio. Un *conjunto difuso* es un par (X,A), donde $A : X \to I$ y $I = [0,1]$. A la función A se le denomina *función de pertenencia*. Si se considera el universo X fijo, se puede identificar el conjunto difuso mediante su función de pertenencia. De esta forma, $A(x)$ representa el grado de pertenencia de x al conjunto difuso A.

Dos conjuntos difusos A y B se dicen iguales si:

$$A = B \iff A(x) = B(x),\ \forall x \in X$$

La relación de inclusión viene dada por:

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

Los conjuntos difusos suelen representarse mediante etiquetas lingüísticas. Por ejemplo, véase la Figura 2.14 en la que se han definido las funciones de pertenencia de tres conjuntos difusos *bajo*, *medio* y *alto* sobre el dominio de la altura de las personas.

Operaciones de intersección: *t-normas*

Una *t-norma* es una función de dos argumentos de la forma:

$$T : I \times I \to I$$

,

que cumple las siguientes propiedades:

- $T(a,1) = a,\ \forall a \in I$
- $T(a,b) \leq T(u,v)$ si $a \leq u,\ b \leq v$,
- $T(a,b) = T(b,a)$,

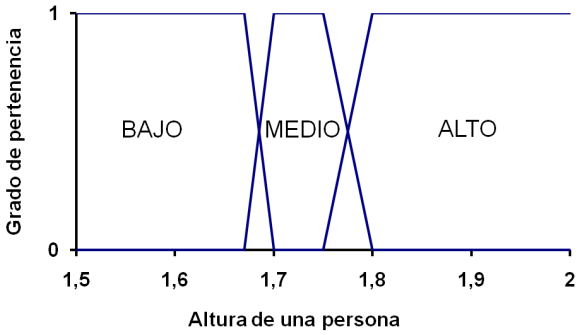


Figura 2.14: Se definen tres conjuntos difusos sobre el dominio de la altura de las personas.

- $T(T(a,b),c) = T(a,T(b,c))$

Por tanto, la intersección de dos conjuntos difusos mediante una *t*-norma T viene dada por:

$$(A \cap B)(x) = T(A(x), B(x))$$

Algunas de las *t*-normas más utilizadas son:

$minimo(a,b) = min(a,b)$,
$producto(a,b) = a \cdot b$,
$t-norma\ de\ Lukasiewicz(a,b) = max(a+b-1,0)$

Operaciones de unión: *t-conormas*

Una *t-conorma* es una función de dos argumentos de la forma:

$$S : I \times I \to I$$

que cumple las siguientes propiedades:

- $S(a,1) = a,\ \forall a \in I$
- $S(a,b) \leq S(u,v)$ si $a \leq u,\ b \leq v$,
- $S(a,b) = S(b,a)$,

- $S(T(a,b),c) = S(a,S(b,c))$

Por tanto, la unión de dos conjuntos difusos mediante una t-conorma S viene dada por:

$$(A \cup B)(x) = S(A(x), B(x))$$

Algunas de las t-conormas más utilizadas son:

$maximo(a,b) = max(a,b)$,
$suma\ algebraica(a,b) = a + b - ab$,
$suma\ acotada\ de\ Lukasiewicz(a,b) = min(1, a+b)$

Operadores de complemento: negación

Una función $C : I \to I$ es una *función complemento* (o *negación*) si y sólo si satisface las siguientes propiedades:

- C es una *involución*. Es decir, dado un conjunto difuso A, $C(C(A(x))) = A(x)$,
- C es estrictamente decreciente,
- C es continua,
- $C(0) = 1$ y $C(1) = 0$.

La negación más utilizada es la llamada función de negación estándar:

$$C(A(x)) = 1 - A(x)$$

2.2.2 Análisis de microarrays

Los resultados de un conjunto de experimentos con microarrays se organizan en las llamadas *matrices de expresión génica*. Las filas (columnas) en estas matrices representan los genes y las columnas (filas) las condiciones experimentales que se estudian. De este modo, cada posición de la matriz contiene el nivel de expresión de un gen bajo una cierta condición experimental. El

análisis de estas matrices suele resultar realmente complejo, dada la gran cantidad de genes (incluso en el organismo más simple), el alto número de condiciones experimentales que se pueden llegar a estudiar y el ruido que afecta a todo el proceso experimental.

En todos los casos es imprescindible realizar una etapa previa de preprocesamiento de la matriz de expresión. Una vez preprocesada, se aplican sobre ella las técnicas de análisis correspondientes. Aunque en algunos casos se utilizan técnicas estadísticas para su análisis, algunas de ellas muy populares como SAM (*Significance Analysis of Microarrays*) [262], en la mayoría de las ocasiones la cantidad de muestras diferentes en el estudio hace tremendamente engorrosa la tarea de análisis mediante estas técnicas. Por ello, se hace necesaria la aplicación de técnicas de minería de datos y, en particular, los algoritmos de clustering son especialmente útiles en este campo. Asimismo, la extracción de reglas de asociación se ha aplicado también satisfactoriamente en gran cantidad de ocasiones para el análisis de microarrays. Dado que estas dos técnicas son de especial interés en este trabajo, se centrará la atención en los principales usos propuestos de estas dos técnicas para el análisis de datos de expresión.

Preprocesamiento

El preprocesamiento de los datos de expresión incluye por lo general diversas tareas, tales como la normalización, relativización de los valores de expresión, tratamiento de datos perdidos y eliminación de datos planos. Los conjuntos de datos de expresión utilizados en este trabajo no presentaban datos perdidos, por lo que se centrará la atención en el resto de tareas.

Normalización

Tal y como ya se ha comentado, el proceso de preparación de los microarrays está compuesto por una serie de tareas complejas sujetas a un error. Este error puede enmascarar el valor biológico en el que se está interesado. La *normalización* es un paso esencial del preprocesamiento para corregir los efectos de las fuentes sistemáticas de variación de origen no biológico. Se

han propuesto una gran cantidad de metodologías de normalización, muchas de ellas dependientes de la plataforma de microarrays utilizada. En este trabajo, se han utilizado dos conjuntos de microarrays: uno descargado de la *Saccharomyces Cerevisae Database* (SGD [84, 6], Sección 3.2.4) y otro proporcionado por el Hospital Universitario Virgen de las Nieves de Granada (Sección 4.3.3). El primero de ellos se descargó ya preprocesado, por lo que no requirió de ninguna acción previa a su análisis. El segundo conjunto fue obtenido mediante la plataforma de arrays Affymetrix Genechip U133 plus 2.0. Existe toda una serie de metodologías de normalización específicas para esta plataforma, por lo que se centrará la atención en ellas [140].

Cada gen en un chip de Affymetrix de este tipo está representado por un conjunto de sondas. Dicho conjunto está formado por 11 pares diferentes de oligos de 25 pb cada uno, que cubren ciertas zonas de la región que se transcribe del gen. Cada par está formado por un oligonucleótido *perfect match* (PM) y otro *mistmatch* (MM). La sonda PM se ajusta exactamente a la secuencia de un genotipo estándar, mientras que la sonda MM difiere en una posición central, la base 13. Esta segunda sonda (MM) está diseñada para distinguir el ruido provocado por la hibridación no-específica de la señal de hibridación específica.

Teniendo en cuenta todo lo anterior, los datos obtenidos con esta plataforma se normalizan en tres etapas: *corrección de fondo*, *normalización* y *agregación*. La *corrección de fondo* permite ajustar los efectos de hibridación no relacionados con la interacción entre sondas y ADN objetivo. Mediante la *normalización* se eliminan errores sistemáticos, de forma que se puedan comparar valores de expresión de distintos arrays. Finalmente, en la *agregación* se combinan las intensidades de las diferentes sondas del conjunto de sondas, de forma que se proporcione un sólo valor de expresión para cada gen.

El primer método de normalización que surgió, establecido inicalmente como el método por defecto para Affymetrix, se denominó *average difference* (AD). Se trata una medida lineal basada en la diferencia PM-MM para corregir el efecto de las uniones no específicas. Esta medida se reemplazó por el actual estándar MAS5.0, el cual utiliza una escala logarítmica más

apropiada y el método de regresión robusta Tukey Biweight [1]. Posteriormente, se demostró que un tercio de los pares de sondas PM-MM dan lugar a señales negativas, quedando patente así que el uso de las sondas MM para la detección de uniones no específicas no es fiable [131, 179]. En este sentido, Irizarry et al. [130] desarrollaron el *robust multi-array average method* (RMA), basado únicamente en los valores proporcionados por las sondas PM. Li y Wong [160] desarrollaron un modelo estadístico sobre los datos para calcular un índice de expresión (MBEI), incorporando dicho índice en la herramienta dChip [159], una de las aplicaciones software más conocidas en este área. Se han propuesto también algunos modelos basados en la energía física de las interacciones, en un intento por modelar la formación de los pares ADN-ARN [240], entre los que cabe destacar el *Positional Dependent Nearest Neighbour* (PDNN) de Zhang et al. [295]. Siguiendo esta idea, Wu et al. [282] desarrollaron el método GCRMA, que intenta combinar la fuerza de algoritmos basados en modelos estocásticos, como el RMA, con el modelado físico de la información de la secuencia. El número de métodos continúa creciendo, no habiendo aún un consenso acerca de qué método es más apropiado y fiable en cada ocasión.

Transformación de valores absolutos de expresión en relativos

En general, no se suele trabajar con valores de expresión absolutos sino relativos. Es decir, los valores de expresión absolutos obtenidos en las muestras objetivo, se comparan con los valores de expresión obtenidos en un conjunto de muestras de referencia, de tal modo que lo que finalmente se obtiene es una medida del cambio de la expresión de cada gen en la muestra objetivo respecto a la muestra de referencia. Los valores relativos se calculan dividiendo cada valor en la muestra objetivo por el valor en las muestras de referencia, y calculando el logaritmo en base 2 del resultado:

$$Expr'_{i,j} = \log_2(Expr_{i,j}/valor\_ref_i),$$

donde:

- $Expr'_{i,j}$ representa el nivel relativo de expresión del gen i en la muestra j,

- $Expr_{i,j}$ representa el nivel absoluto de expresión del gen i en la muestra j,
- $valor\_ref_i$ representa el valor de expresión de referencia para el gen i. Este valor se obtiene normalmente agregando los niveles de expresión del gen correspondiente en diferentes muestras de referencia. Dicha agregación se puede llevar a cabo calculando la mediana o el valor medio.

Al dividir cada valor absoluto por el valor de referencia, aparecerán valores superiores a 1, que indican que los genes están *sobre-expresados*, es decir, producen un nivel de expresión mayor en la muestra experimento, que en la de referencia; por otra parte, aparecerán valores inferiores a 1, indicando que los genes están *sub-expresados*, en el sentido de que el nivel de expresión es menor en la muestra del experimento que en el de referencia. Por esta razón, las dos clases de datos anteriormente indicadas no están representadas en la misma escala, ya que en el caso de la sub-expresión, serán valores que pertenecen al intervalo $]0,1[$, mientras que en el caso de la sobre-expresión tomarán valores en una rango mayor $]1,\infty[$. Por tanto, es necesaria una transformación de los datos, para poder manejar conjuntamente los valores de sobre-expresión y de sub-expresión. Usualmente se elige una transformación logarítmica en base 2. La idea fundamental es que los datos de sub-expresión se concentran en $[0,1]$ y con la transformación logarítmica en base 2, los datos se distribuyen de forma visual más realista, representando los valores positivos, la sobre-expresión, y los negativos la sub-expresión.

A continuación se muestra un ejemplo para tratar de aclarar este proceso. Supónganse los datos absolutos (normalizados) de la Tabla 2.2. Las tres primeras columnas representan muestras de referencia, mientras que el resto se corresponden con muestras objetivo. Se obtienen en primer lugar los niveles de expresión de referencia para cada gen: para esto se calcula la mediana de las tres primeras columnas de cada fila (Tabla 2.3).

A continuación, se dividen los niveles de expresión de cada gen en cada muestra objetivo por sus correspondientes valores de referencia (Tabla 2.4). Tal y como se puede ver, los valores que representan sub-expresión se concen-

Tabla 2.2: Ejemplo de datos de expresión.

Gen	Ref1	Ref2	Ref3	Obj1	Obj2	Obj3	Obj4
$DDR1$	10,002	10,126	10,377	9,332	9,640	10,511	9,030
$RFC2$	7,393	6,483	8,215	6,697	6,677	6,971	6,970
$HSPA6$	6,428	5,659	7,692	6,866	5,957	5,488	5,973
$PAX8$	7,592	8,744	8,707	7,842	7,926	7,219	8,365
$GUCA1A$	3,363	3,291	3,723	3,181	3,212	3,138	3,128
$UBA7$	9,356	9,722	9,424	7,783	8,216	7,808	7,793
$THRA$	5,276	7,512	5,790	5,324	4,991	5,028	5,279
$PTPN21$	4,142	4,030	4,389	4,631	4,314	4,222	4,633
$CCL5$	10,357	10,123	10,311	7,509	8,769	6,444	8,247

Tabla 2.3: Ejemplo de datos de expresión.

Gen	Ref	Obj1	Obj2	Obj3	Obj4
$DDR1$	10,126	10,126	9,408	10,297	9,577
$RFC2$	7,393	6,653	6,423	7,662	6,579
$HSPA6$	6,428	26,351	5,620	6,275	6,299
$PAX8$	8,707	18,025	7,652	7,977	17,814
$GUCA1A$	3,363	3,533	3,611	3,540	3,374
$UBA7$	9,424	7,979	7,023	7,453	7,552
$THRA$	5,790	6,526	5,297	6,455	6,277
$PTPN21$	4,142	3,959	4,341	4,000	3,977
$CCL5$	10,311	9,405	7,831	8,308	7,792

tran en el intervalo $]0,1[$, mientras que los que representan sobre-expresión toman valores mayores que 1.

Tabla 2.4: Ejemplo de datos de expresión relativos.

Gen	Obj1	Obj2	Obj3	Obj4
$DDR1$	1,000	0,929	1,017	0,946
$RFC2$	0,900	0,869	1,036	0,890
$HSPA6$	4,099	0,874	0,976	0,980
$PAX8$	2,070	0,879	0,916	2,046
$GUCA1A$	1,050	1,074	1,052	1,003
$UBA7$	0,847	0,745	0,791	0,801
$THRA$	1,127	0,915	1,115	1,084
$PTPN21$	0,956	1,048	0,966	0,960
$CCL5$	0,912	0,759	0,806	0,756

Finalmente, se calcula el logaritmo en base 2 de cada valor de la tabla, trasladando así los valores sub-expresados al rango de los valores reales negativos, y los niveles sobre-expresados al rango de los valores reales positivos (Tabla 2.5).

Eliminación de los datos planos

A veces aparecen datos planos en la matriz de expresión, es decir, genes que prácticamente no modifican su expresión (cercanos al cero) a lo largo de todos los experimentos. Estos genes no aportan información y dificultan el análisis posterior. El procedimiento general consiste en eliminar directamente aquellos genes que no presenten un número mínimo de "picos". Por ejemplo, eliminar aquellos genes cuyo nivel de expresión relativo no sobrepase los límites del intervalo $]-1,1[$ en al menos el 10 % de los experimentos.

Tabla 2.5: Ejemplo de datos de expresión tras aplicar log_2.

Gen	Obj1	Obj2	Obj3	Obj4
$DDR1$	$0,000$	$-0,106$	$0,024$	$-0,080$
$RFC2$	$-0,152$	$-0,203$	$0,052$	$-0,168$
$HSPA6$	$2,035$	$-0,194$	$-0,035$	$-0,029$
$PAX8$	$1,050$	$-0,186$	$-0,126$	$1,033$
$GUCA1A$	$0,071$	$0,103$	$0,074$	$0,005$
$UBA7$	$-0,240$	$-0,424$	$-0,339$	$-0,320$
$THRA$	$0,173$	$-0,128$	$0,157$	$0,117$
$PTPN21$	$-0,065$	$0,068$	$-0,050$	$-0,059$
$CCL5$	$-0,133$	$-0,397$	$-0,312$	$-0,404$

Clustering y biclustering para el análisis de datos de microarrays

Al aplicar un algoritmo de clustering sobre una matriz de expresión, se pretende obtener grupos de genes que presenten valores de expresión similares a lo largo de las condiciones experimentales estudiadas (genes co-expresados). Así, los genes de un mismo clúster se comportan de manera similar en diferentes circunstancias y, por tanto, probablemente lleven a cabo una función común. Además, los algoritmos de clustering se pueden utilizar también para agrupar condiciones experimentales con perfiles genéticos similares, de forma que pudieran observarse subgrupos desconocidos dentro del conjunto de pacientes/experimentos estudiados. De este modo, los objetivos que se persiguen al realizar este tipo de análisis son variados: anotación funcional de genes, diagnóstico de enfermedades, descubrimiento de genes, desarrollo de medicamentos, detección de subtipos tumorales, etc.

Eisen et al. [86] fueron los primeros investigadores que mostraron el potencial de los algoritmos de clustering para extraer patrones de datos de microarrays. Aplicaron un algoritmo de clustering jerárquico para identificar grupos funcionales de genes. Posteriormente, se desarrollaron otras muchas propuestas para el análisis de microarrays mediante clustering: Smet et al.

[74], Tamayo et al. [243], etc. (un listado más extenso y detallado puede encontrarse en [139]).

Sin embargo, todas estas propuestas agrupan los genes en clústers exclusivos, es decir, un gen puede pertenecer a un sólo clúster. Esto no representa adecuadamente la realidad, ya que un gen puede desempeñar funciones distintas en procesos biológicos diferentes. De esta forma, se han propuesto una serie de estrategias que tratan de resolver este problema, entre las que cabe destacar el algoritmo *Gene Shaving* [114]. Este algoritmo se ha convertido en uno de los más utilizados para el análisis de datos de expresión. La principal ventaja que presenta consiste en que, no sólo busca genes con patrones de expresión similares, sino que trata también de maximizar la varianza entre condiciones experimentales. Por tanto, los resultados muestran genes que presentan comportamientos muy diferentes a lo largo de las distintas condiciones experimentales, ignorando aquellos genes que no participan en procesos activos, así como aquellos que participan en procesos activos permanentemente. De esta forma, los clústers obtenidos con esta técnica son muy útiles para identificar los distintos tipos de muestras y procesos biológicos que producen estas diferencias.

Los algoritmos de clustering difuso han recibido también una especial atención en este campo, dada su capacidad de asignar un gen a distintos clústers con diferentes grados de pertenencia, permitiendo así superar el problema de la "exclusividad" que se planteaba anteriormente. De esta forma, han surgido métodos como el k-medias difuso [77], versiones difusas de las redes SOM (del inglés *Self Organized Maps*) [190], familias de algoritmos de clustering basados en mezclas de modelos gausianos (GMM) [205], etc.

Sin embargo, en la matriz de expresión pueden aparecer decenas de condiciones experimentales heterogéneas, lo que hace que buscar genes que se comporten de forma similar para todas ellas no tenga sentido. Es más, probablemente se obtengan patrones de expresión carentes de interés y se pierdan los más relevantes. Así, los algoritmos de Biclustering han resultado ser una herramienta apropiada para solventar esta limitación [112], ya que permiten identificar grupos de genes co-expresados bajo un subconjunto de muestras

(y no necesariamente todas).

Biclustering

Sean $F = \{1, .., n\}$ y $C = \{1, ..., m\}$ los conjuntos de filas y columnas de una matriz de expresión $A_{n\times m}$. Un biclúster B se define como una submatriz $B = A_{IJ}$ de $A_{n\times m}$, cuyos valores se interrelacionan de acuerdo a un cierto criterio. Encontrar los biclústers de tamaño máximo en una matriz es un problema NP-completo [195], por lo que la mayoría de los algoritmos de biclustering utilizan enfoques heurísticos [167].

El concepto de biclúster fue propuesto por Hartigan en 1975 [112] y aplicado por primera vez al análisis de microarrays por Cheng y Church [60]. Desde entonces, se han desarrollado una gran cantidad de propuestas (un listado más extenso puede encontrarse en [167]). Yang y Wang propusieron el δ-cluster [287], el $p-\delta$-cluster [286] y FLOC [285]. Otros métodos (SAMBA [246], Kluger et al. [149], Cano et al. [48]) modelan la matriz de expresión como un grafo bipartito cuyas dos partes se corresponden con los genes y las condiciones respectivamente, y tratan de dividir el grafo mediante el uso de técnicas estadísticas [246] o espectrales [149, 48]. Lazzeroni y Owen [153] introdujeron el concepto de modelo PLAID, donde la matriz se describe como la suma de diferentes capas que corresponden a los biclústers. Algunos otros métodos intentan mejorar este modelo [260, 261]. Liu y Wang [161] diseñaron un algoritmo de tiempo polinomial para encontrar biclústers cuadrados óptimos con un grado de similitud máximo. Aguilar y Divina [9] propusieron un algoritmo genético para identificar biclusters coherentes no solapados con máxima varianza entre condiciones. En este caso definen una medida de ajuste combinando el residuo cuadrado medio (MSR) [60] y la varianza entre filas del biclúster candidato.

2.2.3 Reglas de asociación

En 1993 Agrawal propuso un algoritmo para la extracción de reglas de asociación de grandes bases de datos [8]. La aplicación inicial de los análisis de asociación consistía en el estudio de relaciones ocultas en las llamadas *bases*

de datos de supermercado (en inglés, *market basket databases*). Estas bases de datos contienen información acerca de los productos que adquieren los clientes en cada compra. De esta forma, una base de datos de supermercado consiste en un conjunto de transacciones, cada una de las cuales contiene los items adquiridos en esa transacción (Tabla 2.6). El objetivo que se pretende al llevar a cabo un análisis de asociación sobre este tipo de bases de datos es obtener relaciones de la forma:

$$\{Leche\} \rightarrow \{Mantequilla\}$$

En esto consiste básicamente una regla de asociación y representa la expresión: *los clientes que compran leche también compran mantequilla*. Este tipo de información puede ser de gran interés para un supermercado ya que, por ejemplo, se puede conseguir un incremento en las ventas simplemente colocando ciertos productos juntos en los expositores.

Tabla 2.6: Ejemplo de una base de datos de supermercado.

ID Transacción	Items
1	Pan, Leche, Mantequilla
2	Cerveza, Huevos, Leche, Mantequilla, Fruta
3	Leche, Mantequilla
...	...

Las reglas de asociación se han aplicado con éxito en muchas otras áreas tales como web mining, publicidad, Bioinformática, etc. Desde que se propusieron por primera vez en 1993, la extracción de reglas de asociación se ha convertido en una de las principales técnicas para la extracción de conocimiento de bases de datos (KDD, del inglés *Knowledge Discovery in Databases*).

Reglas de asociación: definición formal

Sea $I = \{x_1, x_2, ..., x_n\}$ un conjunto de pares atributo-valor o *items*. Sea D una base de datos transaccional, en la que cada *transacción* es un conjunto

de items $T \subseteq I$. Una *regla de asociación* es una expresión de la forma $X \rightarrow Y$, donde X e Y son conjuntos de items (o *itemsets*) tales que $X \cap Y = \{\emptyset\}$. Al itemset X se le denomina *antecedente* de la regla, mientras que a Y se le denomina *consecuente*. Una regla de asociación como esta indica que si X ocurre, entonces es probable que también ocurra Y. A la probabilidad de que ocurra Y dado que ha ocurrido X se le llama *confianza* de la regla. A la probabilidad de que ocurran X e Y conjuntamente se le denomina *soporte* de la regla. Así, el objetivo de los algoritmos clásicos de minería de reglas de asociación, consiste en extraer todas las reglas con soporte y confianza mayores que ciertos umbrales fijados por el usuario.

Se dice que una transacción T *soporta* un itemset $X \subseteq I$, si $X \subseteq T$, es decir, T contiene todos los items de X. De esta forma, el *soporte* de un itemset X se calcula como el porcentaje de transacciones en la base de datos que soportan X, o, en otras palabras, el soporte de un itemset X indica la probabilidad de encontrar dicho itemset en la base de datos. Por lo tanto, el soporte de una regla $X \rightarrow Y$ se puede calcular como:

$$sop(X \rightarrow Y) = sop(X \cup Y),$$

mientras que la confianza se define como:

$$conf(X \rightarrow Y) = \frac{sop(X \rightarrow Y)}{sop(X)}$$

Finalmente, se dice que un itemset X es *frecuente* si su soporte es mayor que el umbral establecido por el usuario.

Por ejemplo, considérese la información de la Tabla 2.7, la cual contiene datos estructurales de un conjunto de genes de la levadura. Esta tabla se puede transformar fácilmente en una tabla transaccional en la que cada fila representa una transacción, y los atributos de cada columna forman los items de la transacción (Tabla 2.8). Sea el siguiente itemset Z:

$$Z = \{(Longitud\ gen = Medio), (Longitud\ intergenico = Medio), (Orientacion\ gen = Tandem)\}$$

Este itemset está soportado por las transacciones (genes) *YALO11W, YALO12W* y *YALO15C*. Hay 13 transacciones en total, por lo que $sop(Z) = 3/13 = 0{,}231$. Considérese ahora la regla de asociación:

$$R = \{(Longitud\ gen = Medio), (Longitud\ intergenico = Medio)\} \rightarrow \\ \rightarrow \{(Orientacion\ gen = Tandem)\}$$

El soporte de R viene dado por:

$$sop(R) = sop(Z) = 0{,}231,$$

y la confianza se puede calcular como:

$$\begin{aligned} conf(R) &= \frac{sop(R)}{sop\left(\left\{\begin{array}{c}(Longitud\ gen = Medio), \\ (Longitud\ intergenico = Medio)\end{array}\right\}\right)} \\ &= \frac{(3/13)}{(4/13)} = 0{,}75 \end{aligned}$$

En resumen, el proceso de minería de reglas de asociación se divide por lo general en dos fases:

- Encontrar el conjunto de itemsets frecuentes. La mayoría de las investigaciones se han centrado en esta etapa, ya que es la fase más costosa en términos computacionales. Implica la búsqueda de itemsets frecuentes entre todas las combinaciones que se puedan formar con los items de la tabla de datos, es decir, cada combinación posible de los items de I (Figura 2.15). Por lo tanto, hay $2^{|I|}$ combinaciones de items posibles, y dado que $|I|$ suele ser grande, la aplicación de técnicas de exploración simples no es factible.
- Obtener, a partir del conjunto de itemsets frecuentes, las reglas de asociación con confianza mayor que el umbral especificado por el usuario. Normalmente se genera un ingente número de reglas, muchas de ellas proporcionando información trivial o redundante. En este sentido, el modelo soporte/confianza ha demostrado ser insuficiente y, debido a esto, se han propuesto medidas de interés adicionales para reglas de asociación.

Tabla 2.7: Ejemplo de tabla de datos.

Gen	Longitud del gen	Longitud del intergénico	Orientación
YAL008W	*Corto*	*Corto*	*Tandem*
YAL003W	*Corto*	*Largo*	*Divergente*
YAL018C	*Medio*	*Largo*	*Divergente*
YAL002W	*Largo*	*Largo*	*Tandem*
YAL009W	*Medio*	*Corto*	*Divergente*
YAL010C	*Medio*	*Corto*	*Divergente*
YAL011W	*Medio*	*Medio*	*Tandem*
YAL012W	*Medio*	*Medio*	*Tandem*
YAL013W	*Medio*	*Medio*	*Divergente*
YAL015C	*Medio*	*Medio*	*Tandem*
YAL017W	*Largo*	*Largo*	*Divergente*
YAL019W	*Largo*	*Corto*	*Tandem*
YAL021C	*Largo*	*Medio*	*Tandem*

Tabla 2.8: Ejemplo de tabla transaccional.

ID Transacción	Items
YAL002W	{ *Longitud gen=Largo*, *Longitud intergenico=Largo*, *Orientación=Tandem*}
YAL003W	{ *Longitud gen=Corto*, *Longitud intergenico=Largo*, *Orientación=Divergente*}
YAL008W	{ *Longitud gen=Corto*, *Longitud intergenico=Largo*, *Orientación=Tandem*}
...	...

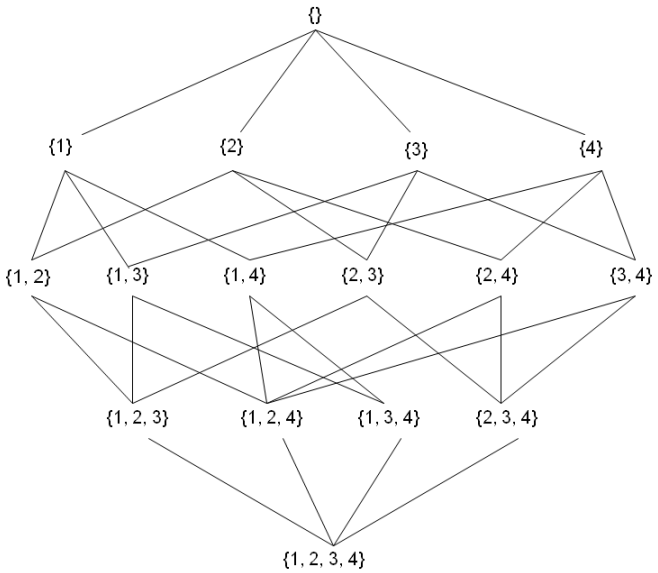


Figura 2.15: Ejemplo de espacio de búsqueda para cuatro items (numerados del 1 al 4)

Algoritmos de extracción de reglas de asociación

Se han propuesto un gran número de algoritmos para la extracción de reglas de asociación. Sin embargo, hasta el momento no existe ninguna implementación cuyo rendimiento mejore el resto de implementaciones sobre cualquier base de datos y con cualquier umbral de soporte [101].

Los algoritmos clásicos de minería de reglas de asociación se pueden dividir en dos grandes categorías que se corresponden con dos estrategias principales para la búsqueda de itemsets frecuentes: *generación de candidatos* y *crecimiento de patrones* (*candidate generation* y *pattern growth* en inglés). La mayoría de los algoritmos clásicos son del tipo generación de candidatos [8, 292, 123]. Este tipo de algoritmos genera conjuntos de itemsets *candidatos* que son posteriormente validados según las restricciones impuestas (por ejemplo, $soporte \geq umbral\_soporte\_minimo$). De este modo, la generación de nuevos itemsets candidatos se basa en un conjunto de itemsets previamente validados. Los principales algoritmos de este tipo son los populares Apriori y Eclat [8, 292].

Al contrario que los algoritmos de generación de candidatos, los métodos de crecimiento de patrones evitan generar estos candidatos intermedios mediante la construcción de estructuras de datos complejas, en las que almacenan de forma comprimida la información del conjunto de datos original. Una vez generada la estructura de datos, y siempre y cuando ésta quepa en memoria, no son necesarios más accesos a la base de datos. Se han propuesto varios algoritmos de este tipo, siendo el más conocido el *Frequent-Pattern Growth* (FP-Growth) [109, 196].

Las subdivisiones dentro de cada clase (generación de candidatos y crecimiento de patrones), se basan en la estrategia seguida por los algoritmos correspondientes para recorrer el espacio de búsqueda (primero en profundidad o primero en anchura) y en las diferentes estructuras de datos que utilizan (hash-trees, enumeration-set trees, prefix trees, FP-trees, H-struct, etc.).

Además de todo esto, se ha propuesto también otro tipo de algoritmos derivados de los clásicos. El objetivo de este otro tipo de algoritmos consiste en generar un conjunto resumido de reglas a partir del cual se pueda obtener el conjunto completo, optimizando así el proceso de búsqueda de itemsets válidos. El conjunto de reglas resultante es por lo tanto más pequeño que el conjunto completo, facilitándose de esta forma la interpretación de las mismas [191, 26, 46, 44].

Como resumen se puede concretar que se han propuesto una gran cantidad de algoritmos de extracción de reglas de asociación, siendo los principales Apriori, Eclat y FP-growth. En general, muchas de las implementaciones que se han propuesto después de estos se basan en Apriori y FP-growth. De hecho, Apriori es el algoritmo más conocido y el que normalmente se usa cuando se aplican las reglas de asociación en algún campo. Se han desarrollado muchas mejoras de este algoritmo, siendo las más eficientes las descritas en las referencias [39, 36]. Información más amplia y detallada de algoritmos de extracción de reglas de asociación puede encontrarse en las referencias [57, 245].

Medidas de interés de reglas de asociación

Como ya se ha comentado, se suele generar un ingente número de reglas, muchas de ellas proporcionando información trivial o redundante. El modelo de soporte/confianza ha demostrado ser insuficiente para tratar este problema. De este modo, se han propuesto estrategias y medidas de interés adicionales para mejorar la interpretabilidad del conjunto de reglas resultante. Sin embargo, el grado de *interés* de un patrón se confunde normalmente con el grado de *precisión* del mismo, de tal forma que la mayoría de la bibliografía se centra en maximizar la precisión de los patrones descubiertos, ignorando otros criterios de calidad igualmente importantes. De hecho, en la práctica, la correlación entre precisión e interés no está tan evidente. Por ejemplo, la afirmación "*los hombres no dan a luz*" es altamente precisa pero nada interesante [91]. De este modo, no hay un convenio extendido respecto a la definición formal del interés de una regla. Geng et al. [97] definieron el grado de interés de un patrón como un compendio de diferentes conceptos tales como *concisión*, *cobertura*, *fiabilidad*, *peculiaridad*, *diversidad*, *novedad*, *sorpresa*, *utilidad* y *accionabilidad*.

Las medidas de interés se pueden dividir en dos clases: objetivas y subjetivas. Las primeras se basan exclusivamente en los datos mientras que las medidas subjetivas tienen en cuenta no sólo los datos sino también el conocimiento del usuario. La mayoría de las medidas objetivas se basan en medidas de probabilidad, siendo funciones de la tabla de contingencia (Tabla 2.9). El soporte y la confianza son ejemplos de este tipo de medidas. Con el objetivo de destacar las ventajas e inconvenientes de cada medida, se han propuesto diferentes propiedades que podrían ser deseables. Por ejemplo, Piatetsky-Shapiro planteó las siguientes tres propiedades para una medida F y una regla $X \rightarrow Y$ [199]:

- $F = 0$ si X e Y son estadísticamente independientes, es decir $sop(XY) = sop(X) \cdot sop(Y)$,
- F crece de forma monótona con $sop(XY)$ cuando $sop(X)$ y $sop(Y)$ permanecen fijos,

Tabla 2.9: Tabla de contingencia para la regla $X \rightarrow Y$.

	Y	$\neg Y$	
X	$n(XY)$	$n(X\neg Y)$	$n(X)$
$\neg X$	$n(\neg XY)$	$n(\neg(XY))$	$n(\neg X)$
	$n(Y)$	$n(\neg Y)$	N

- F decrece de forma monótona con $sop(X)$ (o $sop(Y)$) cuando $sop(XY)$ y $sop(Y)$ (o $sop(X)$) permanecen fijos.

El soporte y la confianza, por ejemplo, no cumplen estas tres propiedades, mientras que otras medidas tales como los factores de certeza o la medida de Klosgen sí las cumplen [30, 148]. En la bibliografía se pueden encontrar otras propiedades adicionales, algunas de las cuales pueden o no ser deseadas por los usuarios [245, 157, 97]. Dado que se han definido una gran cantidad de medidas objetivas y que no existe ninguna mejor que el resto en todas las situaciones posibles, se han desarrollado también diferentes estrategias para seleccionar las medidas más adecuadas en cada caso. Algunas de las más relevantes se pueden encontrar en las referencias [245, 157, 265].

Al contrario que las medidas objetivas, las subjetivas tienen en cuenta no sólo la información de los datos, sino también el conocimiento del usuario. De este modo, se han planteado diferentes enfoques para integrar la información que proporciona el usuario [97]:

- El usuario proporciona una especificación formal de su conocimiento (por ejemplo, mediante predicados lógicos de primer orden) y basándose en esta información el sistema selecciona los patrones inesperados que mostrará al usuario.
- El usuario selecciona interactivamente conjuntos de patrones no interesantes para eliminarlos.
- El usuario especifica ciertas restricciones que permitirán reducir el espacio de búsqueda y el número de patrones obtenidos.

Además de éstas se han propuesto también otras estrategias. Por ejemplo, Berzal et al. definieron el concepto de reglas *muy fuertes* [30]. Una regla $X \rightarrow$

Y se dice *muy fuerte* si su soporte y factor de certeza son mayores que los especificados por el usuario y, además, si el soporte y el factor de certeza de la regla $\neg Y \rightarrow \neg X$ son también mayores que los correspondientes umbrales. La idea subyacente es que esta última regla proporciona una información equivalente a la primera. Se han propuesto otros muchos otros enfoques para reducir el número de reglas basándose en la redundancia de las mismas, una descripción de los principales se puede encontrar en la referencia [57].

Reglas de asociación difusas

Los algoritmos clásicos de extracción de reglas de asociación, dividen los dominios continuos en intervalos crisp para poder manejar las variables correspondientes. Por ejemplo, considérense los datos de la Tabla 2.10. Los atributos *Longitud gen* y *Longitud intergénico* son contínuos y, por tanto, no es factible buscar directamente itemsets frecuentes que contengan estos atributos. Es necesario un paso previo de preprocesamiento para discretizar el dominio, o, en otras palabras, dividir el dominio en intervalos. Una vez llevada a cabo la discretización, cada valor continuo se reemplaza por el intervalo al que pertenece. En la bibliografía se pueden encontrar diversas propuestas para llevar a cabo la discretización automática de los dominios continuos [237, 175].

Sin embargo, al dividir un dominio continuo en intervalos que cubren ciertos rangos de valores surge el "problema del límite brusco" (del inglés, *the sharp boundary problem*). Los elementos que se encuentren cerca de los límites de los intervalos serán ignorados o excesivamente considerados dependiendo del caso. Por ejemplo, reglas como "Si la longitud del gen está en el intervalo $[1541, 14733]$, entonces el contenido en G+C tiende a estar en el intervalo $[0{,}26, 0{,}38]$" y "Los genes largos tienden a tener bajo contenido en G+C" pueden ser ambas interesantes dependiendo de la situación. La primera es más específica, mientras que la segunda es más general en su expresión semántica. Sin embargo, la primera presenta el antes mencionado *problema del límite brusco*, es decir, genes de 1540 pb ó con 0,259 de contenido en G+C no serán considerados. Por el contrario, la segunda regla es más flexible, pudiendo reflejar estos casos límite [58]. Es más, la teoría de

Tabla 2.10: Ejemplo de tabla de datos con variables continuas.

Gen	Longitud del gen	Longitud del intergénico	Orientación
YAL008W	492	280	*Tandem*
YAL003W	1290	742	*Divergente*
YAL018C	885	683	*Divergente*
YAL002W	2217	546	*Tandem*
YAL009W	4299	188	*Divergente*
YAL010C	1965	188	*Divergente*
YAL011W	1107	215	*Tandem*
YAL012W	918	268	*Tandem*
YAL013W	471	250	*Divergente*
YAL015C	2634	250	*Tandem*
YAL017W	330	149	*Divergente*
YAL019W	393	683	*Tandem*
YAL021C	1215	99	*Tandem*

conjuntos difusos ha demostrado ser una metodología superior para mejorar la interpretabilidad de estos intervalos, debido a que modelan los conceptos de un modo más acorde a cómo los humanos expresamos el conocimiento [75]. De este modo, en el caso difuso, los dominios continuos se "fuzzifican" definiendo sobre ellos conjuntos difusos (Figura 2.16). Por tanto, las reglas de asociación difusas son también expresiones de la forma $X \rightarrow Y$, pero en este caso, X e Y son conjuntos de pares atributo-valor difusos.

La forma tradicional de definir los conjuntos difusos consiste en consultar a un experto del dominio, el cual determina cómo deben ser las funciones de pertenencia. Sin embargo, esto requiere acceso al conocimiento del dominio, lo que en muchos casos es difícil o incluso imposible de conseguir. Así, se han propuesto diversos enfoques para definir automáticamente los conjuntos difusos: basados en clustering [62, 106, 93], en algoritmos genéticos [13] y

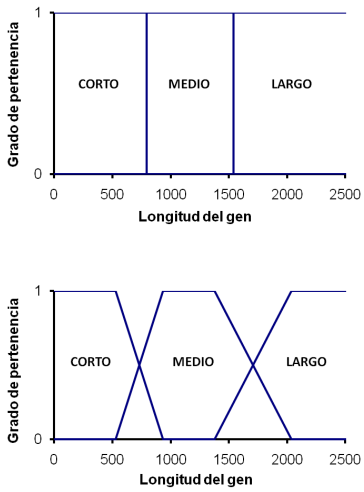


Figura 2.16: Ejemplo de divisiones crisp y difusas.

otros muchos. Aunque estas estrategias son útiles en ciertos casos, hay que considerarlas con cautela, ya que los conjuntos difusos obtenidos podrían ser difíciles de interpretar, es decir, podría resultar difícil asociarlos a una etiqueta (por ejemplo: alto, bajo, largo, etc.).

Al evaluar el interés de una regla de asociación difusa, el procedimiento usual consiste en utilizar adaptaciones difusas del soporte y la confianza. Así, se han planteado diversas generalizaciones de estas dos medidas [81]. A la hora de realizar los cálculos, se suelen reemplazar las operaciones clásicas de conjuntos por las correspondientes operaciones de conjuntos difusos. De este modo, dada una base de datos transaccional D, el grado de pertenencia de una transacción $t \in D$ a un itemset difuso X se calcula como $X(t) = \otimes_{X_i \in X} X_i(t)$, donde $\otimes$ representa una *t-norma* [83]. Por ejemplo, considérese el itemset:

$$X = \{(Longitud\ gen = Largo), (Longitud\ intergenico = Corto)\}$$

y la transacción:

$$YAL010C = \{(Longitud\ gen = 1965), (Longitud\ intergenico = 188), (Orientacion\ gen = Divergente)\}$$

Supóngase que el grado de pertenencia de 1965 al item difuso $Longitud\ gen = Largo$ es 0,6, y que la pertenencia de 188 al item difuso $Longitud\ intergenico = Corto$ es 1. Considérese también que la t-norma escogida es el *minimo*. Entonces, el grado de pertenencia de la transacción $YAL010C$ al itemset difuso X viene dado por:

$$X(YAL010C) = min(0{,}6, 1) = 0{,}6$$

Así pues, teniendo en cuenta todo lo anterior, el soporte difuso de un itemset X se suele calcular como:

$$Supp(X) = \sum_{t \in D} [\otimes_{X_i \in X} X_i(t)]$$

Es decir, como la suma de los grados de pertenencia de las transacciones de la base de datos al itemset X. Finalmente, el soporte difuso y la confianza de una regla $X \rightarrow Y$ vienen dados por:

$$\begin{aligned} Supp(X \rightarrow Y) &= \sum_{t \in D} X(t) \otimes Y(t), \\ Conf(X \rightarrow Y) &= \frac{\sum_{t \in D} X(t) \otimes Y(t)}{\sum_{t \in D} X(t)} \end{aligned}$$

Aunque la mayoría de las propuestas difusas se basan en las extensiones que se acaban de describir, también se pueden encontrar en la bibliografía algunos otros enfoques [75, 81, 100].

Finalmente, es importante comentar que no se le ha prestado gran atención al desarrollo de algoritmos de extracción de reglas de asociación difusas, probablemente debido a que por lo general se pueden adaptar los algoritmos clásicos crisp para el tratamiento de conjuntos difusos [81, 76]. La primera propuesta de minería de reglas de asociación difusas fue descrita por Lee et al. [154]. Los autores presentaron un enfoque sencillo en el que se fijaba un umbral para transformar las transacciones difusas en transacciones crisp, ejecutando posteriormente un algoritmo clásico de extracción de reglas de asociación sobre las nuevas transacciones. Posteriormente, otro autores presentaron algoritmos para la extracción de reglas de asociación difusas tales

como F-APACS y FARM [20, 21], extensiones del algoritmo Equi-depth (EDP) [296] y otros métodos que siguen la filosofía Apriori [106, 117]. Un listado detallado puede encontrarse en la referencia [75].

Aplicaciones previas de las reglas de asociación al análisis de datos genómicos

Como ya se ha comentado, las técnicas de extracción de reglas de asociación se han aplicado en numerosas ocasiones en Bioinformática. Por ejemplo, Thierry-Mieg y Trilling [250] y Oyama et al. [188] describen la utilización de las reglas de asociación para obtener relaciones entre las interacciones de proteínas y las características de las proteínas implicadas en estas interacciones. La idea básica de estos trabajos es simple:

- En la tabla sobre la que se aplica la extracción de reglas aparecen dos filas por cada interacción: una para la interacción protA-protB y otra para protB-protA,
- En las columnas se colocan las características de las proteínas, de modo que en la primera mitad de las columnas se sitúan las características de una de las proteínas de la interacción, mientras que la segunda mitad de las columnas se reserva para las características de la otra.

Con las reglas de asociación obtenidas de una tabla de este tipo, se consigue relacionar unas características con otras, en función de las interacciones que se producen entre las proteínas.

En otro ejemplo de aplicación de las reglas de asociación en Bioinformática, A. Rodríguez et al. [208] utilizan un algoritmo Apriori modificado para fijar relaciones entre secuencias de proteínas y sus características. La mejora del algoritmo Apriori propuesta en este trabajo, consiste básicamente en la eliminación de transacciones e ítems durante el proceso de búsqueda conforme dejan de ser necesarios, de forma que se aumenta la eficiencia del algoritmo.

Bebek et al. [27] integran datos de expresión, diversas bases de datos biológicas (*Gene Ontology* [18, 3] y *Kyoto Encyclopedia of Genes and Genomes* ó KEGG [144]) y reglas de asociación para inferir caminos de señalización

entre dos proteínas dadas. En primer lugar, generan un grafo cuyos nodos representan genes y cuyos arcos enlazan pares de genes que presentan perfiles de expresión similares. Dadas dos proteínas, el sistema busca todos los posibles caminos que conectan los dos genes correspondientes. Con el objetivo de filtrar el conjunto de posibles caminos entre los dos genes, la búsqueda se guía mediante un conjunto de reglas de asociación que relacionan términos GO. Estas reglas representan asociaciones entre anotaciones GO de proteínas que se sabe participan en el mismo camino.

Las reglas de asociación se han utilizado también en algunos casos para mapear las anotaciones de distintas bases de datos. Por ejemplo, Yu et al. [289] desarrollaron PIPA, un sistema para inferir funciones de proteínas. La aplicación anota la función de las proteínas combinando los resultados de múltiples programas y bases de datos, tales como InterPro, *The Conserved Domain Database* y otras. En este caso, se utiliza la extracción de reglas de asociación para mapear automáticamente los diferentes esquemas de clasificación de cada programa/base de datos a términos GO. Otro ejemplo de este tipo es el trabajo de Tveit et al. [264], en el que las reglas de asociación se obtienen para encontrar asociaciones entre términos del *The Medical Subject Headings thesaurus* (MeSH [4]) y términos GO. En el mismo artículo, sin embargo, se proponen otras dos metodologías que parecen dar mejores resultados que las reglas de asociación.

En otras ocasiones, el objetivo último no son las reglas de asociación en sí, sino el conjunto de itemsets. Por ejemplo, en el trabajo reciente de Carmona-Saez [51], el sistema obtiene itemsets estadísticamente sobre-representados en un conjunto de anotaciones de un grupo de genes. Las anotaciones se pueden obtener de diferentes fuentes, tales como GO o KEGG. Klema [146] también trata de obtener conjuntos de itemsets interesantes integrando text-mining, similitud funcional obtenida de anotaciones GO y datos de expresión.

Además de todas las aplicaciones descritas hasta el momento, las reglas de asociación se han utilizado también con éxito en el análisis de micro-arrays. Así, aparecen dos enfoques diferentes (aunque complementarios) para extraer información de una matriz de expresión mediante reglas de asociación:

- Obtener reglas que relacionen los niveles de expresión con cualquier otro tipo de condición/anotación biológica de interés.
- Obtener reglas que describan cómo la expresión de uno o más genes se asocia con la expresión de otro conjunto de genes.

Reglas de asociación que relacionan los niveles de expresión con otras características

Varios autores han seguido esta idea, siendo el trabajo de Creighton y Hanash [70] uno de los más conocidos en este campo. Este trabajo presenta una versión del algoritmo Apriori para extraer reglas de asociación de una matriz de expresión. En esta matriz las filas representan diferentes condiciones experimentales, las columnas genes y cada elemento de la matriz indica el estado del gen para esa condición: sub-expresado, sobre-expresado o no-modificado. Además de la información proporcionada por la matriz de expresión génica, incluyen otra información adicional como ampliación de dicha matriz, con otras columnas que aportan información acerca de las condiciones experimentales. Así por ejemplo, si las distintas condiciones experimentales fueran personas diferentes, una columna nueva podría ser la edad de esas personas. De esta forma, los ítems entre los que se buscan relaciones de asociación tendrían el siguiente aspecto: $\{GenA = sobre\_expresado\}$, $\{Edad > 60 = si\}$, etc. Para limitar la cantidad de reglas obtenidas, los autores reducen la búsqueda a aquellas que tienen un único ítem en el antecedente o en el consecuente.

En esta misma línea, Carmona-Saez et al. [52] extraen reglas de asociación de un conjunto en el que combinan datos de expresión y términos GO. Al igual que en el trabajo anterior, los autores consideran tres posibles valores para los niveles de expresión de los genes: sub-expresados, sobre-expresados o no-modificados. La matriz de expresión está construida de forma que aparece una fila por cada gen y una columna por cada condición experimental. Además, añaden otra columna en la que se incluyen ciertas características de los genes obtenidas a partir de la Gene Ontology (GO).

Para la extracción de las reglas, Carmona-Saez et al. [52] implementan el algoritmo Apriori modificado según se propone en la referencia [208]. De

las reglas obtenidas se seleccionan solamente aquellas que tienen un término GO en el antecedente y varios ítems en el consecuente representando estados de expresión de diferentes genes. Para reducir aún más el número de reglas, eliminan aquellas que consideran redundantes. Finalmente, calculan un p-valor para cada regla que viene dado por un test χ^2, calculado bajo la hipótesis nula de que el antecedente y el consecuente son estadísticamente independientes. Es decir, los autores hacen uso del test χ^2 para asegurar la correlación entre el antecedente y el consecuente de las reglas.

Una aplicación similar fue desarrollada por Martinez et al. [170]. En este trabajo los autores describen GenMiner, una herramienta que facilita la extracción de reglas de asociación en una tabla de datos que integra niveles de expresión génica, anotaciones y cualquier otra condición biológica. En la tabla se incluyen anotaciones de distintas bases de datos, tales como términos GO, anotaciones KEGG, anotaciones bibliográficas, etc. GenMiner se basa en el modelo soporte-confianza e implementa una versión del algoritmo Close [191]. Los autores explican que los datos que procesa GenMiner están altamente relacionados y que, por lo tanto, el algoritmo Apriori consume demasiado tiempo y memoria. Además, según comentan, Apriori generaría demasiadas reglas, muchas de ellas redundantes. Los autores sostienen que Close es un algoritmo diseñado específicamente para tratar este tipo de datos. Finalmente, al contrario de lo que se describe en el trabajo de Carmona-Saez et al. [52], los autores no fijan ningún tipo de plantilla para filtrar el conjunto de reglas resultante, sino que permiten que se generen todas las reglas, independientemente de los atributos que aparezcan en el antecedente y el consecuente, ya que consideran que todas las reglas proporcionan información relevante para el experto.

Reglas de asociación que relacionan genes y sus niveles de expresión
Esta estrategia busca asociaciones de la forma:

$$\{GenA = expresionA, GenB = expresionB, ...\} \rightarrow$$
$$\{GenC = expresionC, GenD = expresionD, ...\},$$

donde *expresionA*, *expresionB*, etc., son valores discretos de expresión que repesentan etiquetas lingüísticas tales como *sobre_expresado*, *sub_expresado*, o *no_modificado*. En otros casos lo que se pretende capturar es la tendencia del nivel de expresión entre muestras y, por tanto, *expresionA*, *expresionB*, etc., representan el incremento o decremento del nivel de expresión entre muestras. Esto último se puede conseguir simplemente sustituyendo los valores originales de cada muestra por las diferencias de valores entre muestras. Como ejemplo, véanse las Tablas 2.11 y 2.12.

Tabla 2.11: Ejemplo de matriz de expresión.

Tiempo	**GenA**	**GenB**	**...**
Tiempo1	exprA1	exprB1	...
Tiempo2	exprA2	exprB2	...
Tiempo3	exprA3	exprB3	...
Tiempo4	exprA4	exprB4	...
...	...	...	...
Tiempoi	exprAi	exprBi	...
Tiempoi+1	exprAi+1	exprBi+1	...
...	...	...	...

El trabajo de Ponzoni et al. [202] se puede enmarcar en este tipo de enfoque. Los autores obtienen un conjunto de reglas de la forma:

$$\{GenA = +/-\} \rightarrow \{GenB = +/-\},$$

que representan uno de los tres tipos siguientes de asociación:

- *Simultanea*: el nivel de expresión del *GenB* en el tiempo i depende del nivel de expresión del *GenA* en ese mismo instante de tiempo.
- *Retrasada*: el nivel de expresión del *GenB* en el tiempo i depende del nivel de expresión del *GenA* en el tiempo $i-1$.
- *Basada en el cambio*: cuando el nivel de expresión del *GenA* cambia de estado, el nivel de expresión del *GenB* también cambia su estado.

Tabla 2.12: Se ha calculado la diferencia entre tiempos adyacentes de la Tabla 2.11

Tiempo	**GenA**	**GenB**	**...**
Tiempo2-Tiempo1	exprA2-exprA1	exprB2-exprB1	...
Tiempo3-Tiempo2	exprA3-exprA2	exprB3-exprB2	...
Tiempo4-Tiempo3	exprA4-exprA3	exprB4-exprB3	...
...	...	...	...
Tiempo(i+1)-Tiempoi	exprA(i+1)-exprAi	exprB(i+1)-exprBi	...
...	...	...	...

La principal novedad de la metodología que proponen estos autores consiste en el cálculo de umbrales adaptativos para la discretización de los niveles de expresión. Los autores argumentan que el nivel de expresión que requiere un *genR* para activar (inhibir) el *genT1*, no es necesariamente el mismo que el requerido por el mismo *genR* para activar (inhibir) otro *genT2*. Por lo tanto, proponen una metodología para calcular umbrales específicos de regulación para cada par de genes.

Un trabajo similar es el de McIntosh et al. [174]. En este caso, el algoritmo hace uso de una estructura de datos en forma de árbol que permite evitar cualquier restricción de soporte, ya que es capaz de restringir el espacio de búsqueda estimando la confianza de las reglas que están a punto de generarse. Tan sólo se consideran reglas de la forma:

$$\{GenA = expresado/no\_expresado\} \rightarrow \{GenB = expresado/no\_expresado, \\ GenC = expresado/no\_expresado, ...\}$$

Es decir, reglas con un único item (gen) en el antecedente y varios items (genes) en el consecuente.

Otros trabajos que podrían ser incluidos en este apartado se pueden encontrar en las referencias [28, 263, 98, 118, 72].

2.2.4 Detección de TFBSs

Tal y como se describió en la Sección 2.1.4, parece ser que no existe un código de apareamiento sencillo aminoácido-base que determine la unión del factor de transcripción a una secuencia de ADN determinada. Sin embargo, algunos tipos de interacción aminoácido-base aparecen con mucha más frecuencia que otros. De esta forma, y con el objetivo de representar de forma computacional las preferencias de unión de los TFs, surgieron los llamados *motivos de regulación*. Dado un TF, sus preferencias de unión se representan mediante una matriz que indica, para cada posición del TFBS, la afinidad que presenta cada nucleótido. Dichas matrices se derivan normalmente de alineamientos de los TFBSs conocidos para ese TF. Aunque se han propuesto diferentes representaciones para los motivos [186], la más extendida consiste en estas matrices que recogen para cada nucleótido su frecuencia de aparición en cada posición (PFMs, del inglés *position frequency matrices*), o bien matrices de pesos que proporcionan un valor ponderado de afinidad para cada posición de la unión entre el motivo y una secuencia de ADN (PWMs, del inglés *position weighted matrices*). Por ejemplo, supóngase que se ha comprobado experimentalmente que un TF dado se une a las siguientes secuencias de ADN:

AATAACGGAA
AATAACGGAA
CATAACGGAA
GCTAACGGCA
TGAAACTTGG
TACAACTGAA

La PFM correspondiente se obtiene contando las ocurrencias de A, C, G y T en cada una de las posiciones (Tabla 2.13).

La identificación de TFBSs en un conjunto de secuencias de ADN es un campo de investigación muy activo. Así, existen dos tendencias principales: 1) técnicas que identifican nuevos motivos y 2) técnicas que detectan TFBSs a partir de motivos ya conocidos. Las primeras, tratan de encontrar patrones de subsecuencias significativos en el conjunto de secuencias. Algunos de los

Tabla 2.13: Ejemplo de PFM.

Posición	A	C	G	T
1	2	1	1	2
2	4	1	1	0
3	1	1	0	4
4	6	0	0	0
5	6	0	0	0
6	0	6	0	0
7	0	0	4	2
8	0	0	5	1
9	4	1	1	0
10	5	0	1	0

métodos de este tipo más conocidos son MEME [23], *Gibbs sampling* [152] y AlignACE [125]. Un listado más completo puede encontrarse en [73].

Por otra parte, los métodos de detección tratan de inferir TFBSs en secuencias de ADN a partir de motivos ya conocidos. Estos métodos son los que se utilizarán en este trabajo, dado son más apropiados para los objetivos que se persiguen, tal y como se verá posteriormente. Los métodos pioneros de este tipo son Patser [119] y el propuesto por Staden [238], siendo probablemente Patser el más popular. Dada una secuencia de ADN del tamaño del motivo en cuestión, la idea intuitiva de Patser consiste en medir cuánto se ajusta dicha secuencia al patrón definido por la matriz del motivo. Para esto, se calcula en primer lugar la probabilidad de que la matriz ocurriera por azar. A continuación, se calcula cuánto se vería modificada esta probabilidad si se añadiera la secuencia en cuestión al motivo. A partir de este factor de modificación se define una medida de ajuste, de tal forma que cuanto mayor sea el valor de la medida, menor será la probabilidad de que la matriz ocurra por azar y, por tanto, más cercana está la secuencia de ADN al patrón descrito por la matriz original. Otro método bastante conocido es el descrito por Sandelin

et al. [216], en el que los autores combinan la técnica propuesta por Staden [238] con información filogenética. Más recientemente se han descrito otros enfoques que tratan de modelar las dependencias entre posiciones del TFBS. Por ejemplo, Tomovic y Oakeley [257] propusieron un método que incorpora en el valor de la medida final el grado de interdependencia posicional. Otro ejemplo es el de Fatemeh et al. [89], quienes desarrollaron un método basado en el contenido de información conjunta y en la información mutua. En este último caso, las dependencias posicionales se tienen en cuenta considerando todas las posibles parejas de posiciones.

2.2.5 Detección de módulos de regulación

Los factores de transcripción no suelen, por lo general, unirse de forma independiente al ADN. Como ya se comentó previamente (Sección 2.1.4), los TFBSs tienden a agruparse en los llamados módulos de regulación (CRMs). Se ha desarrollado una amplia variedad de técnicas para el estudio de los CRMs *in silico*, que se diferencian tanto en las estrategias que utilizan, como en el objetivo específico que persiguen.

De forma global, se pueden apreciar diferentes estrategias para reducir el espacio de búsqueda y mejorar la significación de los módulos, entre las que cabe destacar el estudio de secuencias conservadas entre especies, el agrupamiento de elementos reguladores cercanos y el estudio de las regiones promotoras [267]. Muchas de las propuestas se centran en el análisis de secuencias no-codificadoras que se conservan entre especies [108, 7, 224, 34]. Estas secuencias son buenas candidatas para contener elementos reguladores [111]. De hecho, se ha demostrado que, en general, los TFBSs tienden a estar más conservados que el ADN que los rodea [177]. Por otra parte, el agrupamiento de elementos reguladores se ha considerado también en numerosas ocasiones [92, 142, 7, 224, 34, 176]. Se piensa que los CRMs tienden a abarcar varios cientos de pares de bases [17], por lo que la aparición de grupos de TFBSs en una pequeña zona del genoma se considera un indicador fiable de su funcionalidad. Finalmente, el estudio de las regiones promotoras se ha llevado a cabo también con éxito en repetidas ocasiones [241, 224, 201, 198, 212, 24].

Según el objetivo pretendido, las técnicas computacionales para la detección de CRMs se pueden dividir en las siguientes tres clases [266]: 1) métodos que escanean secuencias (o genomas completos) buscando CRMs que siguen un modelo predefinido, 2) métodos que buscan CRMs similares en un conjunto de genes relacionados (por ejemplo co-expresados o co-regulados) y 3) métodos que escanean secuencias (o genomas completos) buscando grupos de TFBSs formados por cualquier combinación de TFs. A continuación se describen algunas de las principales propuestas de cada clase. Existen otros métodos que no aparecen descritos en las secciones siguientes pero que podrían incluirse en las mismas, un listado y descripción detallada de éstos pueden encontrase en [266, 147, 267].

Búsqueda de CRMs predefinidos

El objetivo de este tipo de métodos es identificar CRMs que contengan sitios de unión para una combinación específica de PWMs. Para ello, hacen uso tanto de bibliotecas de motivos conocidos como de bibliotecas de agrupamientos de TFBSs (para una combinación de TFs específica en la que se centra la atención). En este apartado cabe destacar el *Enhancer Element Locator (EEL)* propuesto por Hallikas et al. [108]. Este programa utiliza alineamientos de TFBSs previamente inferidos en dos especies diferentes para predecir CRMs. En primer lugar, se consideran las secuencias de las dos especies por separado y se predicen potenciales TFBSs en ambas utilizando las PWMs. A continuación, los TFBSs inferidos en ambas especies se alinean mediante el algoritmo Smith-Waterman [232]. Es decir, una vez inferidos los TFBSs a partir de las secuencias, éstas no se vuelven a utilizar: EEL no asume que cada TFBS está conservado, sino que lo que tiende a mantenerse es el orden de los sitios de unión funcionales. Adicionalmente, la función de evaluación tiene en cuenta la conservación de las distancias entre los TFBSs del CRM. Las entradas del programa en este caso son: dos secuencias homólogas de ADN, el conjunto de PWMs de interés y una serie de parámetros necesarios para el alineamiento.

Otro método de este tipo es *Cister* [92]. En este caso el algoritmo requiere proporcionar una secuencia, el conjunto de PWMs, el número esperado de

TFBSs en los módulos, las distancias esperadas entre los sitios de unión del módulo y la distancia esperada entre módulos. A partir de esta información el método construye un modelo oculto de Markov (HMM) con tres estados básicos: *motivo*, *fondo intra-modulo* y *fondo inter-modulo*. Las probabilidades de transición entre estos estados siguen distribuciones geométricas de acuerdo con los parámetros de entrada. A partir de este HMM, se puede calcular la probabilidad de que cada base de la secuencia de entrada fuera generada por un estado *módulo*, en contraposición a que fuera generada por un estado *inter-modulo*.

El programa MSCAN propuesto por Johansson et al. [142] puede enmarcarse también en esta clase de procedimientos. Dada una secuencia y las PWMs de interés, MSCAN evalúa la significación estadística de combinaciones TFBSs no solapados. La significación de las combinaciones se calcula en una ventana de tamaño fijo que se va desplazando a lo largo de dicha secuencia (el tamaño de la ventana es un parámetro de entrada). Cada PWM proporcionada se compara con cada posición de la ventana para obtener un grado de ajuste. A partir de este grado de ajuste se calcula un p-valor, que indica la probabilidad de obtener un valor de ajuste igual o mayor. Para una ventana dada, MSCAN calcula valores de significación para todas las combinaciones de TFBSs (de hasta k elementos, donde k es otro parámetro de entrada) y selecciona la combinación óptima. Se proporciona una predicción de salida si el p-valor calculado es menor que un umbral especificado por el usuario.

Construcción de CRMs a partir de genes relacionados

En este caso se buscan CRMs similares para un conjunto de genes co-regulados o co-expresados. Dado este conjunto de genes, se analizan sus regiones reguladoras promotoras buscando sitios de unión para una serie de TFs. A partir de los TFBSs inferidos, construyen o seleccionan una combinación de PWMs de interés. Así, estos métodos combinan el agrupamiento de los sitios de unión con la asunción de que patrones de expresión similares vienen dados por elementos de regulación similares.

Es de especial interés en este trabajo la propuesta de Sun et al. [241]. Estos autores aplican la minería de itemsets *cerrados* frecuentes para encontrar co-ocurrencias de TFs. Más concretamente, esta metodología se centra en un conjunto de genes que se saben co-expresados o co-regulados. Se seleccionan las secuencias promotoras de dichos genes (hasta $1000pb$ antes del punto de inicio de la transcripción) y se aplica el algoritmo CHARM [291, 94], considerando que cada promotor forma una transacción y que los TFBSs correspondientes son los items. La fiabilidad de los itemsets se evalúa mediante el soporte de los mismos y mediante un p-valor calculado tal y como se describe en la referencia [94].

Otra propuesta interesante es la de Aerts et al. [7]. Estos autores desarrollaron *ModuleSearcher*, un programa que, dado un conjunto de genes co-regulados y sus correspondientes secuencias, procede buscando combinaciones de TFBSs en ventanas de dichas secuencias. Para cada ventana, *ModuleSearcher* encuentra la mejor combinación de TFBSs mediante un algoritmo A^*. Con el objetivo de disminuir el nivel de ruido, se consideran sólo regiones no-codificadoras conservadas entre especies. En el caso particular del trabajo referenciado en [7] se centran en regiones conservadas entre el genoma humano y el de ratón.

Sharan et al. desarrollaron CREME [224], un paquete software para identificar y visualizar CRMs en los promotores de un conjunto de genes. Los autores parten de un conjunto de TFBSs previamente identificados y de sus correspondientes motivos. En primer lugar, seleccionan sólo los motivos que aparecen sobre-representados en el conjunto de promotores con respecto a un conjunto de secuencias de fondo. En el mismo trabajo [224] se describen distintos tests estadísticos para estimar esta sobre-representación. Una vez hecho esto, y mediante un algoritmo hash que se propone en dicho trabajo, se buscan las combinaciones de los motivos seleccionados que co-ocurren en clusters dentro de los promotores. Dada una combinación, se calcula su frecuencia esperada a partir de las ocurrencias de los motivos que la componen. La significación estadística de cada co-ocurrencia se mide, por tanto, en comparación con dicha frecuencia esperada. Finalmente, la metodología se aplica de nuevo tan sólo a regiones de los promotores conservadas entre

especies. En el caso particular del trabajo referenciado en [224], los autores consideran regiones conservadas entre el genoma humano y el de ratón.

Búsqueda de CRMs en el genoma completo

En este apartado se incluyen métodos que escanean secuencias (o genomas completos), buscando grupos de sitios de unión para cualquier combinación de TFs. Estas técnicas no requieren ningún modelo predefinido, conjunto determinado de PWMs o conjunto de genes de interés. Es más, este tipo de métodos hacen, por lo general, pocas asunciones acerca de los CRMs que buscan, lo que los convierte, por tanto, en los métodos cuya aplicabilidad es más general.

En primer lugar, cabe destacar PReMod [34]. Este método se basa, de nuevo, en la conservación filogenética de los TFBSs pertenecientes a un CRM. Los autores parten de 481 PWMs obtenidas de TRANSFAC [277] y de una serie de secuencias no-codificadoras. Estas secuencias se derivaron previamente del alineamiento de los genomas de humano, ratón y rata (utilizan MULTIZ [35] para obtener estos alineamientos). En primer lugar, se consideran las secuencias alineadas de cada especie independientemente, y se utiliza una medida de probabilidad logarítmica para detectar potenciales sitios de unión para cada PWM. A continuación, se combinan los TFBSs inferidos para cada especie, calculando para esto una especie de media ponderada que proporciona una "puntuación" para los sitios de unión alineados. Seguidamente, se detectan regiones de hasta $2kb$ que aparecen significativamente enriquecidas en hasta 5 TFs. La "puntuación" de cada uno de estos módulos se calcula a partir de las "puntuaciones" de cada sitio de unión, asignándole además un p-valor a cada módulo.

Otro método interesante es el propuesto recientemente por Morgan et al. [176]. En este caso la metodología se aplica sobre la secuencia completa del genoma humano. El proceso comienza detectando posibles sitios de unión en el genoma completo para 83 PWMs obtenidas de TRANSFAC [277]. La detección de los posibles TFBSs se lleva a cabo mediante la herramienta Patser (Sección 2.2.4). A continuación, se divide la secuencia completa del genoma

en intervalos consecutivos de $100pb$. A partir de cada uno de estos intervalos, forman una transacción donde los items son los TFs correspondientes. Sobre la base de datos transaccional así construida, aplican un algoritmo de extracción de reglas de asociación, limitando las reglas obtenidas a aquellas que presentan un sólo item en el antecedente y el consecuente. Con el objetivo de evitar el solapamiento de TFBSs, dado un par de TFs A y B, eliminan los sitios de unión de ambos TFs que aparezcan solapados, así como los TFBSs de otros TFs que aparezcan solapados con estos. Finalmente, la bondad de las reglas se evalúa mediante el soporte, la confianza y un p-valor calculado a partir de la distribución hipergeométrica.

Más trabajos de esta clase son los referenciados en [201, 198, 212, 24]. Todos estos llevan a cabo sus análisis sobre el genoma de la levadura, centrándose en el estudio de las regiones promotoras y combinando las coocurrencias de TFs obtenidas con datos de expresión. En el primero de ellos se limitan al estudio de combinaciones de dos TFs, analizando la influencia de la presencia de estas combinaciones en la expresión de los genes objetivo. Los otros tres trabajos siguen ideas similares, extendiendo el estudio a combinaciones de más de dos TFs y aplicando en cada caso diferentes estrategias. Es destacable el uso del algoritmo FPClose [104] en el trabajo de Pham et al. [198], dado que en el Capítulo 5 se propondrá una estrategia basada en la minería de itemsets difusos.

CAPÍTULO 3

Reglas de asociación para el análisis de datos biológicos: estudio del genoma de la levadura

3.1 Introducción

La ingente cantidad de información genómica generada ha impulsado el desarrollo de técnicas computacionales capaces de analizarla [143, 180, 31]. Sin embargo, no se trata sólo del volumen de información creciente, sino de los diferentes tipos de datos recopilados, las relaciones entre ellos y otras peculiaridades de los mismos, tales como la imprecisión, valores perdidos, etc. En este capítulo se presenta una metodología basada en reglas de asociación difusas que permite evitar algunas de las principales dificultades que presenta el análisis de este tipo de información. La metodología se aplica sobre datos estructurales y funcionales del genoma de la levadura (*Saccharomyces Cerevisae*), obteniendo un gran número de asociaciones interesantes, muchas de ellas en concordancia con investigaciones previas en este área, y otras que podrían contribuir al entendimiento de las relaciones genómicas estructurales-funcionales.

Motivación

Hasta hace poco se ha hecho mucho hincapié en aspectos relacionados con herramientas orientadas a la automatización, prestándose poca atención a la integración de información de diversas fuentes. La razón pudiera estar en la relativa simplicidad de los datos de los que se disponía. Sin embargo, con la aparición de las grandes masas de datos derivadas de las tecnologías de alto rendimiento, es necesario avanzar hacia la integración de los datos. No es suficiente ya con conocer qué es un genoma: debemos entender lo que significan sus componentes, cómo funcionan y cómo se relacionan entre sí y con el todo. No obstante, en el momento del desarrollo de este trabajo, la mayoría de los estudios previamente publicados sólo tenían en cuenta una única fuente de datos (por ejemplo, la matriz de expresión). La aparición de diversas propuestas integrativas tras la publicación de este trabajo [163, 51, 146, 170], demuestra la importancia objetiva de este tipo de análisis. Algunas de estas propuestas se han descrito previamente en la Sección 2.2.3.

Así pues, no se trata sólo de un problema relacionado con el volumen de información, sino también de los diferentes tipos de datos generados, en otras palabras, de la heterogeneidad de los mismos. Estos datos pueden venir dados en forma de ontologías, secuencias, medidas, etc. Aunque en la actualidad van surgiendo trabajos que afrontan este problema, aún no existen suficientes propuestas capaces de manejar la heterogeneidad que presenta este tipo de datos.

Por otra parte, es bien conocido que la información biológica tiende a ser imprecisa y a presentar un cierto grado de incertidumbre. Como norma general se suele hacer uso de técnicas clásicas para el análisis de datos biológicos. Sin embargo, existen otras metodologías que han demostrado ser más apropiadas para el tratamiento de este tipo de información (por ejemplo las metodologías difusas) y cuya aplicación no es habitual.

Lo habitual para abordar este tipo de análisis ha sido la aplicación de técnicas tradicionales estadísticas al análisis de datos biológicos. Como ejemplos concretos de este tipo de trabajos véanse, en relación con el estudio que se lleva a cabo en este capítulo, los artículos referenciados en [168, 169, 67,

136]. Sin embargo, la naturaleza de las técnicas estadísticas hace difícil la integración de datos heterogéneos en el análisis. Además, este tipo de técnicas permiten estudiar tan sólo unas pocas variables simultáneamente.

Propuesta

En este capítulo se presenta una aproximación difusa-integrativa al estudio del genoma de la levadura mediante la extracción de reglas de asociación difusas. Más concretamente, se ha desarrollado y aplicado una versión difusa del algoritmo Top-Down Frequent-Pattern Growth (TD-FP Growth) [271], para estudiar las relaciones existentes entre diversas características estructurales y funcionales del genoma de la levadura.

Tal y como se ha comentado anteriormente, los conjuntos difusos son especialmente apropiados para modelar datos imprecisos y ruidosos, mientras que las reglas de asociación manejan con cierta facilidad datos heterogéneos. Por tanto, las técnicas de extracción de reglas de asociación difusas son particularmente apropiadas para llevar a cabo este estudio. Además, al contrario que enfoques previos que sólo permiten estudiar un conjunto muy limitado de variables, la extracción de reglas de asociación permite analizar las relaciones entre un gran número de variables de distinto tipo (niveles de expresión, anotaciones de la Gene Ontology, características estructurales, etc.).

Se escogió el genoma de la levadura *Saccharomyces cerevisae* como objeto de estudio en este trabajo, ya que la gran cantidad de investigaciones llevadas a cabo sobre este organismo, han generado conjuntos de datos de alta calidad y bibliografía abundante describiendo las tendencias y patrones en la organización genética. Las propiedades únicas de este organismo, junto con sus muchas aplicaciones industriales, lo han convertido en uno de los organismos favoritos para la investigación biológica. Así, el genoma de la levadura (*Saccharomyces cerevisae*) fue el primer genoma eucariota en secuenciarse [82]. Desde entonces, el trabajo con este organismo ha ido abriendo paso al resto de estudios en genómica estructural y funcional, estableciendo un estándar en biología celular y molecular, y facilitando de este modo estudios similares en otros organismos [82, 102, 55, 279].

La mayoría de la información acerca del genoma de la levadura se encuentra almacenada en bases de datos como la *Saccharomyces Genome Database* (SGD [84, 6]) y la *Comprehensive Yeast Genome Database* (CYGD [2]). Los datos más recientes, resultantes de las últimas investigaciones sobre este organismo no han sido aún incluidos en las bases de datos y, por tanto, han tenido que ser recopilados de la bibliografía pertinente.

Se ha desarrollado el algoritmo de extracción de reglas de asociación difusas Fuzzy Top-Down Frequent-Pattern Growth (FTD-FP-Growth), para su posterior aplicación sobre una base de datos en la que se integró información estructural y funcional del genoma de la levadura. Entre las reglas obtenidas se encuentran la mayoría de las tendencias previamente descritas entre estas variables, lo que valida el método, y además nuevas asociaciones que podrían contribuir al entendimiento de las relaciones genómicas estructurales-funcionales.

3.2 Construcción de la Base de Datos

3.2.1 Información estructural

La secuencia del genoma de la levadura, así como las "anotaciones" correspondientes se descargaron del servidor ftp de la SGD (versión de Febrero de 2007). De esta forma, se obtuvieron las variables que se detallan a continuación.

Cromosoma

Número del cromosoma en el que se encuentra el gen. El genoma de la levadura está compuesto por 16 cromosomas y cada gen se encuentra en uno de estos 16 cromosomas.

Longitud del gen

Número de bases que forman el gen (incluyendo las tres bases que indican el comienzo del gen y las tres que indican el final).

Orientación del gen

Según la hebra de ADN (Watson o Crick) en la que se encuentre el gen, éste se leerá de izquierda a derecha ó de derecha a izquierda. La zona junto al gen en dirección contraria a la dirección de lectura se denomina región *promotora* del gen. La orientación de un gen se determina teniendo en cuenta con qué linda la región promotora de éste:

- Dado un gen, si su región promotora linda con la región promotora de otro gen, se dice que su orientación es *divergente* (Figura 3.1).

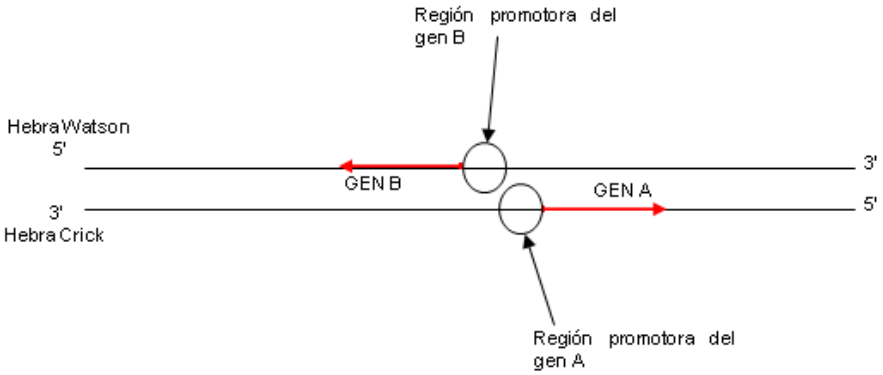


Figura 3.1: La dirección de las flechas que representan los genes indica la dirección de lectura de los mismos. La orientación del GEN A y del GEN B es divergente.

- Dado un gen, si su región promotora linda con el final de otro gen, se dice que su orientación es en *tandem* (Figura 3.2). La orientación de los genes que se encuentran en el inicio o en el final del cromosoma, y cuya región promotora no linda ni con una región promotora ni con el final de otro gen no puede ser determinada (Figura 3.3). Además, hay que tener en cuenta que los cromosomas están formados por dos brazos separados por el *centrómero* (en las figuras anteriores se ha omitido el centrómero para mayor claridad). La orientación de aquellos genes cuya región promotora linda con el centrómero tampoco puede

obtenerse (Figura 3.4). No se consideraron en este estudio aquellos genes cuya orientación no podía ser determinada.

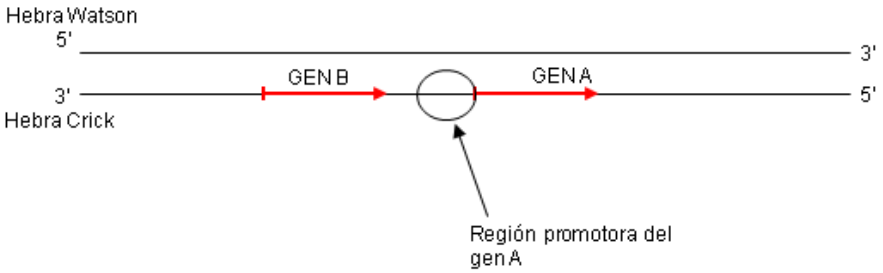


Figura 3.2: La orientación del GEN A es tandem.

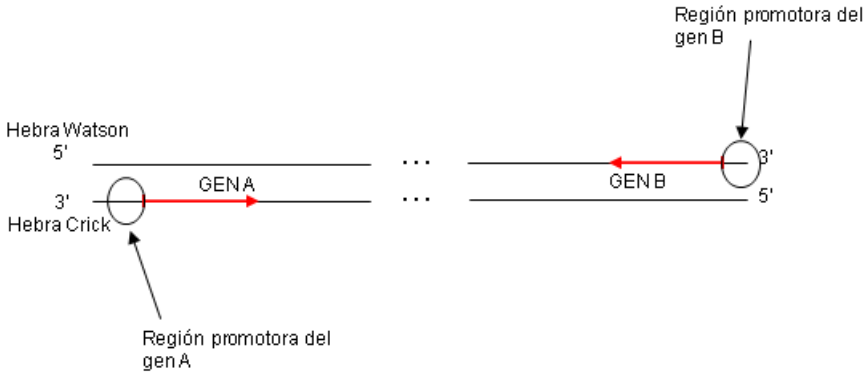


Figura 3.3: La orientación de los genes A y B no puede obtenerse.

Número de intrones

Se refiere al número de regiones no codificadoras que hay en el gen.

Proporción de G+C

Se trata de la proporción de Guanina y Citosina que hay en el gen, es decir, el número de Gs y Cs que hay en la cadena dividido por el número de bases que forman el gen.

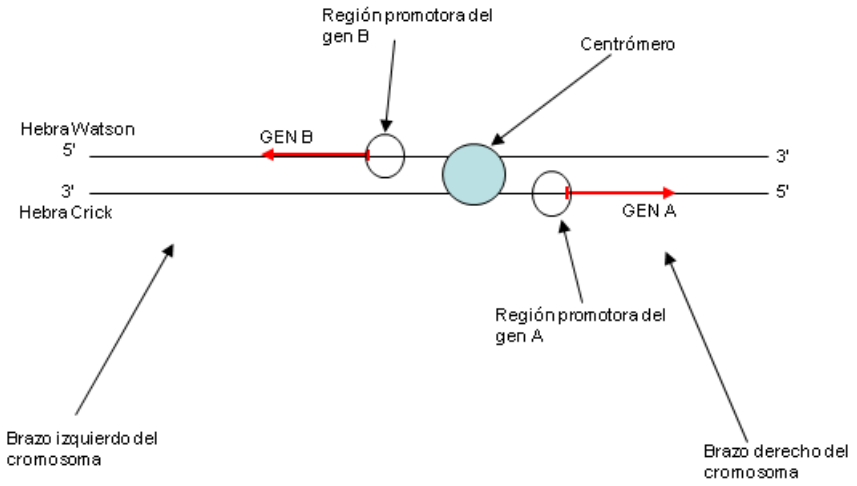


Figura 3.4: La orientación de los genes A y B no puede obtenerse.

GC3s

Cada tres bases del gen codifican un aminoácido; a estos conjuntos de tres bases se les denomina codones. El índice GC3s es la fracción de codones que son sinónimos por su tercera posición y que tienen C ó G en tercera posición. A continuación se explica más detalladamente con un ejemplo.

Se ilustra el cálculo del GC3s en el siguiente gen:
GEN: ATG GTA CTG ACC TAT ATA TCA ATC TAC ACT TAA

Esta secuencia tiene 33 bases lo que significa que hay 11 codones incluido el de inicio (ATG) y el de terminación (TAA). A partir de la Tabla 3.1 se puede conocer qué aminoácidos codifica cada codón (en la tabla se muestran los codones en el ARN que lleva U en vez de T). Al total de codones se le resta el de metionina (Met = ATG) que no tiene sinónimos, esto es, el aminoácido metionina está codificado por un único codón, ATG. Asimismo, se elimina el codón de terminación. Dado que se han contabilizado 11 codones, y se han eliminado 1 de Met +1 de Term = 2 codones, quedan $11-2=9$ codones. De los que quedan, se contabilizan los que acaban en G o C (que tienen G ó C en

Tabla 3.1: Código genético universal

AA	Codón	AA	Codón	AA	Codón	AA	Codón
Phe	UUU	Ser	UCU	Tyr	UAU	Cys	UGU
	UUC		UUC		UAC		UGC
Leu	UUA		UCA	TER	UUA	TER	UGA
	UUG		UCG		UAG	Trp	UGG
	CUU	Pro	CCU	His	CAU	Arg	CGU
	CUC		CCC		CAC		CGC
	CUA		CCA	Gln	CAA		CGA
	CUG		CCG		CAG		CGG
Ile	AUU	Thr	ACU	Asn	AAU	Ser	AGU
	AUC		ACC		AAC		AGC
	AUA		ACA	Lys	AAA	Arg	AGA
Met	AUG		ACG		AAG		AGG
Val	GUU	Ala	GCU	Asp	GAU	Gly	GGU
	GUC		GCC		GAC		GGC
	GUA		GCA	Glu	GAA		GGA
	GUG		GCG		GAG		GGG

la tercera posición): 1 CUG, 1 ACC, 1 AUC y 1 UAC, resultando 4 codones. La fracción 4/9 proporciona el índice $GC3s = 0'444$.

Longitud del intergénico

A la región situada entre el comienzo de un gen y el gen adyacente se le denomina *intergénico* del gen. La longitud del intergénico es el número de bases que lo forman (Figura 3.5). No se consideraron en el análisis aquellos genes cuyo intergénico no es *puro*, es decir, se descartaron aquellos intergénicos que presentan en su secuencia un elemento codificador (ej. tARN, snARN, snoARN, genes ARN, etc.).

G+C del intergénico

Tiene el mismo significado que el G+C del gen sólo que para el intergénico, es decir, es la proporción de G+C que hay en el intergénico del gen.

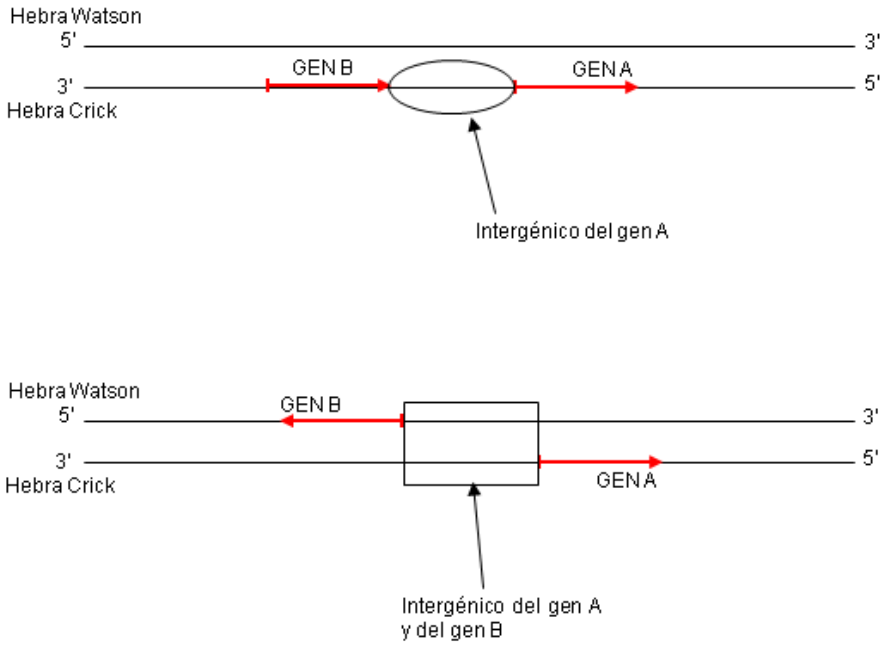


Figura 3.5: La figura muestra el intergénico del GEN A y B en dos situaciones diferentes.

3.2.2 Características funcionales

Respecto a la actividad de un gen aparecen dos características funcionales clave:

- Cantidad del producto final del gen presente en la célula,
- Capacidad para modificar su nivel de expresión como respuesta a condiciones variables.

La primera de estas dos magnitudes fue medida por Ghaemmaghami et al. [126], quienes obtuvieron una medida bastante precisa del número de moléculas protéicas por célula para el 75 % de los genes de la levadura durante su crecimiento normal. La segunda magnitud fue determinada por Tirosh et al. [256] a partir de un conjunto de perfiles de expresión que comprendía más de 1500 condiciones experimentales, lo que les permitió medir la respuesta de cada gen a condiciones experimentales cambiantes. A cada gen se le asignó un *grado de reacción* basado en la variabilidad de su perfil de expresión, definido como:

$$Grado\_reaccion(genA) = \sum_{i=1}^{n} \log_2^2 \left(\frac{exprA_i}{expr\_refA} \right),$$

donde:

- $exprA_i$ es el nivel de expresión del $genA$ bajo la condición experimental i,
- $expr\_refA$ es el nivel de expresión del $genA$ en la muestra de referencia,
- n es el número de condiciones experimentales.

Considérese la matriz de expresión que se muestra en la Tabla 3.2 formada por 10 genes (filas) y 5 condiciones experimentales (columnas). El grado de reacción del GEN3 viene dado por:

$$\begin{aligned} Grado\_reaccion(GEN3) &= \sum_{i=1}^{5} GEN3_i^2 = \\ &= 0{,}008^2 + 0{,}349^2 + (-0{,}212)^2 + 0{,}004^2 + 0{,}031^2 = \\ &= 0{,}168 \end{aligned}$$

Finalmente, relacionada con el nivel de expresión de un gen y con su grado de reacción, se encuentra la presencia (o ausencia) de la caja TATA, un elemento conservado en el promotor de los genes que participa en la iniciación de la transcripción. Tirosh et al. [256] describieron la presencia de la caja TATA en 585 genes de la levadura y su ausencia en 2492 genes. Asimismo, señalaron que dicha secuencia tiende a estar presente en el promotor de genes con ciertas funciones particulares.

3.2.3 Anotaciones de la Gene Ontology

La versión de la Gene Ontology de Febrero de 2007 fue descargada en formato MySQL para su utilización en el estudio. Para cada gen se obtuvo la lista de términos GO en los que aparecía explícita o implícitamente anotado. Se descartaron aquellas anotaciones con código de evidencia “IEA”, dado que no representan una fuente fiable de información.

Tabla 3.2: Ejemplo de matriz de expresión (cada valor es el $log_2(expr/ref)$).

ID GEN	COND1	COND2	COND3	COND4	COND5
GEN1	-1.064	-0.112	-0.340	-0.008	-1.040
GEN2	-0.844	-0.456	-0.277	0.035	-0.818
GEN3	0.008	0.349	-0.212	0.004	0.031
GEN4	1.221	-0.646	-0.032	-0.274	1.460
GEN5	0.628	-0.749	-0.192	0.201	0.391
GEN6	1.470	0.839	-0.312	-0.137	1.770
GEN7	0.703	0.167	0.145	-0.035	0.774
GEN8	1.364	0.271	0.240	-0.039	1.919
GEN9	1.922	-0.064	-0.630	-0.033	1.451
GEN10	3.083	-0.352	1.873	0.089	3.577

3.2.4 Datos obtenidos de experimentos con microarrays

La replicación del ADN, la división de los cromosomas y la mitosis, son eventos que definen una periodicidad fundamental en el ciclo celular eucariota. La transición precisa entre estos estados es crítica para garantizar la integridad y la supervivencia celular. De hecho, la desregulación celular desencadena inestabilidad genómica [113], y se cree que juega un papel fundamental en la etiología de cánceres tanto hereditarios como esporádicos [127, 272, 225, 280, 107]. Así, se han observado fluctuaciones en el nivel de mARN transcrito de diversos genes involucrados en diferentes procesos celulares, como el control de la transcripción [278, 184], la capacidad de respuesta a estímulos externos [293, 183], o la localización celular de proteínas [221]. Además, estudios genéticos han demostrado que la actividad de las proteínas reguladoras del ciclo celular es necesaria para la reparación del ADN [189, 276, 275], la meiosis [135, 268] y el desarrollo multicelular [103, 252, 253, 79]. Todas estas observaciones sugieren que todas las células eucariotas experimentan cambios fisiológicos importantes durante el ciclo

celular, y que una serie de eventos biológicos dependen del mantenimiento de esta periodicidad.

Cho et al. estudiaron en el trabajo referenciado en [63] los cambios en los niveles de expresión de 2879 genes de la levadura, durante dos ciclos celulares completos capturados en 17 instantes de tiempo. El trabajo señalado es una referencia indiscutible después de 11 años de su publicación, y los resultados siguen siendo aún un referente crucial en este campo. Así, se incluyó la información de este conjunto de datos en nuestro estudio, para buscar relaciones entre patrones de expresión en el ciclo celular y el resto de características del análisis.

Además del conjunto de datos anterior, se consideraron también los valores de expresión obtenidos por Gasch et al. [95]. En este último trabajo los autores usaron microarrays de ADN para analizar los cambios de expresión génica en células de levadura bajo diferentes condiciones ambientales externas. Los organismos celulares requieren una serie de condiciones internas específicas para su óptimo desarrollo y funcionamiento. Así, los distintos organismos han desarrollado miles de estrategias diferentes para mantener su estado interno ante variaciones de las condiciones externas. En el caso de las levaduras por ejemplo, éstas tienen que hacer frente con frecuencia a variaciones en el tipo y en la cantidad de nutrientes que hay en el entorno, fluctuaciones de temperatura, osmolaridad y acidez de su entorno, así como a la presencia variable de agentes nocivos como radiaciones o productos químicos tóxicos. De esta forma, cuando las condiciones ambientales externas se modifican de forma repentina, la célula tiene que ajustar rápidamente su expresión génica para adaptarse a las nuevas condiciones. El complejo sistema celular para detectar y responder a la variación de las condiciones ambientales externas es todavía ampliamente desconocido.

3.2.5 Preprocesamiento de los datos

Previo a la aplicación del algoritmo de extracción de reglas de asociación difusas fue necesario el preprocesamiento de los datos. En primer lugar, se utilizaron algoritmos de bicluster para extraer patrones de expresión de los

datos de experimentos con microarrays. Además, fue necesario definir conjuntos difusos sobre los dominios de aquellas variables cuyos dominios son continuos. Finalmente, se realizó una selección de las anotaciones GO introducidas en el análisis para mejorar la interpretabilidad de los resultados.

Extracción de patrones de expresión de los experimentos con microarrays

Tal y como se ha comentado anteriormente, en este estudio se ha hecho uso de los datos de expresión generados por los trabajos de Cho et al. [63] y Gasch et al. [95]. El primero de ellos proporcionó una matriz de expresión compuesta por los niveles de expresión de 2879 genes en 17 instantes de tiempo, que cubren dos ciclos celulares completos. El segundo de los trabajos generó una matriz de expresión de 6152 genes y 172 condiciones experimentales. Ambos conjuntos de datos fueron tratados de forma independiente en este trabajo.

La mayoría de las propuestas previas para analizar datos de microarrays mediante la extracción de reglas de asociación, discretizan los niveles de expresión definiendo etiquetas lingüísticas crisp tales como *SOBRE-EXPRESADO* ó *SUB-EXPRESADO* [70, 52]. Sin embargo, esta estrategia presenta algunos problemas: en primer lugar, es necesario determinar los umbrales que definan dichas etiquetas lingüísticas, lo que conlleva a su vez cierta pérdida de información que acompaña a todo proceso de discretización. Pero además, y probablemente el inconveniente más importante, es necesario introducir una nueva variable en el conjunto de datos por cada condición experimental de la matriz de expresión, es decir, 17 nuevas variables en el caso de Cho et al. y 172 en el caso de Gasch et al. Esto significa que se generarían una ingente cantidad de itemsets y reglas relacionando niveles de expresión, lo que dificultaría en gran medida la interpretación del conjunto de reglas resultante, siendo realmente complicado identificar perfiles de expresión y relacionarlos con el resto de características.

En una primera aproximación, se investigó el uso de algoritmos de clústering para obtener grupos de genes con perfiles de expresión similares [164]. De esta forma, tan sólo una nueva variable indicando el(los) cluster(s) a los

que pertenecía cada gen tenía que ser introducida en la tabla de datos, con lo que se evitaba generar miles de reglas relacionando niveles de expresión. Sin embargo, tal y como ya se ha comentado, los algoritmos de clústering proporcionan grupos de genes que se comportan de forma similar bajo **todas** las condiciones experimentales del estudio. Dado que este hecho no representa adecuadamente la realidad, se decidió utilizar algoritmos de biclústering [167]. Mediante este tipo de técnicas se obtienen grupos (biclusters) de genes que se comportan de forma similar bajo ciertas de las condiciones experimentales (no necesariamente todas ellas).

Distintos algoritmos de biclústering podrían dar lugar a diferentes conjuntos de biclusters [203, 242]. Aunque los resultados obtenidos por dos algoritmos diferentes probablemente coincidirían en gran medida, podrían aparecer pequeñas diferencias entre los biclústers comunes, pero además, podrían encontrarse biclusters con uno de los algoritmos que no se obtienen con el otro y viceversa. Así, con el objetivo de capturar adecuadamente los perfiles de expresión existentes, se utilizaron dos algoritmos de biclústering diferentes. Los dos métodos empleados en este caso fueron el *Gene-&Sample shaving* y el *EDA biclustering algorithm*, cuyo buen rendimiento en el análisis de microarrays ha sido previamente demostrado [49]. El primero de los métodos utiliza el cálculo de componentes principales (PCA) para identificar los biclústers, extendiendo así el algoritmo *Gene Shaving* propuesto por Hastie et al. [114]. El segundo hace uso de un tipo particular de algoritmos evolutivos (*Estimation of Distribution Algorithms*, o EDAs), para identificar biclústers en matrices de expresión génica. Ambos algoritmos buscan biclústers que presenten amplias variaciones de niveles de expresión entre muestras. De esta forma, los grupos obtenidos con estos métodos están formados por genes que presentan comportamientos muy diferentes en las distintas muestras (genes que participen en procesos constantemente activados, así como aquellos que no participen en ninguno de los procesos activos son ignorados), por lo que pueden ser muy útiles para identificar los distintos tipos de muestras y características que producen estas diferencias.

Definición de conjuntos difusos en dominios continuos

De todas las variables descritas hasta el momento, varias de ellas toman sus valores en dominios continuos (Tabla 4.7). En algunos casos, aunque el dominio de la variable no es estríctamente continuo (por ejemplo longitud del gen, longitud del intergénico, etc.), se consideraron como tales, ya que por ejemplo, no tiene sentido contar ocurrencias en la base de datos de genes que presenten exactamente el mismo número de bases. Así, se definieron tres conjuntos difusos en los correspondientes dominios representando las etiquetas lingüísticas BAJO, MEDIO y ALTO. Las definiciones se llevaron a cabo utilizando los percentiles p_{20}, p_{40}, p_{60} y p_{80}, siguiendo la recomendación del experto, tal y como se muestra en la Figura 3.6. Asimismo, se llevaron a cabo experimentos definiendo más conjuntos difusos (4, 5 y 6) en cada dominio (Figura 3.6). Sin embargo, los resultados obtenidos no mejoraron los conjuntos de reglas obtenidos con sólo 3 etiquetas. El uso de sólo 3 etiquetas lingüísticas proporciona conjuntos de reglas más claros y mejora el rendimiento de la metodología, ya que se genera una menor cantidad de itemsets. Así, se decidió que no era necesaria una mayor granularidad en el análisis y tan sólo 3 conjuntos difusos se utilizaron finalmente. La Figura 3.7 muestra la distribución de los valores sobre los que se definieron los conjuntos difusos, así como los valores de los percentiles p_{20}, p_{40}, p_{60} y p_{80}. Las líneas verticales rojas indican la localización aproximada de los percentiles 20, 40, 60 y 80. Parte del extremo derecho de los histogramas para el grado de reacción y la abundancia de proteínas ha tenido que ser omitido. La Sección omitida es similar al extremo de las gráficas que se muestran, es decir, la parte que no se muestra presenta también pocos puntos muy dispersos

Anotaciones GO

Para cada gen es necesario incluir la lista de nodos GO en los que aparece anotado. En este punto se pueden considerar varias estrategias:

- *Incluir en la lista únicamente los términos en los que los genes están explícitamente anotados*. Sin embargo, esta metodología puede acarrear diversos problemas:

Tabla 3.3: Resumen de variables incluidas en el estudio.

VARIABLE	POSIBLES VALORES
Longitud del gen	Z^+
Proporción de G+C en el gen	$[0,1]$
GC3s	$[0,1]$
Longitud del intergénico	Z^+
Proporción de G+C en el intergénico	$[0,1]$
Orientación	$Divergente, Tandem$
Cantidad de proteína	$[0,\infty[$
Grado de reacción	$[0,\infty[$
Caja TATA	si, no
Anotaciones GO	Términos GO
Biclusters	Z^+

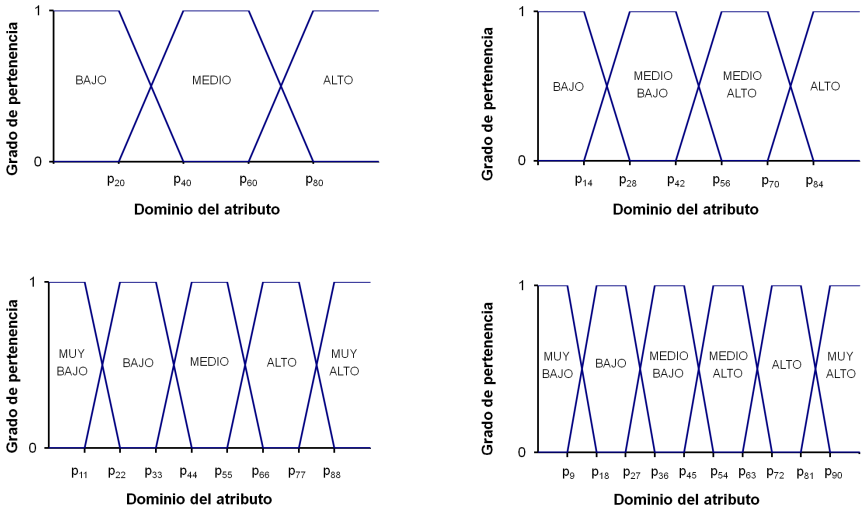


Figura 3.6: Definiciones de 3, 4, 5 y 6 conjuntos difusos.

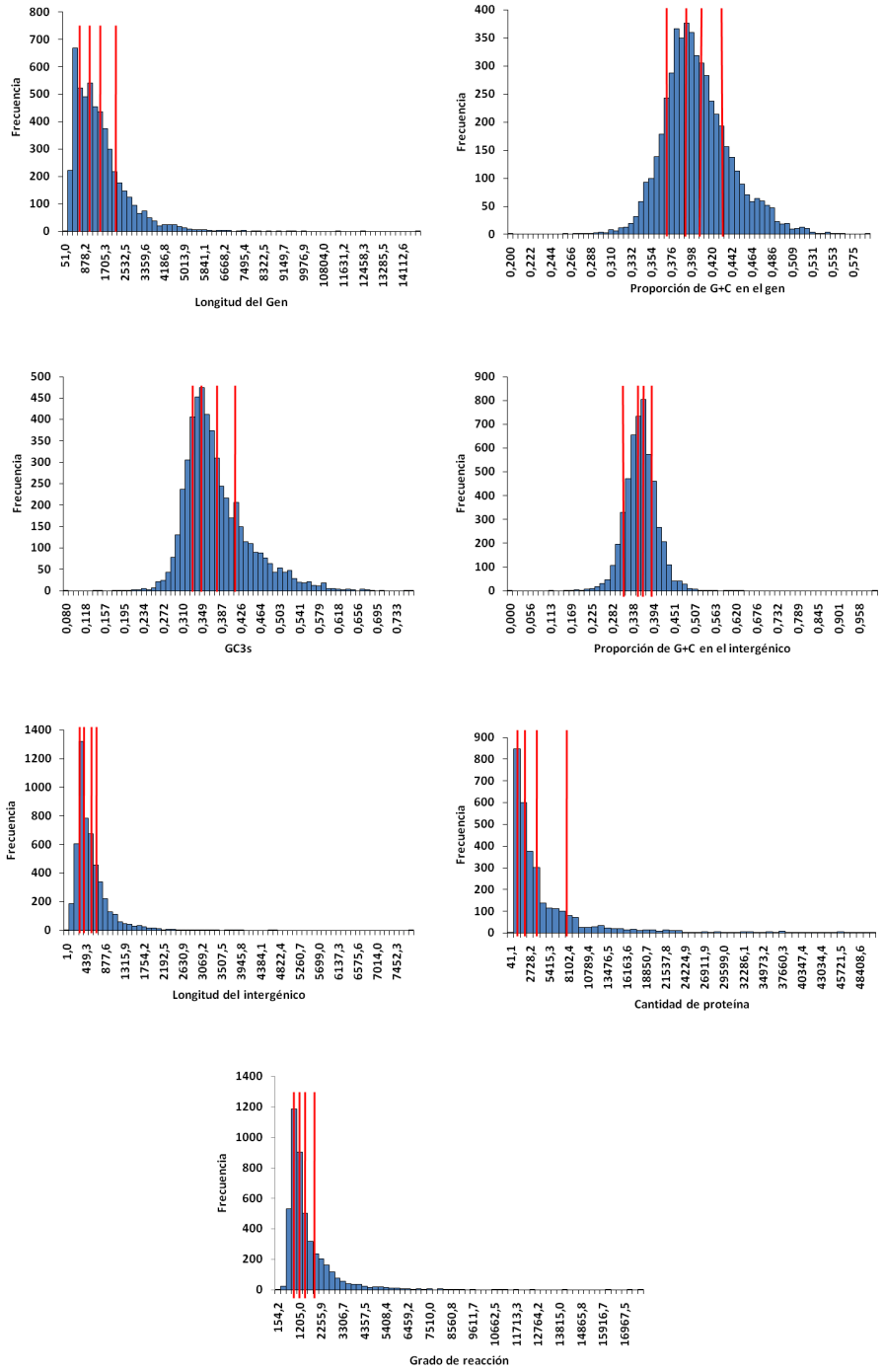


Figura 3.7: Distribución de los valores de los atributos sobre los que se definieron los conjuntos difusos.

  - Al considerarse únicamente los términos en los que los genes aparecen anotados explícitamente, éstos términos serán muy específicos, lo que significará que tendrán pocos genes anotados y, por tanto, los items correspondientes no podrán ser frecuentes. Esto puede ocurrir en una gran cantidad de casos [52].
  - Supóngase un conjunto de genes anotados en un término T, y otro conjunto diferente de genes anotados en un término $T^{'}$, donde $T^{'}$ es un ancestro de T. Al contar las ocurrencias de los itemsets que contienen $T^{'}$ en el conjunto de datos, no se tendrían en cuenta aquellos genes anotados en T, ya que estos últimos tan sólo contienen el término T en sus transacciones. Dado que cada término hereda las propiedades de todos sus ancestros, todos los genes anotados en T deben considerarse al calcular la frecuencia del término $T^{'}$, ya que de no ser así se estaría perdiendo información importante.

- *Incluir todos los términos en los que los genes aparecen explícita o implícitamente anotados.* Esta estrategia permite evitar todos los anteriores problemas y es de hecho la estrategia seguida por Martinez et al. [170]. Sin embargo, surge un inconveniente importante al seguir esta metodología: si todos los ancestros se incluyen en el analisis, se estarán considerando términos muy generales (por ej. *molecular_function*, *biological_process*, *cellular_component*, etc.). Estos términos son tan generales que no proporcionan ninguna información. Es más, ralentizan la búsqueda de reglas de asociación y entorpecen la interpretación del conjunto de reglas final, ya que generan una gran cantidad de reglas triviales o carentes interés.
- *Fijar un nivel de GO e incluir tan sólo los términos de este nivel. Las anotaciones en términos que aparecen por debajo del nivel seleccionado se mapean a los nodos correspondientes de ese nivel, y aquellas de términos superiores se descartan*. Algunas aplicaciones como FatiGO [10] han seguido esta estrategia. En principio parece que el nivel 3 representa un buen compromiso entre calidad de la información y número de genes

anotados [172]. Sin embargo, esta opción presenta el problema de que la especificidad de los nodos GO de un mismo nivel no es homogénea. Es decir, en un mismo nivel puede haber términos más generales y términos más específicos, por lo que se correría el riesgo de perder información [16].

- *Incluir todos los términos en los que están anotados los genes (explítica o implícitamente), determinar la información aporta cada uno y descartar aquellos no informativos*. Esta parece ser la opción más adecuada y es de hecho la que se ha escogido. Asumiendo que un término aporta más información cuanto más específico es, el contenido de información (CI) de un nodo se puede calcular mediante la expresión:

$$CI(T) = \frac{-\log P(T)}{-\log P(min)},$$

donde $P(T)$ representa la probabilidad de que *ocurra* el nodo T. Se dice que un nodo *ocurre* cuando aparece un gen anotado en este nodo o en alguno de sus descendientes. Por tanto, la probabilidad de que ocurra el nodo se calcula como el número de anotaciones que hay en él y por debajo de él, dividido entre el total de anotaciones (Figura 3.8). El denominador de la expresión ($-\log P(min)$) se utiliza como factor de normalización, de modo que el valor del contenido de información se encuentre entre 0 y 1. $P(min)$ representa la probabilidad mínima de que ocurra un nodo, lo que sucederá con aquellos nodos que tengan una única anotación, es decir:

$$P(min) = 1/NumeroTotalDeAnotaciones$$

Obsérvese que cuanto más profundo se encuentre un término en la ontología, mayor será su CI. Este hecho se debe a la estructura ontológica de GO: si el número de anotaciones decrece, la probabilidad de ocurrencia de los términos también decrece y, por tanto, su CI tiende a aumentar (Figura 3.8). Tras estudiar los resultados obtenidos, se ha determinado que un valor apropiado para el umbral del contenido de información es 0,2.

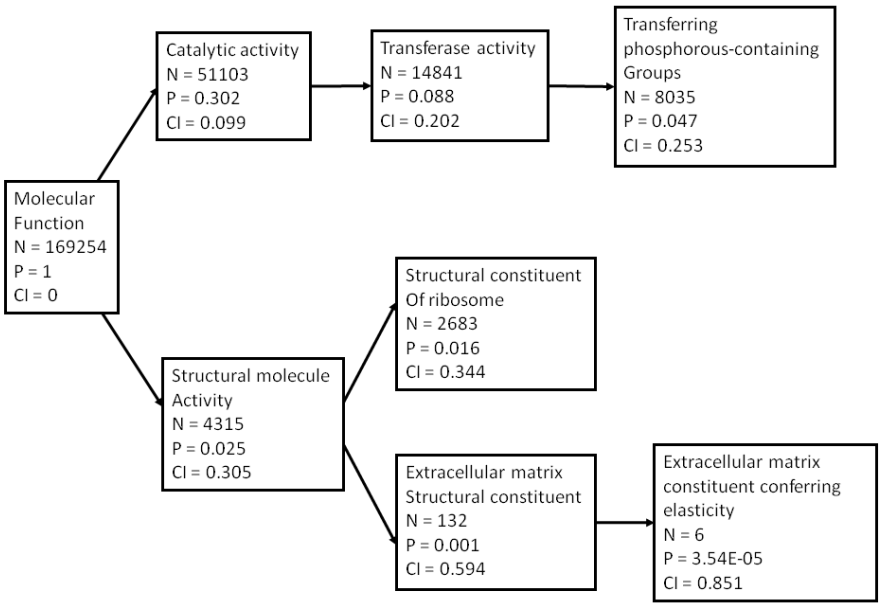


Figura 3.8: Fragmento de la ontología *molecular_function* de GO. Cada nodo se etiqueta con su nombre, el número de anotaciones en él y bajo él (N), la probabilidad derivada del número de anotaciones (P) y su contenido de información (CI). El valor de $NumeroTotalDeAnotaciones$ utilizado para calcular las probabilidades se corresponde con el número de anotaciones del nodo raíz de la ontología, es decir, 169524 en este caso.

3.3 Extracción de reglas de asociación difusas

Los datos descritos en la Sección anterior se integraron en una tabla de 4363 genes y 13 variables. Dicha tabla puede considerarse como una tabla transaccional en la que cada gen representa una transacción y los pares *atributo-valor* representan los items de las transacciones.

En una primera fase se llevó a cabo una implementación difusa y la posterior aplicación del algoritmo Apriori. Sin embargo, el número de ítems que se generan, principalmente debido a la gran cantidad de anotaciones GO, y por tanto de itemsets que se forman es ingente, haciendo imposible el uso de este algoritmo. Por ello, se desarrolló y aplicó una versión difusa del algoritmo Top-Down Frequent-Pattern Growth (Fuzzy TD FP-Growth) [163].

3.3.1 Algoritmo Fuzzy Top-Down Frequent-Pattern Growth

Tal y como ya se ha comentado, la idea de este algoritmo consiste en acelerar el proceso de búsqueda de itemsets frecuentes. Han et. al. [109] proponen el uso de una estructura de datos, el FP-tree, para almacenar la frecuencia de todas las combinaciones de ítems presentes en la tabla, de tal modo que sólo es necesario recorrer la tabla dos veces. Una vez almacenada toda la información referente a las frecuencias en este árbol, tan sólo hay que recorrerlo de una manera especial para poder conseguir todos los conjuntos frecuentes de ítems. En dicho trabajo se propone un algoritmo bastante complejo para recorrer el FP-tree, el cual requiere de la creación de una gran cantidad de estructuras de datos intermedias. Posteriormente, Wang et. al. [271] propusieron un método para recorrer el árbol que no necesita de la creación de estas estructuras de datos intermedias. Éste se denomina Top-Down Frequent-Pattern Growth (TD-FP-Growth). Dicho algoritmo en su forma original es tan sólo apropiado para manejar datos crisp, por lo que se tuvo que desarrollar una versión difusa del mismo capaz de tratar itemsets difusos, el Fuzzy Top-Down Frequent-Pattern Growth (FTD-FP-Growth). El algoritmo se puede resumir en las siguientes etapas:

1. Escaneo de la base de datos para la obtención de una lista de items frecuentes.
2. Construcción del Fuzzy Frequent-Pattern tree (FFP-tree).
3. Obtención de la lista de itemsets frecuentes recorriendo para ello el Fuzzy Frequent-Pattern tree.

Obtención de la lista frecuente de items

Tal y como ya se ha comentado, el primer paso consiste en llevar a cabo un escaneo inicial de la base de datos para obtener una lista con los items frecuentes presentes en la base de datos. A continuación, esta lista es ordenada en función del soporte decreciente de los items. Esta disposición de los items en la lista determinará el orden en el que posteriormente se introducirán los mismos en el FFP-tree. La razón para esta ordenación decreciente del soporte se basa en que se ha demostrado que, aunque no existe una disposición

Tabla 3.4: Ejemplo de lista ordenada de items.

Índice	Item	Soporte
1	{Longitud del intergénico = Corto}	8,48
2	{Orientación = Tandem}	7
3	{Orientación = Divergente}	6
4	{Longitud del gen = Medio}	4,95
5	{Longitud del gen = Corto}	4,59
6	{Longitud del gen = Largo}	4,13
7	{Longitud del intergénico = Largo}	2,97
8	{Longitud del intergénico = Medio}	1,93
...	...	...

óptima para todos los casos, el orden decreciente de soporte suele generar estructuras de FFP-trees más eficientes [109]. La Tabla 3.4 muestra un ejemplo de lista de items frecuentes.

A cada item se le asigna un código que se corresponde con el índice de la posición en la que se encuentra en la lista ordenada. Así, por ejemplo, en la lista de la Tabla 3.4 Longitud del gen = Medio tiene el código 4 y el item Orientación = Tandem tiene el código 2.

Construcción del Fuzzy Frequent-Pattern tree

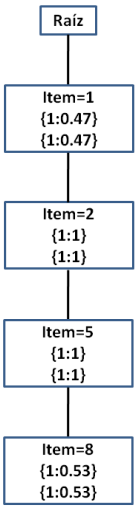
Una vez se dispone de la lista se van tomando una a una las transacciones de la tabla de datos, e introduciendo en el árbol de forma ordenada los items de la lista que aparecen en cada fila. Cada nodo del árbol representará un item en una posición determinada de un conjunto(s) de items y contendrá un par de listas de pertenencias, de forma que cada vez que se recorra ese nodo durante el proceso de creación del árbol se actualizarán dichas pertenencias. Por ejemplo, la Figura 3.9 muestra de forma secuencial la inserción de las tres primeras transacciones de la Tabla 3.5.

Tabla 3.5: Ejemplo de tabla transaccional difusa. Los valores que aparecen en cada item tras los dos puntos indican la pertenencia de la transacción al item difuso.

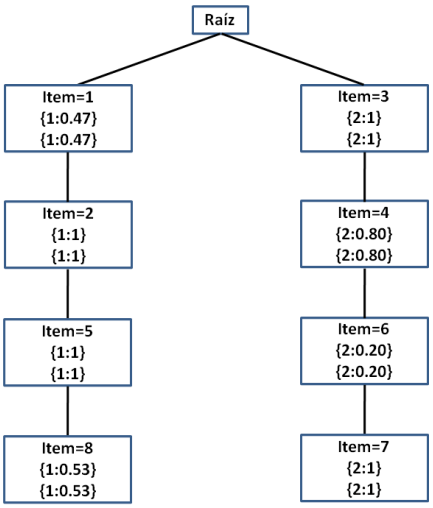
ID Transacción	Items
*YAL*008*W*	{ *Long. gen*=*Corto* : 1, *Long. interg.*=*Medio* : 0,53, *Long. interg.*=*Corto* : 0,47, *Orientación*=*Tandem*}
*YAL*003*W*	{ *Long. gen*=*Largo* : 0,20, *Long. gen*=*Medio* : 0,80, *Long. interg.*=*Largo* : 1, *Orientación*=*Divergente*}
*YAL*018*C*	{ *Long. gen*=*Medio* : 0,67, *Long. gen*=*Corto* : 0,33, *Long. interg.*=*Largo* : 0,82, *Long. interg.*=*Medio* : 0,18, *Orientación*=*Divergente*}
*YAL*002*W*	{ *Long. gen*=*Largo* : 1, *Long. interg.*=*Medio* : 0,67, *Long. interg.*=*Largo* : 0,33, *Orientación*=*Tandem*}
*YAL*009*W*	{ *Long. gen*=*Largo* : 1, *Long. interg.*=*Corto* : 1, *Orientación*=*Divergente*}
*YAL*010*C*	{ *Long. gen*=*Largo* : 0,93, *Long. gen*=*Medio* : 0,07, *Long. interg.*=*Corto* : 1, *Orientación*=*Divergente*}

Nótese que, como ya se ha comentado, cada nodo almacena dos listas de pertenencias. Estas dos listas de pertenencias, que inicialmente contendrán los mismos valores, aparecen representadas en el árbol mediante parejas de la forma $transaccion : pertenencia$. Estos pares, indican la pertenencia del item correspondiente, a las diferentes transacciones que "van recorriendo" ese nodo durante la construcción del árbol. Así por ejemplo, el par $1 : 0{,}33$ del nodo correspondiente al item 7, indica que la pertenencia del item 7 en la transacción 1 es 0,33.

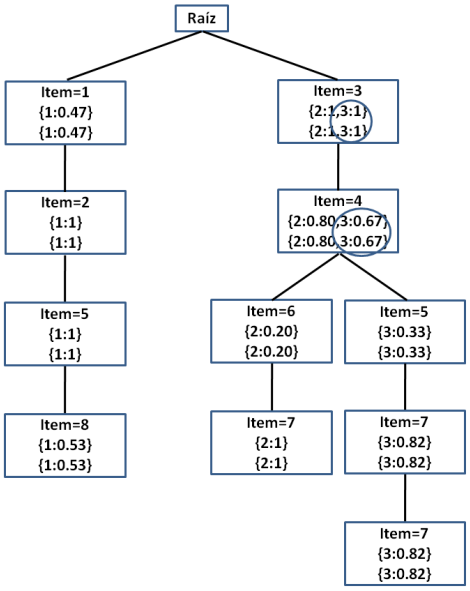
Además de todo lo comentado hasta el momento, conforme se genera el árbol se va construyendo una tabla cabecera H. Esta tabla contiene una entrada por cada item introducido en el árbol. Cada entrada de la tabla contiene la lista de pertenencias del item correspondiente, así como una lista de punteros a los nodos de dicho item (Figura 3.10). En el Algoritmo 1 se muestra



(a) Primera transacción.



(b) Segunda transacción.



(c) Tercera transacción.

Figura 3.9: Se introducen las tres primeras transacciones de la Tabla 3.5 en el FFP-tree.

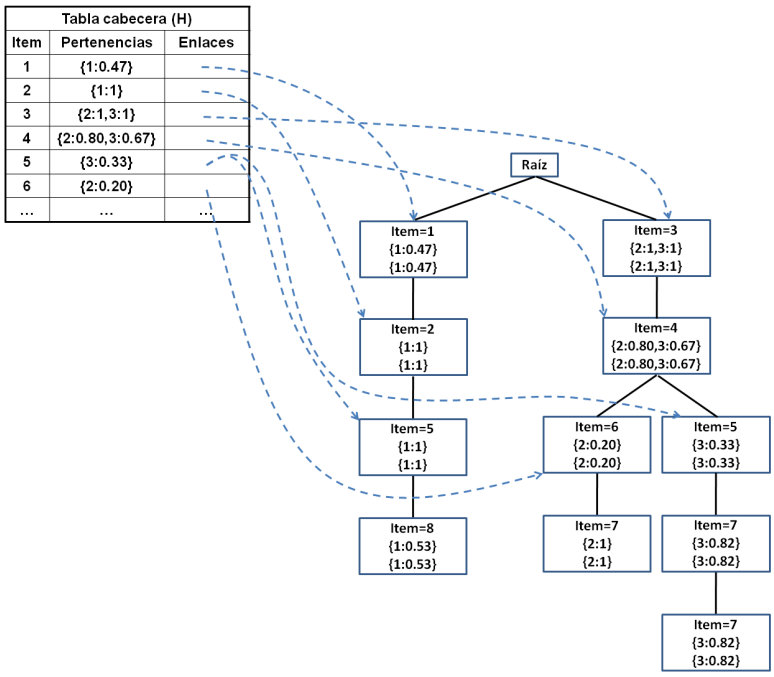


Figura 3.10: Árbol junto con la tabla cabecera.

el pseudocódigo para construir el FFP-tree a partir de la tabla transaccional.

Generación de itemsets frecuentes a partir del FFP-tree

El procedimiento parte de la tabla H y del FFP-tree construidos mediante el algoritmo que se muestra en el pseudocódigo anterior. Los items que se encuentran en la tabla H deben aparecer ordenados según sus códigos.

Se irán considerando una a una las entradas de la tabla H. Mediante los enlaces de cada ítem en H se podrán localizar en el árbol los nodos de todos los ítems. Partiendo de estos nodos y recorriendo el árbol hacia arriba, se irán obteniendo todos los itemsets frecuentes cuyo último ítem es el de la entrada en H que se está considerando. El Algoritmo 2 muestra en pseudocódigo el proceso. Nótese que de las dos listas de pertenencias que contiene cada nodo, una de ellas es considerada “auxiliar” y se va modificando conforme se desarrolla el proceso, mientras que la otra permanece inalterada hasta el final.

Algoritmo 1 Pseudocódigo para construir el FFP-tree

```
raiz=nodo()
H=[] {Tabla cabecera}

Para cada item hacer
  H[item].pertenencias=pertenencias del item para cada transacción
  H[item].enlaces = []
fin

Para cada transacción de la tabla hacer
  nodoActual = raiz
  Para cada item de la lista ordenada hacer
    Si el item está en la transacción actual entonces
      Si el nodo actual no tiene un hijo con este item entonces
        aux=nodo()
        aux.item=item
        aux.pertenencias={transaccion:pertenencia}
        aux.pertenencias_aux={transaccion:pertenencia}
        aux.padre=nodoActual
        H[item].enlaces+=referencia a aux
        nodoActual=aux
      en caso contrario
        nodoActual.hijos[item].pertenencias+={transaccion:pertenencia}
        nodoActual.hijos[item].pertenencias_aux+={transaccion:pertenencia}
        nodoActual=nodoActual.hijos[item]
      fin
    fin
  fin
fin
```

Algoritmo 2 Función buscaItemsetsFrecuentes(X, H)

X: parámetro de entrada, contendrá una lista con los items que forman el itemset actual. En la llamada inicial a la función contendrá una lista vacía.
H: parámetro de entrada, contendrá la tabla cabecera.

Lista=[] {Aquí se guardarán los itemsets frecuentes}

Para cada entrada I en H **hacer**
  Si H[I].contador $\geq$ soporteMinimo **entonces**
    Lista = Lista + IX
    Crea una nueva tabla cabecera H_I llamando a crearSubTabla(I) (ver Algoritmo 3)
    aux=buscaItemsFrecuentes(IX, H_I)
    Lista+=aux
  fin
fin

Devolver Lista

Algoritmo 3 Función crearSubTabla(I)

I: parámetro de entrada, contendrá una entrada de la tabla de cabecera.

Para cada nodo u en las referencias de I.enlaces, subir por el árbol desde u, y cada vez que se recorra un nodo v **hacer**
  aux = min{u.pertenencias_aux, v.pertenencias}
  Si no hay una entrada en H_I para el item del nodo v **entonces**
    H_I[v.item].pertenencias=aux
    H_I[v.item].enlaces+=referencia a v
    v.pertenencias_aux=u.pertenencias_aux
  en caso contrario
    Actualizar H_I[v.item].pertenencias con los valores de aux
    Si en esta llamada a crearSubTabla no se había visitado v **entonces**
      v.pertenencias_aux = aux
      H_I[v.item].enlaces+=referencia a v
    en caso contrario
      Actualizar v.pertenencias_aux con los valores de aux
    fin
  fin
fin

Se realizarán algunas etapas del procedimiento partiendo del estado de la Figura 3.10. Para simplificar considérese que el soporte mínimo es 0. En primer lugar se procesa la primera entrada de la tabla H, es decir, la correspondiente al item 1 (Figura 3.11). El único nodo que hay en el árbol correspondiente al item 1 es el que aparece coloreado en rojo. Puesto que es un item frecuente, se procede a crear la subtabla correspondiente. Dado que este nodo se encuentra en el primer nivel del árbol, al ascender se llega en el primer paso al nodo raíz, por lo que no hay que crear ninguna subtabla y se pasa a la siguiente entrada de H (Figura 3.12). Nuevamente aparece un sólo nodo para el item 2, y puesto que el item es frecuente, se crea la subtabla H_2. Para ello se procede a subir en el árbol. El primer nodo que aparece al subir es el que está coloreado en verde. Dado que no hay una entrada en H_2 que contenga el item 1, se crea dicha entrada:

1. El item de la nueva entrada en H_2 se corresponde con el item del nodo recorrido, es decir, en este caso el item 1.
2. El vector de pertenencias de la nueva entrada se calcula como el mínimo entre las pertenencias auxiliares del nodo de partida y las pertenencias del nodo recorrido:

$$pertenencias\_en\_H_2 = min(\{1:1\},\{1:0{,}47\}) = \{1:0{,}47\}$$

   De igual forma, el vector de pertenencias auxiliar del nodo recorrido pasa a ser también este vector mínimo (Figura 3.13).
3. Se añade una referencia en la nueva entrada de la tabla apuntando al nodo recorrido.

A continuación se sigue ascendiendo en el árbol. Como ya se ha alcanzado el nodo raíz, se hace una llamada recursiva de la función buscaItemsetsFrecuentes con el item 2 y la tabla cabecera H_2 como parámetros. Se procede por tanto a procesar las entradas de la tabla H_2: ésta tan sólo tiene una entrada para el item 1 y, como su contador es mayor que el mínimo soporte (recuérdese que se está suponiendo soporte mínimo 0 para simplificar), el itemset $\{1,2\}$ es frecuente y se añade a la lista de salida. En este momento se procede a crear la nueva subtabla H_{21}. Como el único nodo al que hay

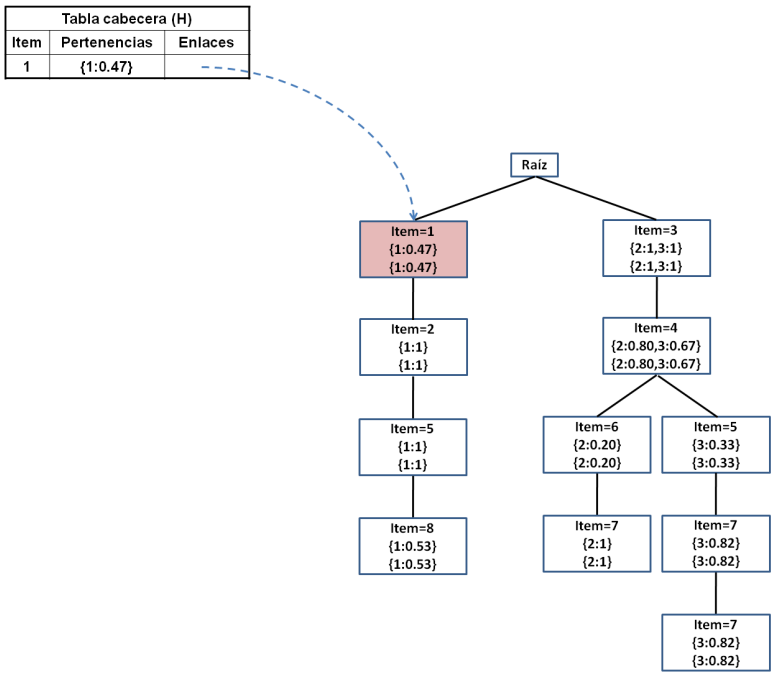


Figura 3.11: Recorriendo el FFP-tree, primera entrada de H.

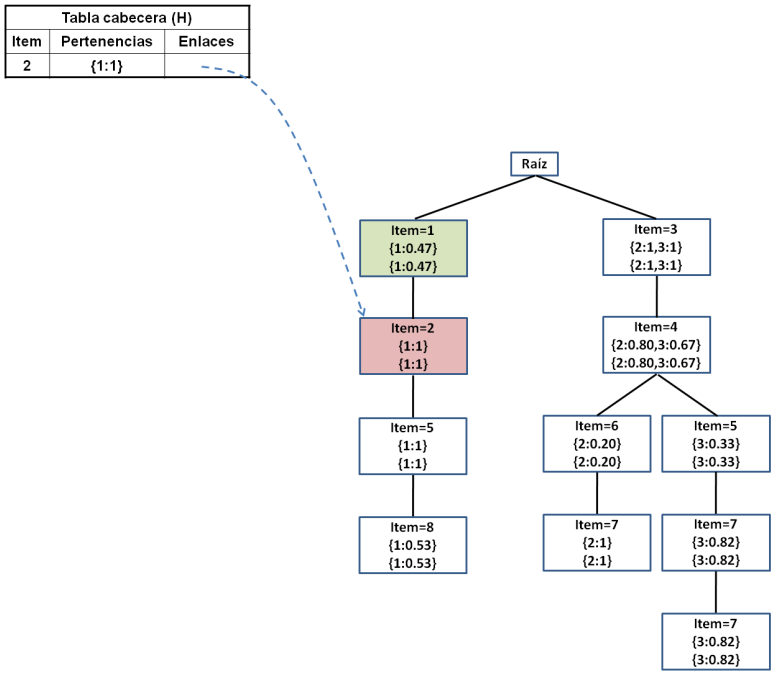


Figura 3.12: Recorriendo el FFP-tree, segunda entrada de H.

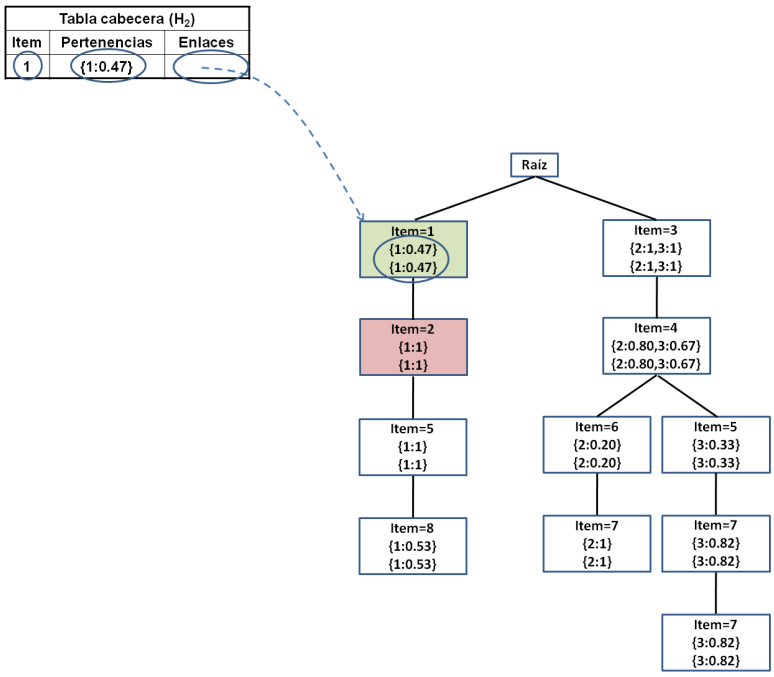


Figura 3.13: Recorriendo el FFP-tree, creando la tabla H_2.

un enlace desde la primera entrada de la tabla H_2 cuelga directamente del nodo raíz, se detiene el proceso y no se crea la tabla H_{21}. Llegado a este punto se continuaría procesando la siguiente entrada de la tabla H, y así hasta que todas las entradas de esta tabla y de todas las subtablas que se fueran creando hubieran sido procesadas. Al finalizar todo el proceso se habrá obtenido la lista con los itemsets frecuentes. Tan sólo queda obtener las reglas de asociación a partir de los mismos.

3.3.2 Obtención y procesamiento de las reglas de asociación difusas

Tal y como se ha comentado con anterioridad, la última fase en la extracción de reglas de asociación consiste en obtener dichas reglas a partir de los itemsets frecuentes. En este apartado se describen las medidas empleadas para evaluar la calidad de las reglas, así como el post-procesamiento llevado a cabo sobre el conjunto final de reglas.

Obtención de las reglas de asociación a partir de los itemsets frecuentes

El procedimiento para obtener las reglas de asociación a partir de los itemsets es común a todos los algoritmos de extracción de reglas de asociación, por lo que no describirá detalladamente. Dado un itemset, la idea consiste en formar subconjuntos con sus elementos, cada uno de los cuales formará el consecuente de una regla. Cada una de estas reglas se completa colocando en el antecedente el resto de ítems del itemset que no aparecen en el consecuente. Así, en principio, se generará una regla por cada subconjunto de items obtenido. Los valores de confianza y soporte difusos se calculan para cada una de estas reglas, y tan sólo se generan aquellas que presenten valores superiores a los umbrales. Sin embargo, la eficiencia del procedimiento puede mejorarse, ya que no es necesario considerar absolutamente todos los subconjuntos de items de un itemset. Una información más detallada se encuentra en el trabajo referenciado en [8]. Finalmente, tan sólo queda mencionar en este apartado que los valores de soporte y confianza difusos fueron calculados siguiendo el procedimiento descrito por M. Delgado et al. [76].

Deficiencias del modelo soporte/confianza

Tal y como se comentó en la introducción, el número de reglas que se generan es enorme, muchas de ellas redundantes o carentes de información. El modelo de soporte/confianza para medir la calidad de las mismas ha demostrado ser insuficiente y, debido a esto, se han propuesto gran cantidad de estrategias y medidas de interés que mejoran la interpretabilidad del conjunto de reglas. Así, se han propuesto diferentes propiedades que podrían ser deseables para una medida de calidad según diferentes situaciones [244, 97]. En este caso, se consideró que las medidas que se utilizaran debían cumplir las siguientes propiedades:

- Capacidad de discriminar la independencia de los atributos. Es decir, si $sop(X \rightarrow Y) = sop(X) \cdot sop(Y)$, la medida debe valer 0.
- Debe ser monótona creciente cuando crezca el $sop(X \rightarrow Y)$ y el resto de parámetros permanezcan invariables.

- Si $sop(X)$ crece y el resto de parámetros permanecen invariables, la medida debe decrecer. Asimismo, si $sop(Y)$ crece y el resto de parámetros permanecen invariables, el valor de la medida debe decrecer.
- No debe ser simétrica. Es decir, el valor de la medida para $X \rightarrow Y$, debe ser diferente que su valor para $Y \rightarrow X$
- Debe decrecer si $sop(X \neg Y)$ crece.
- Debe ser creciente con el soporte si los márgenes de la tabla de contingencia (Tabla 2.9) permanecen fijos.
- Debe guardar alguna relación con el número de registros que no contienen ni X ni Y.

Tan sólo se encontraron tres de las medidas propuestas hasta el momento que cumplieran todas estas propiedades: *added value*, *Klosgen's measure* y *certainty factors* (ó factores de certeza en español, FC)[97, 76, 148]. La definición de las tres medidas se muestra en la Tabla 3.6. Tal y como se puede observar, la medida de Klosgen incluye a la *add value*, ya que se trata de una combinación de ésta con el soporte de la regla. Sin embargo, no parece intuitivo el hecho de que la medida tome valores en el intervalo $]-1,1[$, ya que en ningún caso puede llegar a tomar el valor -1 o el valor 1. En otras palabras, no parece intuitivo por ejemplo, el hecho de que dada una regla "perfecta", no haya un valor máximo definido de la medida de calidad para la misma (por ejemplo 1). Por ello, se decidió finalmente hacer uso de los factores de certeza. Por tanto, tan sólo se generaron aquellas reglas con valores de soporte, confianza y factor de certeza mayores que los umbrales especificados por el usuario.

Finalmente, era necesario proporcionar un valor que mostrara de forma global la relevancia del conjunto de relaciones obtenido. Con este objetivo, se estimó el número de reglas obtenidas por azar: se generaron 100 conjuntos de datos aleatorios independientes a partir del conjunto de datos original y se aplicó la metodología de reglas de asociación sobre cada uno de ellos. El número estimado de reglas falsas se calculó como la media del número de reglas obtenido para cada uno de estos 100 conjuntos de datos aleatorios. De

Tabla 3.6: Medidas que cumplen las propiedades deseadas.

Medida	Fórmula
Add value	$conf(X \rightarrow Y) - sop(Y)$
Klosgen	$\sqrt{sop(X \rightarrow Y)}(conf(X \rightarrow Y) - sop(Y))$
Certainty Factor	$\frac{conf(X \rightarrow Y) - sop(Y)}{1 - sop(Y)}$

esta forma, calculamos un *False Discovery Rate* (FDR) que permitió verificar la calidad del conjunto de reglas.

Filtrado de las reglas utilizando la jerarquía GO

Aún considerando todo lo anterior, se obtienen una gran cantidad de reglas relacionando nodos GO con el resto de variables, muchas de ellas representando prácticamente la misma información. Estas reglas pueden fusionarse en una sola más general sin perder información relevante, reduciendo así de forma considerable el número de reglas obtenidas.

En primer lugar, se buscan grupos de reglas que contengan algún nodo GO y que compartan el resto de sus items. Para cada grupo, se busca un término GO que sea ancestro común del resto de nodos GO que aparecen en dicho grupo de reglas. Tan sólo la regla que contiene el ancestro común se mantiene y el resto se descarta. La estrategia se basa en el hecho de que cada término GO hereda los atributos de todos sus ancestros. Como además se está asegurando que todos los términos incluidos en el análisis son suficientemente informativos (mediante el umbral del CI), la experiencia ha mostrado que el ancestro común representa el término más intuitivo. Al eliminar el resto de reglas se obtiene un conjunto más pequeño, más claro y por tanto más fácilmente interpretable (Figura 3.14).

Aunque existe la posibilidad de perder algo de información al llevar a cabo el filtrado, esta pérdida se compensa por la ganancia en claridad del conjunto de reglas final. En cualquier caso, si durante el análisis del conjunto de reglas resultante se encuentra una relación de interés especial, el proceso

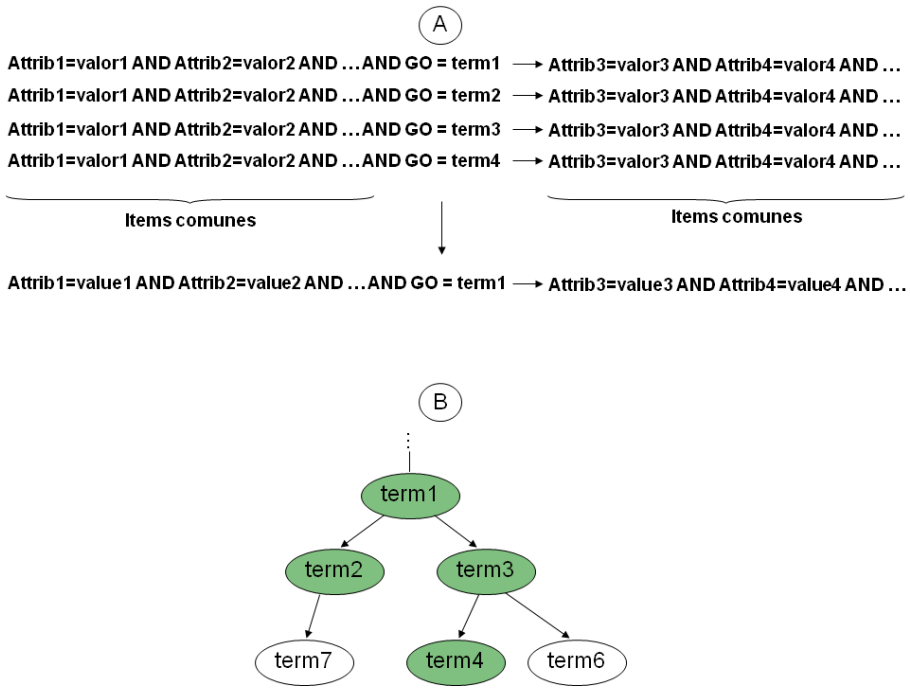


Figura 3.14: Filtrado de un conjunto de reglas utilizando la jerarquía GO. La parte A muestra un grupo de reglas que comparten todos sus items menos el término GO. Estas 4 reglas se fusionan en la que es un poco más general de todas ellas. La parte B de la figura muestra la distribución de los términos en la ontología.

de filtrado puede omitirse para obtener tanta información como sea posible referente a dicha asociación.

3.4 Biological data analysis by Fuzzy Association Rule mining: BioFAR

Tan importante como desarrollar una metodología, es facilitar el uso de la misma a los potenciales usuarios. En este sentido, y dado que no se debe esperar que alguien que no esté estrechamente relacionado con las Ciencias de la Computación implemente complejas técnicas computacionales, toda la metodología anteriormente expuesta fue desarrollada e implementada en una aplicación web de libre acceso: BioFAR (http://genome.ugr.es/biofar).

Figura 3.15: Página principal de BioFAR

Al menos hasta donde nosotros conocemos, no existe ninguna otra aplicación web basada en la extracción de reglas de asociación difusas y además orientada al análisis de datos genómicos. Con esta aplicación se pretende promover el uso de técnicas difusas en Bioinformática, proporcionando una herramienta que permite estudiar las relaciones existentes entre un elevado número de variables. La aplicación está implementada en Python, presenta una interfaz simple (Figura 3.15) y su correspondiente documentación, con el objetivo de proporcionar una herramienta fácil de usar para cualquier persona que no esté estrechamente relacionada con las Ciencias de la Computación.

Los datos de entrada que hay que proporcionar son fundamentalmente tres:

- *La tabla de datos* conteniendo la información que se analizará. Las columnas en esta tabla deben representar las variables del estudio. Dicha tabla puede proporcionarse en forma de un fichero de texto delimitado por tabuladores ó bien un fichero *csv*, los cuales pueden generarse fácilmente a partir de una hoja de cálculo de Excel. Se pueden incluir

variables tanto numéricas como categóricas, incluyendo listas de términos GO, que en este caso se proporcionarán en forma de identificadores GO separados por comas (por ejemplo GO:xxxxxxx,GO:yyyyyyy,...).

- *Las definiciones de los conjuntos difusos*, es decir, cuántas etiquetas lingüísticas hay que definir en cada variable numérica y cómo deben ser definidas. BioFAR permite llevar a cabo tres tipos diferentes de definiciones: 1) de forma automática, utilizando el algoritmo de clústering c-means difuso [237, 122], 2) de forma automática, definiendo funciones trapezoidales mediante percentiles y 3) de forma manual, fijando los puntos de inicio y final de las funciones trapezoidales.

  Además, los valores de expresión se podrán considerar como un caso especial de variable. Es decir, si una variable representa un valor relativo de expresión ($log_2(objetivo/control)$), el sistema permite definiciones especiales de conjuntos difusos sobre ésta. La razón para esto, está basada en que por lo general es aceptado que valores del $log_2(ratio)$ mayores que 1 representan sobre-expresión, y que valores menores que -1 representan sub-expresión.

- *Los umbrales de calidad de las reglas*, es decir, los umbrales para el soporte, la confianza y los factores de certeza. En caso de que se incluyan términos GO en el análisis, se debe proporcionar también el umbral del contenido de información.

Los resultados se devuelven en forma de fichero de texto. Asimismo, para facilitar la interpretación del conjunto de reglas final, se proporciona también un simple programa de filtrado de reglas (implementado también en Python, el código fuente y un instalador de Windows se pueden descargar de la página web de BioFAR). Este software permite fijar patrones de reglas de interés o modificar los umbrales de calidad para filtrar reglas no deseadas de forma rápida. Así, el programa toma como entrada el fichero de texto proporcionado por el servidor y genera un conjunto filtrado de reglas en forma de fichero *.html*. Los términos GO que aparecen en este fichero html se enlazan con la base de datos AmiGO [50], de forma que se puede obtener rápidamente una descripción detallada de los mismos. Finalmente, este soft-

ware permite también llevar a cabo el filtrado GO descrito previamente en la Sección 3.3.2.

3.5 Resultados

En esta Sección se describen los diferentes experimentos llevados a cabo así como los resultados obtenidos. Los umbrales de FC, confianza y soporte se seleccionaron de forma que el FDR resultante fuera bajo y se obtuviera un número asequible de reglas, ya que el conjunto resultante debía ser fácilmente interpretable por un experto. Los valores de los umbrales utilizados en cada experimento, el número total de reglas obtenidas, así como el FDR calculado en cada caso se detallan en la Tabla 3.7.

Al fijar los umbrales de calidad de las reglas, el usuario debe tener cierta información acerca del tipo de datos que se van a analizar. Por ejemplo, debería tener en cuenta que si el soporte de un itemset es muy alto, es muy probable que éste aparezca en el consecuente de asociaciones obtenidas simplemente por casualidad, ya que aparece en muchas transacciones de la tabla de datos [76]. Por tanto, se deberían esperar valores bajos del FC para las reglas que contengan estos itemsets (por ejemplo las reglas en la Tabla 3.8). Es decir, es necesario disponer de algún conocimiento previo acerca de la distribución de los datos, para poder determinar qué valores de las medidas de calidad se pueden considerar aceptables y, de acuerdo con esto, modificar los umbrales de calidad de las reglas. Sin embargo, en la mayoría de los casos esto no representa un problema, ya que este tipo de información se conoce. Por ejemplo, en este trabajo en concreto era fácil saber *a priori*, que el soporte de los itemsets $Orientacion = Divergente$ y $Orientacion = Tandem$ es aproximadamente 0,5, así como que el soporte de los items que contienen nodos GO es por lo general bastante bajo (normalmente menor que 0,01).

Tal y como se comentó en la Sección 3.3.2, la calidad global de los conjuntos de reglas obtenidos se midió mediante el FDR. Observando la Tabla 3.7 se puede ver que los FDRs obtenidos son bastante bajos, lo que indica que tan sólo unas pocas reglas se habrían generado por casualidad, y por

Tabla 3.7: Resumen de los experimentos.

Variables	Umbral FC & Conf.	Umbral soporte	Número total de reglas	FDR
Variables estructurales	0,1	0,01	24	0,093
Molecular Function & Variables estructurales	0,4	0,004	20	0,042
Biological Process & Variables estructurales	0,5	0,004	7	0,050
Cellular Component & Variables estructurales	0,5	0,004	12	0,011
Cantidad proteína & Capacidad reacción & Caja TATA	0,1	0,002	15	0,000
Cantidad proteína & Variables estructurales	0,1	0,002	4	0,040
Cantidad proteína & Molecular Function	0,2	0,002	19	0,109
Cantidad proteína & Biological Process	0,4	0,002	21	0,005
Cantidad proteína & Cellular Component	0,3	0,002	14	0,011
Capacidad reacción & Variables estructurales	0,1	0,002	10	0,044
Capacidad reacción & Molecular Function	0,3	0,002	23	0,069
Capacidad reacción & Biological Process	0,6	0,002	19	0,002
Capacidad reacción & Cellular Component	0,4	0,002	19	0,011
Caja TATA & Variables estructurales	0,1	0,002	8	0,098
Caja TATA & Molecular Function	0,3	0,002	26	0,213
Caja TATA & Biological Process	0,5	0,002	15	0,131
Caja TATA & Cellular Component	0,3	0,002	12	0,260
Cho et al.- EDA (agrupamiento 1)	0,4	0,001	23	0,318
Cho et al. - EDA (agrupamiento 2)	0,4	0,001	6	0,115
Cho et al. - G&S SHAVING (agrupamiento 1)	0,6	0,002	45	0,006
Cho et al. - G&S SHAVING (agrupamiento 2)	0,6	0,002	36	0,003
Gasch et al. - EDA (agrupamiento 1)	0,4	0,001	17	0,005
Gasch et al. - EDA (agrupamiento 2)	0,4	0,001	21	0,004
Gasch et al. - G&S SHAVING (agrupamiento 1)	0,6	0,001	56	0,023
Gasch et al. - G&S SHAVING (agrupamiento 2)	0,7	0,001	35	0,019

Tabla 3.8: Reglas relacionando variables estructurales.

Sop.	Conf.	FC	Regla
0,12	0,40	0,15	$Long.gen = Bajo \rightarrow G+C = Alto$
0,12	0,38	0,14	$G+C = Bajo \rightarrow Long.gen = Alto$
0,12	0,41	0,16	$G+C = Alto \rightarrow Long.gen = Bajo$
0,12	0,40	0,14	$Long.gen = Alto \rightarrow G+C = Bajo$
0,13	0,41	0,17	$Long.interg. = Bajo \rightarrow GCinterg. = Bajo$
0,13	0,43	0,18	$GCinterg. = Bajo \rightarrow Long.interg. = Bajo$
0,13	0,44	0,21	$GCinterg. = Alto \rightarrow Long.interg. = Alto$
0,13	0,44	0,22	$Long.interg. = Alto \rightarrow GCinterg. = Alto$
0,18	0,63	0,24	$Long.interg.. = Alto \rightarrow Orient. = Div.$
0,23	0,56	0,15	$Long.interg.. = MED \rightarrow Orient. = Tan.$
0,20	0,40	0,16	$Orient. = Tan. \rightarrow GCinterg. = Bajo$
0,20	0,68	0,37	$GCinterg. = Bajo \rightarrow orient = Tan.$
0,19	0,36	0,10	$Orient. = Div. \rightarrow GCinterg. = Alto$
0,19	0,65	0,27	$GCinterg. = Alto \rightarrow Orient. = Div.$
0,13	0,42	0,17	$GCinterg. = Bajo \rightarrow G+C = Bajo$
0,13	0,41	0,17	$G+C = Bajo \rightarrow GCinterg. = Bajo$
0,14	0,46	0,23	$G+C = Alto \rightarrow GCinterg. = Alto$
0,14	0,46	0,23	$GCinterg. = Alto \rightarrow G+C = Alto$
0,038	0,48	0,12	$Crom. = 16 \rightarrow Long.interg.. = MED$
0,010	0,41	0,17	$Crom. = 3 \rightarrow G+C = Alto$
0,015	0,39	0,14	$Crom. = 9 \rightarrow GCinterg. = Alto$

tanto que la mayoría de ellas representan asociaciones biológicas reales, demostrando de esta forma la validez de la metodología. No era el objetivo de este trabajo proporcionar una interpretación biológica de todas las asociaciones obtenidas, sino mostrar que se pueden obtener asociaciones relevantes. De esta forma, trabajos futuros incluirán una interpretación biológica más profunda de los conjuntos de reglas obtenidos.

Las reglas que se muestran fueron seleccionadas de acuerdo con el conocimiento experto y la información extraída de la bibliografía. Asimismo se obtuvieron reglas con más de un item en el antecedente/consecuente que serán consideradas en trabajos futuros. Nótese que las tres ontologías GO se estudiaron por separado. Finalmente, para la extracción de reglas relacionando biclústers se estudiaron diferentes agrupamientos obtenidos con cada uno de los dos algoritmos, ya que modificando los parámetros de dichos algoritmos se pueden obtener agrupamientos ligeramente diferentes. Se seleccionaron los mejores agrupamientos de acuerdo con la medida GAP [114].

3.5.1 Variables estructurales

Las reglas obtenidas capturaron todas las relaciones previamente publicadas entre la longitud y la composición de los genes y los intergénicos (Tabla 3.8). De hecho, la primera descripción del genoma de la levadura llevada a cabo por Dujon (1996), ya reseñaba que los intergénicos de aquellos genes orientados de forma divergente, son más largos y con más contenido en G+C que aquellos entre genes orientados en tandem [82].

Las reglas de la Tabla 3.8 también muestran cierta correlación negativa entre la longitud y el contenido de G+C de los genes de la levadura (Spearman's $r = -0{,}25$, $p < 0{,}0001$) [168]. Se podría cuestionar la validez de dichas reglas, ya que los valores de confianza y factor de certeza que presentan son bajos: $\sim 0{,}40$ y $\sim 0{,}14$ respectivamente. Sin embargo, estos valores eran esperados, ya que la correlación de Spearman obtenida por Marin et al. es $-0{,}25$, indicando que el contenido en G+C y la longitud no son independientes y que existe cierta correlación negativa, justo lo que indican las reglas de la Tabla 3.8. De forma análoga, en esta misma tabla aparece la

correlación positiva existente entre la longitud y el contenido en G+C de los intergénicos, además de la relación existente entre el contenido en G+C de los genes y sus intergénicos [169].

El significado biológico de todas estas asociaciones no ha sido desvelado aún, si bien algunas de ellas son parcialmente comprendidas. Así, la mayor longitud de los intergénicos divergentes está ciertamente relacionada con la presencia de dos promotores, o promotores parcialmente compartidos entre los correspondientes genes. La relación positiva entre la longitud del intergénico y su contenido en G+C, se debe probablemente al efecto de la recombinación meiótica, que ocurre de forma predominante en los intergénicos divergentes, y que incrementa el contenido en G+C a través de la reparación sesgada de los pares GC [33, 90, 99]. De la misma forma, la correlación entre el contenido en G+C de los intergénicos y sus genes vecinos, podría deberse a mutaciones sesgadas de los pares GC durante la reparación del ADN, y a una ventaja selectiva a favor de una mayor apertura de la cromatina [90].

3.5.2 Cantidad de proteína y capacidad de reacción

Tal y como se puede observar en la Tabla 3.9, la abundancia de proteína en la célula aparece en correlación negativa con la longitud de los genes correspondientes. Este resultado era esperado, ya que diversos trabajos previos describieron una correlación negativa entre la longitud de los genes y los niveles de mRNA [67, 136, 168, 274]. Otra regla de la Tabla 3.9 relaciona proteínas abundantes con genes que presentan un alto contenido en G+C. Dicha asociación corrobora en este caso los resultados de Marin et al. [168], en los que se destacaba la correlación positiva entre el contenido en G+C de los genes y su nivel de transcripción. Asimismo, la capacidad de reacción aparece positivamente relacionada con el contenido en G+C de los genes, con su longitud y con el contenido en G+C de los intergénicos. Finalmente, hay que resaltar las asociaciones econtradas entre la presencia de la caja TATA, la longitud del gen y el contenido en G+C de las regiones intergénicas.

Todos estos resultados indican que, durante la evolución, las regiones de ADN codificadoras de proteínas han tendido a hacerse más cortas y a enriquecerse en G+C, con el objetivo de incrementar la concentración de mARN y

Tabla 3.9: Reglas que contienen las variables funcionales.

Sop.	Conf.	FC	Regla
0,092	0,48	0,12	$C.proteina = Alto \rightarrow Long.gen = Medio$
0,087	0,45	0,22	$C.proteina = Bajo \rightarrow Long.gen = Alto$
0,10	0,40	0,16	$C.reaccion = Alto \rightarrow G+C = Alto$
0,10	0,35	0,13	$G+C = Alto \rightarrow C.reaccion = Alto$
0,11	0,39	0,14	$C.reaccion = Bajo \rightarrow G+C = Bajo$
0,074	0,40	0,15	$C.proteina = Alto \rightarrow G+C = Alto$
0,096	0,37	0,12	$C.reaccion = Alto \rightarrow GCinterg. = Alto$
0,11	0,44	0,21	$C.reaccion = Alto \rightarrow Long.interg. = Alto$
0,11	0,38	0,17	$Long.interg. = Alto \rightarrow C.reaccion = Alto$
0,10	0,37	0,10	$C.reaccion = Bajo \rightarrow Long.interg. = Bajo$
0,055	0,41	0,17	$TATA = yes \rightarrow GCinterg. = Alto$
0,058	0,44	0,21	$TATA = yes \rightarrow Long.interg. = Alto$

la capacidad de reacción de los genes. Así, el acortamiento del mARN parece estar relacionado con la presión selectiva para reducir el tamaño de proteínas abundantes, de forma que se minimicen los costes de la transcripción y la traslación. Conforme la ARN-polimerasa se desplaza a lo largo de la cadena de ADN, produce cambios en la densidad super-helicoidal de la misma [187, 42, 194, 154], de manera que cuanto más corto es un gen, menor cambio se produce en la densidad super-helicoidal. Así, los cambios en el superenrollamiento de la doble cadena de ADN por delante de la ARN-polimerasa, serán mayores durante la transcripción de genes largos que durante la transcripción de genes cortos. Por lo tanto, la progresión de la ARN-polimerasa es menos eficiente durante la transcripción de genes largos.

La relación entre el contenido en G+C y la expresión de los genes es menos intuitiva. Dado que cuanto menor sea el contenido en G+C del gen, menor es la eficiencia de la transcripción, no se puede argumentar que la relación esté determinada por el mayor coste energético del proceso de aper-

tura de regiones de dsADN ricas en G+C. En lugar de esto, una estructura de cromatina diferente puede estar determinando la eficiencia de la transcripción. La importancia de la estructura de cromatina en la modulación de la transcripción se ha demostrado en estudios de alto rendimiento (*high-throughput*) sobre la transcripción del genoma de la levadura bajo supresión de la histona H4 [283, 156, 284]. En este sentido, se ha observado que el ADN podría presentar información estructural que determina su capacidad para interaccionar con la ADN-topoisomerasa I y los nucleosomas [54]. Asimismo, se ha demostrado que la posición de los nucleosomas viene determinada por diferentes segmentos de ADN de acuerdo con su contenido en G+C [273]. Además, se sabe también que un cambio estructural en secuencias G+C alternativas causa un bloqueo transcripcional y superenrollamiento negativo [193].

Estos resultados se ajustan al modelo de organización de la cromatina propuesto por Filipski y Mucha [90] (véanse también las referencias que aparecen en el artículo). Dicho modelo propone que regiones intergénicas (ricas en G+C) entre ORFs divergentes ocupan posiciones externas en la doble cadena de ADN, facilitando así una conformación abierta de la cromatina que, a su vez, facilita la recombinación y ofrece mayores posibilidades de regulación. El hecho de que ORFs orientadas divergentemente presentan un mayor potencial de regulación fue descubierto por Cho et al. [63], quienes analizaron las variaciones de nivel de mARN durante el ciclo celular (ver Seccion 3.2.4). Observaron que, de entre los genes regulados por el ciclo celular (ocupando posiciones adyacentes), aparece un exceso de genes divergentes (51 %) frente a aquellos orientados en tandem (38 %) o de forma convergente (11 %).

3.5.3 Términos GO

Se obtuvieron también reglas relacionando términos GO con el resto de variables (Tabla 3.10). Warringer y Blomberg [274] mostraron que las anotaciones GO dependen de la longitud de los genes. Por lo tanto, este hecho debería ser capturado por nuestra metodología. En efecto, se ha podido observar

Tabla 3.10: Reglas que contienen términos GO. Primera aproximación.

Sop.	Conf.	FC	Regla
0,0041	0,88	0,84	$GO = DNA\ helicase\ activity \rightarrow$ $\rightarrow Long.gen = Alto$
0,0017	1	1	$GO = cytochrome\text{-}c\ oxidase\ activity \rightarrow$ $\rightarrow Long.gen = Bajo$
0,023	0,57	0,39	$GO = plasma\ membrane \rightarrow$ $\rightarrow Long.gen = Alto$

que más del 60 % de las reglas extraídas con un item en el antecedente y el consecuente, y que relacionan nodos GO con el resto de variables estructurales, presentan "Longitud del gen" en el consecuente. Además, estos mismos autores encontraron ciertas funciones GO significativamente representadas en diferentes tamaños de proteínas; por ejemplo, el término "DNA helicase activity" aparece sobrerrepresentado en el conjunto de las proteínas más largas (más de 771 aminoácidos), mientras que el término "cytochrome-c oxidase activity" aparece sobrerrepresentado entre las proteínas más pequeñas (menos de 202 aminoácidos). Las reglas de la Tabla 3.10 muestran estas peculiaridades.

Sin embargo, tal y como se comentó en la Sección 3.3.2, se obtuvieron una gran cantidad de reglas relacionando términos GO, muchas de ellas redundantes. Por ello, se llevó a cabo el filtrado GO. La Tabla 3.11 muestra el número de reglas antes y después del filtrado GO, así como la tasa de reducción en el número de reglas para cada experimento. Tal y como se puede observar, la tasa media de reducción es del 38,8 %, siendo 68 % el máximo valor de la misma. Así, el tamaño de muchos de los conjuntos de reglas se redujo a casi la mitad de su tamaño original. Algunas de las reglas obtenidas tras el filtrado se muestran en la Tabla 3.12. Por ejemplo, dicha tabla muestra una regla relacionando el término "structural constituent of ribosome" con proteínas pequeñas, relación que da título al trabajo de Godfried et al. [37].

Tabla 3.11: Influencia del filtrado GO en el número de reglas.

Variables	Número de reglas antes	Número de reglas después	Tasa de reducción de reglas
Molecular Function & Variables estructurales	38	20	47%
Biological Process & Variables estructurales	11	7	36%
Cellular Component & Variables estructurales	24	12	50%
Cantidad proteína & Molecular Function	34	19	44%
Cantidad proteína & Biological Process	37	21	43%
Cantidad proteína & Cellular Component	23	14	39%
Capacidad reacción & Molecular Function	45	23	49%
Capacidad reacción & Biological Process	28	19	32%
Capacidad reacción & Cellular Component	50	19	62%
Caja TATA & Molecular Function	53	26	51%
Caja TATA & Biological Process	17	15	12%
Caja TATA & Cellular Component	37	12	68%
Cho et al. - EDA (agrupamiento 1)	24	23	4%
Cho et al. - EDA (agrupamiento 2)	6	6	0%
Cho et al. - G&S SHAVING (agrupamiento 1)	98	45	54%
Cho et al. - G&S SHAVING (agrupamiento 2)	79	36	54%
Gasch et al. - EDA (agrupamiento 1)	21	17	19%
Gasch et al. - EDA agrupamiento 2)	25	21	16%
Gasch et al. - G&S SHAVING (agrupamiento 1)	95	56	41%
Gasch et al. - G&S SHAVING (agrupamiento 2)	77	35	55%

Tabla 3.12: Reglas que contienen términos GO. Segunda aproximación.

Sop.	Conf.	FC	Regla
0,028	0,77	0,67	$GO = structural\ constituent\ of\ ribosome \rightarrow$ $\rightarrow Long.gen = Bajo$
0,01	0,78	0,69	$GO = helicase\ activity \rightarrow$ $\rightarrow Long.gen = Alto$

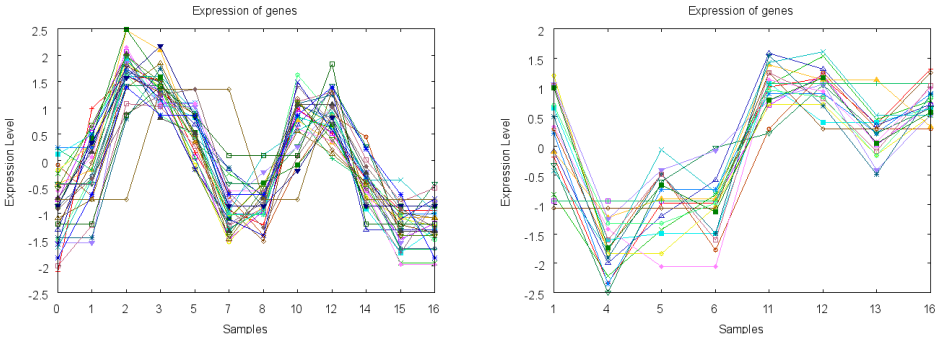
3.5.4 Datos de expresión

En esta Sección se muestran relaciones entre los patrones de expresión encontrados y el resto de características. El objetivo de este trabajo no consistía en proporcionar una interpretación biológica de todos los perfiles de expresión encontrados, sino mostrar cómo la metodología es capaz de revelar asociaciones interesantes entre dichos perfiles de expresión y el resto de características de forma intuitiva y gráfica. Siguiendo estas ideas se seleccionaron seis de los biclústers que presentaban un perfil claro y que aparecían en reglas interesantes y fiables. Las relaciones seleccionadas se muestran en la Tabla 3.13.

Los primeros cuatro biclústers de la Tabla 3.13 representan perfiles de expresión obtenidos de los experimentos del ciclo celular. Las reglas que aparecen en dicha tabla indican que el biclúster 1 está formado por genes cuyo producto se localiza en el núcleo y en orgánulos sin membrana (la definición de orgánulos sin membrana incluye los ribosomas, el citoesqueleto y los cromosomas). Este biclúster fue obtenido por el EDA biclustering algorithm y la Figura 3.16a muestra el perfil de expresión que representa. Tal y como se puede ver, el biclúster 1 contiene genes sobre-expresados al principio del ciclo celular y sub-expresados al final. Es clara la periodicidad de los niveles de expresión de estos genes a lo largo de los dos ciclos celulares. El siguiente biclúster que aparece referenciado en la Tabla 3.13 es el 2, el cual se obtuvo también por medio del EDA biclustering algorithm. Las ORFs asociadas a este biclúster presentan longitud media, una alta capacidad de reacción y llevan a cabo una función oxidorreductasa. El patrón de expresión representado por

Tabla 3.13: Reglas que contienen biclústers.

Sop.	Conf.	FC	Regla
0,0029	0,54	0,45	$bicluster = 1 \rightarrow GO =$ *non-membrane-bound organelle*
0,0033	0,61	0,45	$bicluster = 1 \rightarrow GO = nucleus$
0,0018	0,68	0,46	$bicluster = 2 \rightarrow Long.gen = Medio$
0,0022	0,80	0,74	$bicluster = 2 \rightarrow C.reaccion = Alto$
0,0012	0,43	0,40	$bicluster = 2 \rightarrow GO =$ *oxidoreductase activity*
0,0039	0,65	0,5	$bicluster = 3 \rightarrow GO = nucleus$
0,0029	0,48	0,44	$bicluster = 3 \rightarrow GO =$ *DNA metabolism*
0,0033	0,81	0,73	$bicluster = 4 \rightarrow Long.gen = LOW$
0,0036	0,89	0,85	$bicluster = 4 \rightarrow G + C = Alto$
0,0037	0,90	0,89	$bicluster = 4 \rightarrow GO =$ *non-membrane-bound organelle*
0,0037	0,90	0,89	$bicluster = 4 \rightarrow GO = biosynthesis$
0,0037	0,90	0,87	$bicluster = 4 \rightarrow GO =$ *protein complex*
0,0035	0,86	0,78	$bicluster = 4 \rightarrow GO =$ *organelle part*
0,0035	0,86	0,85	$bicluster = 4 \rightarrow GO = cytosol$
0,0035	0,86	0,85	$bicluster = 4 \rightarrow GO =$ *structural molecule activity*
0,0107	0,92	0,89	$bicluster = 5 \rightarrow Long.gen = Alto$
0,0073	0,63	0,41	$bicluster = 5 \rightarrow C.reaccion = Medio$
0,0019	0,71	0,69	$bicluster = 6 \rightarrow Crom. = II$
0,0017	0,64	0,61	$bicluster = 6 \rightarrow GO =$ *macromolecule biosynthesis*
0,0017	0,64	0,62	$bicluster = 6 \rightarrow GO =$ *cytosol*

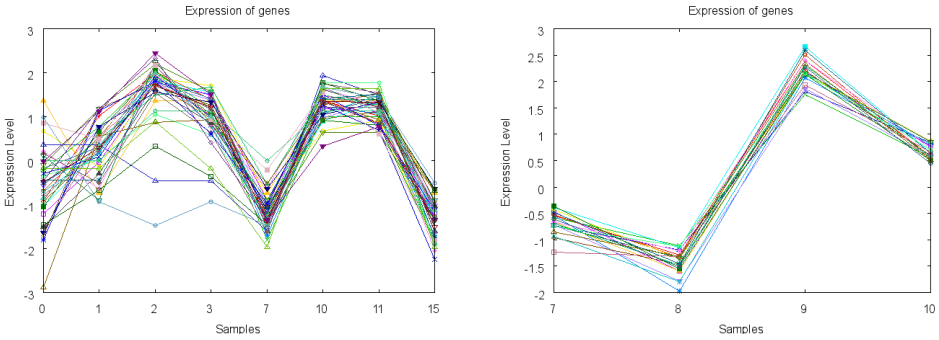


(a) Bicluster 1 **(b)** Bicluster 2

Figura 3.16: Patrones de expresión representados por los biclústers 1 y 2.

este grupo de genes se puede ver en la Figura 3.16b. Las siguientes dos reglas de la Tabla 3.13 hacen referencia al biclúster 3, el cuál fue también obtenido por el EDA biclustering algorithm. Las ORFs que aparecen en el biclúster 3 dan lugar a proteínas que llevan a cabo sus funciones en el núcleo y participan en el metabolismo de ADN. Tal y como se observa en la Figura 3.17a, se puede confirmar la correspondencia entre el proceso biológico "metabolismo de ADN" y el comportamiento de los genes que pertenecen a este grupo: estos genes aparecen sobre-expresados en la fase S del ciclo celular (muestras 2-3 y 10-12) en las que tiene lugar la replicación del ADN. Finalmente, se muestran algunas relaciones referentes al biclúster 4 (Figura 3.17b). Dicho biclúster fue obtenido en este caso mediante el Gene & Sample Shaving biclustering algorithm. Representa ORFs cuya expresión varia de forma brusca de sub-expresadas a sobre-expresadas cuando tiene lugar el cambio de ciclo celular (puntos de tiempo 7 a 10). Las reglas de la tabla 3.13 relacionan el biclúster 4 con ORFs cortas y con alto contenido en G+C, lo que además concuerda con los resultados de la Sección 3.5.1.

Las últimas 5 reglas contienen patrones de expresión obtenidos del conjunto de datos de Gasch et al. El biclúster 5 fue obtenido por el EDA biclustering algorithm, mientras que el biclúster 6 fue obtenido por el Gene & Sample Shaving algorithm (Figura 3.18). En este caso el conjunto de datos lo forman una amplia variedad de experimentos y, por lo tanto, los bi-



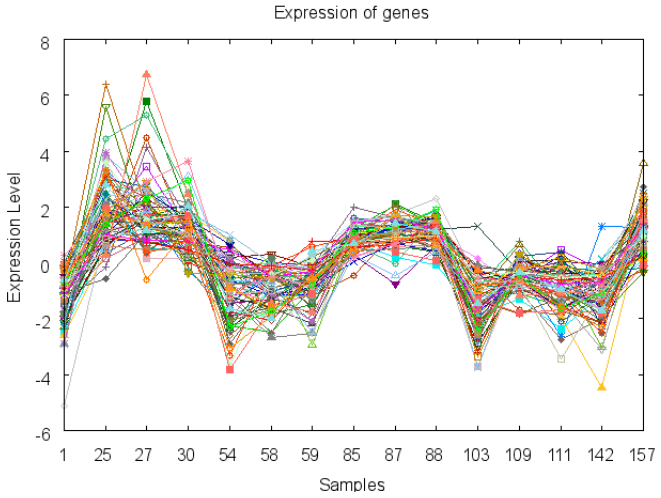
(a) Bicluster 3 **(b)** Bicluster 4

Figura 3.17: Patrones de expresión representados por los biclústers 3 y 4.

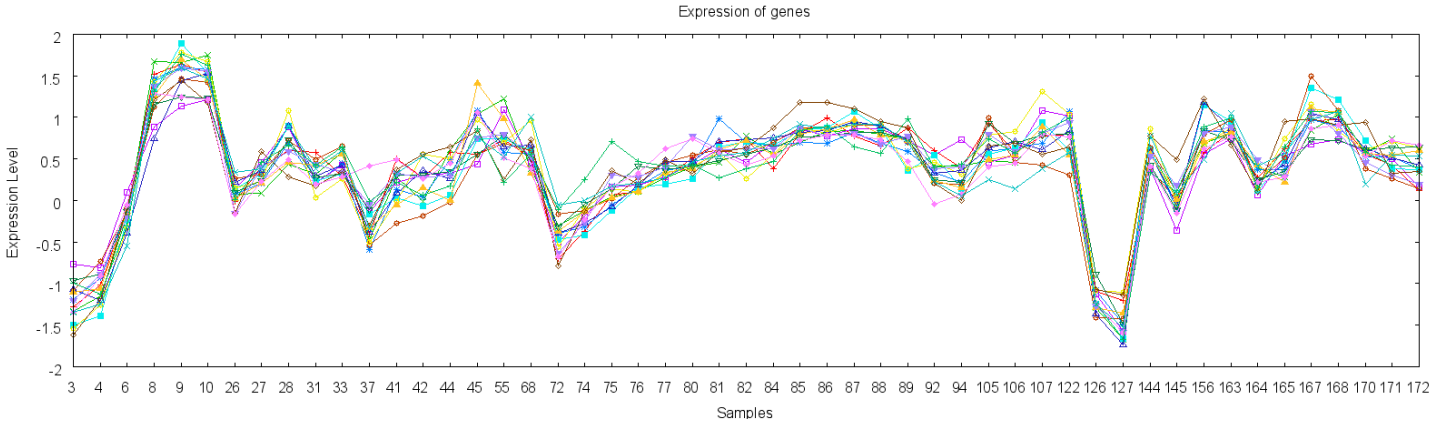
clústers obtenidos contienen columnas de experimentos muy diferentes. Por ejemplo, el biclúster 6 representa el perfil de expresión de 74 genes bajo 15 condiciones experimentales que pertenecen a 9 conjuntos diferentes de experimentos. Los genes que pertenecen a este biclúster son largos y tienden a presentar una capacidad de reacción media. Finalmente, las últimas tres reglas contienen el biclúster 6. Este biclúster resulta especialmente interesante ya que comprende un elevado número de variables (51) y presenta un perfil de expresión muy claro. Las asociaciones encontradas describen los genes de este grupo como pertenecientes al cromosoma II y anotados en los términos *macromolecule biosynthesis* y *cytosol*.

3.5.5 Comparación de los resultados crisp y difusos

Uno de los aspectos que se plantearon fue, si aplicando una metodología crisp, se obtendrían los mismos resultados que con la metodología difusa que se había aplicado previamente. En los experimentos realizados, se observaron diferencias entre los resultados difusos obtenidos y sus equivalentes crisp. Para la extracción de las reglas crisp, los dominios continuos se dividieron en tres intervalos utilizando para ello los percentiles p_{33} y p_{66}. Así, se obtuvieron dos conjuntos de reglas adicionales: uno por el método difuso y otro por el método crisp. Se utilizaron los mismo umbrales para ambos métodos: 0,004 para el soporte y 0,5 para la confianza y el FC. Así, se obtuvieron



(a) Bicluster 5



(b) Bicluster 6

Figura 3.18: Patrones de expresión representados por los biclústers 5 y 6.

Tabla 3.14: Resultado de los ANOVAs para comparar los resultados crisp y difusos.

Medida de calidad	$p-value$	**Media-Crisp**	**Media-Difusa**
Soporte	$1,80E-018$	0,0080	0,0073
Confianza	$1,13E-082$	0,777	0,757
Factor de certeza	$1,47E-049$	0,622	0,606

22893 reglas con el algoritmo difuso mientras que 27304 fueron generadas por el algoritmo crisp.

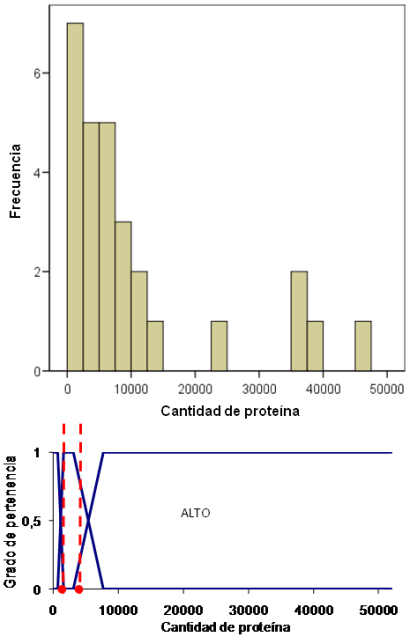
En primer lugar, se compararon los valores de soporte, confianza y FC de las reglas presentes en ambos conjuntos, encontrándose que 11655 de las reglas eran comunes a ambos métodos. Para determinar si los valores de las medidas obtenidas presentaban alguna diferencia significativa, se llevaron a cabo tres ANOVAs (Tabla 3.14). Como se puede observar, aparecen diferencias significativas para las tres medidas. Estas variaciones eran esperadas, dada la forma en que la metodología difusa modela los límites entre etiquetas adyacentes. Las medias de las medidas crisp son mayores que las de sus correspondientes difusas, lo que indica que las medidas crisp tienden a ser mayores que las difusas. Dado que se ha demostrado que las técnicas difusas modelan de una forma más adecuada los conceptos lingüísticos y mejoran el tratamiento de los datos imprecisos, parece ser que el algoritmo crisp tiende a proporcionar valores de calidad mayores de lo que realmente son, mostrando así la necesidad de usar técnicas difusas.

En la Tabla 3.15 se muestran algunos ejemplos concretos de reglas que ilustran la necesidad de utilizar metodologías difusas. Las reglas que aparecen en dicha tabla presentan valores de calidad que varían significativamente entre la versión crisp y difusa. Por ejemplo, los valores difusos de soporte, confianza y FC de la primera regla de la Tabla 3.15 son considerablemente menores que los equivalentes crisp. Este hecho se puede comprender analizando la Figura 3.19a. Dicha figura muestra cómo los genes anotados en el término *electron transport*, se distribuyen a lo largo del dominio de la va-

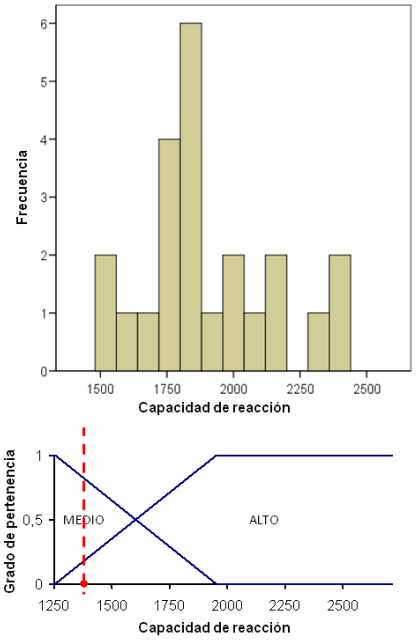
riable *cantidad de proteína*. Muestra asimismo, cómo se definen las etiquetas lingüísticas en el algoritmo difuso y crisp. Observando el histograma se puede ver que aparecen muchos genes en el borde que separa las etiquetas *Medio* y *Alto*. La mayoría de estos genes son considerados como *Largos* por el algoritmo crisp, mientras que son "un poco" *Medios* y "un poco" *Largos* en el difuso. Esto lleva consigo que el algoritmo difuso cuente menos genes *Largos* anotados en *electron transport* y que, por lo tanto, obtenga valores más bajos de soporte, confianza y FC. El mismo razonamiento es válido para las dos reglas siguientes de la Tabla 3.15. Sus correspondientes gráficos se muestran en las Figuras 3.19b y 3.20a. En el caso de la última regla de la Tabla 3.15, los valores de soporte, confianza y FC son ligeramente mayores en la versión difusa que en la versión crisp. Observando la Figura 3.20b, se puede apreciar que muchos de los genes localizados en el cromosoma 16, presentan valores de longitud del intergénico en el límite que divide los conjuntos *Medio-Bajo* y *Medio-Alto*. En este caso, el conjunto difuso *Medio* no sólo incluye "completamente" (es decir, grado de pertenencia 1) casi todos los genes que están incluidos en el conjunto crisp *Medio*, sino que además aparecen muchos genes que pertenecen a él con un menor grado de pertenencia y que no están incluidos en el conjunto crisp. Esto provoca un incremento en el valor del soporte difuso, ya que el número de genes localizados en el cromosoma 16 es el mismo en el algoritmo difuso y en el crisp. Dado que incrementa el número de genes del cromosoma 16 que tienen longitud de intergénico *Medio*, los valores de confianza y FC también aumentan.

Tabla 3.15: Algunas de las reglas obtenidas con el algoritmo crisp y difuso.

Sop. Crisp.	**Sop Dif**	**Conf. Crisp.**	**Conf Dif**	**FC Crisp**	**FC Dif**	**Regla de asociación**
0,0039	0,0030	0,70	0,53	0,60	0,43	*electron transport* $\rightarrow$ $\rightarrow C.proteina = Alto$
0,0044	0,0036	1	0,81	1	0,75	*snoRNA binding* $\rightarrow$ $\rightarrow C.reaccion = Alto$
0,0055	0,0044	0,71	0,56	0,58	0,41	$bicluster = 5 \rightarrow$ $\rightarrow C.reaccion = Alto$
0,0032	0,038	0,39	0,48	0,09	0,12	$Crom. = 16 \rightarrow$ $Long.interg. = Medio$



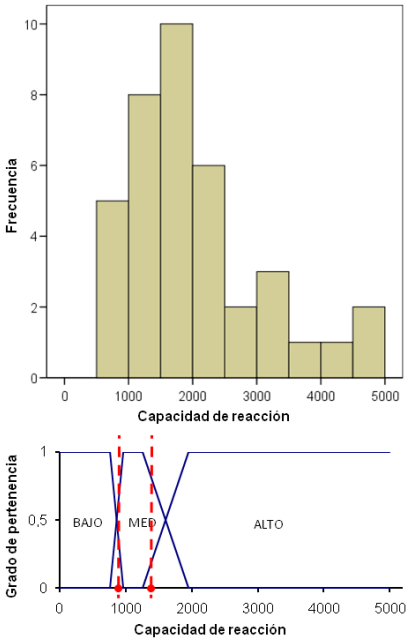
(a) Distribución de los genes anotados en el término *electron transport* a lo largo del dominio de la variable *cantidad de proteína*. Las líneas discontinuas rojas indican los percentiles p_{33} y p_{66}.



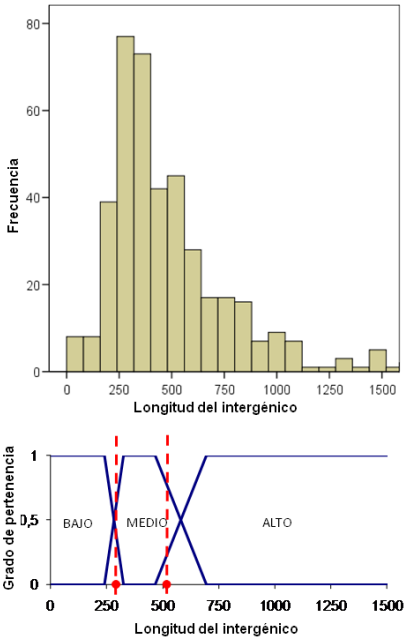
(b) Distribución de los genes anotados en el término *snoRNA binding* a lo largo del dominio de la variable *cantidad de proteína*. La línea discontinua roja indica el percentile p_{66}.

Figura 3.19: Comparación entre los resultados crisp y difuso.

.



(a) El histograma muestra la distribución de los genes del biclúster 5 en el dominio de la variable *capacidad de reacción*. Las líneas discontinuas rojas indican los percentiles p_{33} y p_{66}.



(b) El histograma muestra la distribución de los genes del cromosoma 16 en el dominio de la variable *longitud del intergénico*. Las líneas discontinuas rojas indican los percentiles p_{33} y p_{66}.

Figura 3.20: Comparación entre los resultados crisp y difuso.

3.6 Conclusiones

En este trabajo se ha propuesto una metodología para la integración y el análisis de información biológica heterogénea. El principal aspecto de esta metodología consiste en el uso de una versión difusa del algoritmo de extracción de reglas de asociación TD-FP-Growth. Se ha construido y utilizado un conjunto de datos basado en el genoma de la levadura para validar la metodología propuesta.

Los resultados revelan interesantes asociaciones entre diversas características estructurales y funcionales del genoma de la levadura, muchas de ellas en concordancia con trabajos previos en este área. De esta forma, se demuestra que la metodología es útil para obtener asociaciones interesantes y fiables, las cuales podrían servir como hipótesis para una posterior validación en el laboratorio. Asimismo, las reglas de asociación difusas han mostrado ser una herramienta intuitiva para describir relaciones biológicas utilizando etiquetas lingüísticas y unos pocos parámetros fácilmente interpretables (soporte, confianza y factor de certeza).

Los resultados evidencian la importancia de utilizar técnicas que modelen los límites de una manera más realista a como lo hacen las técnicas clásicas crisp. La definición apropiada de los conceptos introducidos en el análisis es crucial, ya que determina la interpretación que se hará de los resultados. Asimismo, la presencia de ruido e imprecisión en los datos biológicos hace más necesario aún utilizar definiciones difusas de estos conceptos y aplicar sobre ellos técnicas de *Soft Computing*.

CAPÍTULO 4

Relaciones entre perfiles de expresión y factores de pronóstico en cáncer de mama

4.1 Introducción

La experimentación llevada a cabo en el Capítulo anterior proporcionó gran cantidad de información acerca del rendimiento de la metodología, tanto en términos del tamaño del conjunto de datos que es capaz de manejar, como en valores que se pueden esperar de las medidas de calidad, o el tipo de reglas que se desean obtener (por ejemplo, el tamaño máximo de las reglas que podrían ser de interés). En este capítulo se presenta una aplicación de la metodología desarrollada en el capítulo anterior al estudio del cáncer de mama. Los resultados muestran una gran cantidad de asociaciones interesantes, muchas de ellas en concordancia con bibliografía previa y otras que podrían dar lugar a nuevos marcadores en el cáncer de mama.

Motivación

El cáncer de mama es un problema de salud pública de gran importancia. Se trata del segundo cáncer más común en el mundo (una de cada ocho

mujeres lo padece), y de la quinta causa más común de muerte por cáncer. Su alta incidencia y mortalidad lo han convertido en el centro de atención de gran cantidad de investigaciones. Sin embargo, no se sabe aún por qué algunas personas (principalmente mujeres) desarrollan el cáncer de mama. La etiología del cáncer de mama es aún hoy desconocida, aunque se sabe que las hormonas, factores genéticos y condiciones ambientales tienen un papel principal en su desarrollo [133]. De hecho, las nuevas tecnologías están permitiendo entender e identificar algunos de los factores de pronóstico en el cáncer de mama. Es más, hasta ahora no ha existido nunca tanta información disponible que permita estudiar esta enfermedad. Sin embargo, transformar esta ingente cantidad de datos biomédicos en información útil no es una tarea trivial en absoluto.

Tal y como se comentó previamente, existen una gran variedad de factores de pronóstico asociados al cáncer de mama (estadio del tumor, estado del HER2, estado de los receptores hormonales, etc.). Además, se han descrito distintos subtipos de tumores basándose en su perfil de expresión [197, 206, 233, 234, 235]. No existe sin embargo, al menos hasta donde nosotros conocemos, ningún otro trabajo que estudie las relaciones entre la expresión del genoma completo y los factores de pronóstico en el cáncer de mama. Es más, en el reciente trabajo de Sims [227], se insta a la comunidad científica a investigar los vínculos existentes entre la expresión de los genes y el estado de los factores de pronóstico conocidos. Los avances en este sentido, podrían mejorar los tratamientos que se aplican a pacientes cuyo tumor presenta un pronóstico poco claro [227].

Propuesta

En este contexto, se propone aplicar la metodología desarrollada en el capítulo anterior, para desvelar relaciones entre patrones de expresión y factores de pronóstico en el cáncer de mama. Así, se ha creado una base de datos que integra información de los principales factores de pronóstico en cáncer de mama y datos de expresión del genoma completo. En este nuevo estudio se dispone de un conjunto de datos de características similares al del capítulo anterior: grandes dimensiones, datos heterogéneos, imprecisión e

incertidumbre en los mismos. Aunque presenta algunas particularidades destacables que se comentarán posteriormente, las reglas de asociación difusas son, por tanto, una herramienta adecuada para llevar a cabo el estudio.

4.2 El cáncer

En un organismo pluricelular cada célula actúa de manera socialmente responsable, dividiéndose, manteniéndose quiescente, diferenciándose o muriendo cuánto sea necesario para el bien del organismo. En el cuerpo humano, formado por más de 10^{14} células, debido al propio proceso de replicación celular, miles de millones de estas células sufren mutaciones cada día. Lo más peligroso, es que una mutación de este tipo puede proporcionar a una célula una ventaja selectiva, permitiendo que se divida más vigorosamente que sus vecinas y convirtiéndose en un clon mutante en crecimiento. Cualquier mutación que origine un comportamiento no altruista arriesgará el futuro de un organismo. Ciclos repetidos de mutaciones, competición y selección natural que actúan dentro de una población de células somáticas provocarán problemas que empeorarán con el tiempo. Éstos son los ingredientes básicos del cáncer. Así, se puede establecer una lista de comportamientos generales clave en las células cancerosas:

- *Ignorar las señales internas y externas que regulan la proliferación celular*.
- *Evitar la muerte programada o apoptosis*. El número de células se encuentra regulado no sólo por los controles de proliferación celular, sino también por el ritmo de la muerte celular. Aquellas células que ya no son necesarias, se autoeliminan activando un programa intracelular de muerte, llamado *muerte celular programada* o *apoptosis*. Si se generan demasiadas células, la tasa de apoptosis aumenta equilibrando esta superproducción. Una de las propiedades más importantes de las células cancerosas es que han perdido esta capacidad de “suicidarse”.
- *Sortear las limitaciones programadas para la proliferación, eludiendo la senescencia replicativa y evitando la diferenciación*. En promedio, cada

división normal de una célula madre genera una célula madre hija y una célula que está condenada a la diferenciación terminal y a la suspensión de la división celular. Por tanto, para que una célula madre transformada genere un clon de progenie en crecimiento constante, es necesario que las pautas básicas estén alteradas: o bien las células madre fracasan en la producción de una célula que no sea célula madre en cada división, o bien el proceso de diferenciación está descontrolado de manera que las células hijas retienen cierta capacidad de dividirse indefinidamente. Además, la mayoría de las células humanas sólo realizan un número limitado de divisiones celulares, proceso conocido como *senescencia replicativa*. Muchas células cancerosas eluden este tipo de autolimitación de la proliferación celular.

- *Ser genéticamente inestables*. La inmensa mayoría de los cánceres humanos presentan tasas de mutación sumamente elevadas, por lo que se dice que son *genéticamente inestables*. Aunque la inestabilidad genética afecta a la salud celular, podría proporcionarle a la célula propiedades adicionales o cambios que induzcan mutaciones adicionales que le confieran algún tipo de ventaja competitiva. Parece que existe este nivel "óptimo" de inestabilidad genética para el desarrollo del cáncer, transformando una célula lo suficiente como para que evolucione peligrosamente, pero no tanto como para que muera.
- *Tener capacidad invasiva o capacidad de generar metástasis*, lo que implica, generalmente, la habilidad de liberarse y entrar en el torrente sanguíneo o en los vasos linfáticos.
- *Sobrevivir y proliferar en entornos ajenos*, pudiendo así formar tumores secundarios o *metástasis* en otros lugares del cuerpo.

4.2.1 El cáncer de mama

Esta enfermedad puede definirse como la proliferación acelerada, desordenada y no controlada de células pertenecientes a distintos tejidos de una glándula mamaria. La etiología del cáncer de mama aún hoy es desconocida, no obstante se sabe que las hormonas, factores genéticos y ambientales,

desempeñan un importante papel en su aparición y desarrollo [133]. A medida que se van conociendo mejor los factores pronósticos de la enfermedad, se puede ir identificando el riesgo particular de cada caso de cáncer de mama, lo que está ayudando a seleccionar el tratamiento más adecuado para cada paciente. De hecho, gracias a los avances de la Biología Molecular, se están identificando cuáles son los factores pronósticos de la enfermedad en un paciente concreto, y por eso cada vez se tiende más a realizar tratamientos individualizados.

El cáncer de mama metastático está asociado con factores de riesgo, tales como nódulos axilares linfoides[1] afectados, ausencia de receptores de estrógenos y progesterona, incremento de la fase S del ciclo celular, grado nuclear elevado[2], estadio avanzado y sobre-expresión del oncogen[3] HER2 [229]. Esta sobre-expresión se ha observado en el 25 % − 30 % de los tumores de mama y se correlaciona con la ausencia de receptores estrogénicos y de progesterona [230, 294]. Además, el procedimiento rutinario de diagnóstico del cáncer de mama incluye la determinación del estado de otros biomarcadores tales como el ki67 ó el p53.

Etapas del tumor

Para poder determinar el tratamiento más adecuado del cáncer de mama, es importante conocer en qué fase se encuentra el tumor. El sistema que con mayor frecuencia se emplea para su clasificación es el TNM. Estas siglas hacen referencia a tres aspectos del cáncer: la T se refiere al tamaño del tumor, la N a la infiltración de los ganglios linfáticos y la M a la afectación metastásica o no de otros órganos. En función de estos aspectos, el cáncer de mama se agrupa en las siguientes etapas o estadios:

[1] Nódulos o ganglios linfáticos: estructuras en forma de alubia que ayudan a filtrar el exceso de fluidos, bacterias, y los subproductos de las infecciones. La mayoría de los nódulos linfáticos están agrupados en áreas específicas del cuerpo, como la boca, la nuca, el antebrazo, la axila y la ingle.

[2] Grado nuclear: evaluación del tamaño y la forma del núcleo de las células tumorales y del porcentaje de las células tumorales que están en proceso de multiplicarse o crecer. Los cánceres de grado nuclear bajo crecen y se diseminan más lentamente que los cánceres de grado nuclear alto.

[3] Oncogén: versión mutada/desregulada de cualquier gen normal que favorece la proliferación, inhibe la apoptosis, o potencia la invasividad o capacidad metastásica.

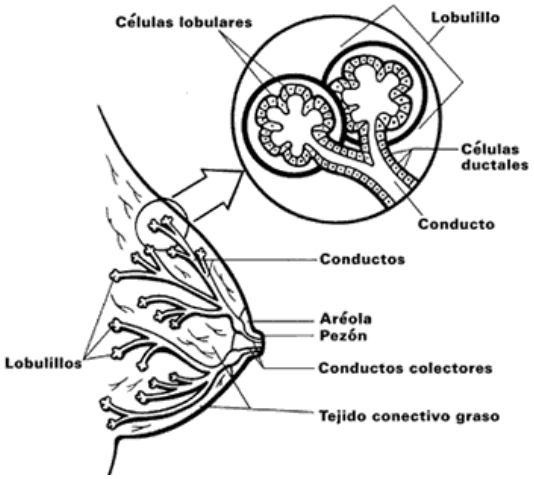


Figura 4.1: Esquema de una mama.

- Estadio 0. Son lesiones precursoras, incluidas el carcinoma[4] *in situ*. Las células tumorales están localizadas exclusivamente en la pared de los lobulillos[5] o de los conductos[6] (Figura 4.1).
- Estadio 1. El tamaño del tumor es inferior a dos centímetros. No hay afectación de ganglios linfáticos ni metástasis a distancia.
- Estadio 2. Tumor entre 2 y 5 centímetros, con o sin afectación de ganglios axilares.
- Estadio 3. El tumor afecta a ganglios axilares y/o piel y pared torácica (músculos o costillas).
- Estadio 4. El cáncer se ha diseminado, afectando a otros órganos como hueso o hígado.

HER2(receptor-2 del factor de crecimiento epidérmico humano)

El oncogén HER2, también conocido como c-erbB2, se encuentra localizado en el cromosoma 17q21.1[7] y codifica una proteína que desempeña un papel

[4]Carcinoma: cáncer que deriva de células epiteliales.
[5]Lobulillos: glándulas productoras de leche.
[6]Conductos: pequeños tubos que llevan la leche de los lobulillos al pezón
[7]El 17 indica el cromosoma. El resto de valores indican la posición en el cromosoma. El centrómero divide al cromosoma en dos partes: el brazo corto (p) y el brazo largo (q). Los

importante en el crecimiento y el desarrollo de las células epiteliales, como es el caso de las glándulas mamarias. En cada célula normal hay dos copias del gen HER2 (una por cada cromosoma 17) y aproximadamente 50.000 copias de la proteína en la superficie celular, mientras que en una célula tumoral se pueden encontrar más de dos copias del gen y más de 1 millón de copias de la proteína en la superficie celular, favoreciendo la formación de heterodímeros de HER2 y convirtiendo a la célula en extremadamente sensible a factores de crecimiento [229, 228].

La sobre-expresión de HER2 se ha observado en el $25\,\% - 30\,\%$ de los tumores de mama. En la mayoría de los casos esta sobre-expresión de la proteína refleja amplificación[8] del gen HER2, aunque en el $3\,\% - 7\,\%$ de los casos la sobre-expresión puede ocurrir en ausencia de amplificación génica, posiblemente debido, entre otras causas, a una polisomía[9] del cromosoma 17 [230, 270, 288]. Así, la importancia de la determinación del estado de HER2 radica en que aquellos pacientes que tienen cáncer de mama con amplificación de HER2 presentan menor supervivencia [229, 228, 204, 223], una velocidad de crecimiento tumoral elevada, un riesgo mayor de recurrencia tras la cirugía y una respuesta pobre al tratamiento quimioterapéutico convencional [210].

Desde hace unos años se dispone de un anticuerpo monoclonal[10] humanizado llamado trastuzumab (Herceptin®), que es capaz de bloquear de manera especifica la acción del oncogen HER2 [53]. Este hecho interrumpe las señales de transducción[11], inhibiendo principalmente la proliferación de las

números tras las letras representan la posición sobre el brazo: banda 2, sub-banda 1, y, tras el punto, sub-sub-banda 1. Las bandas son visibles bajo el microscopio cuando el cromosoma está adecuadamente teñido. Cada banda se numera empezando por 1 por la más cercana al centrómero. Sub-bandas y sub-sub-bandas son visibles a altas resoluciones.

[8] Se dice que hay amplificación de un gen cuando aparecen más de una copia de dicho gen por cromosoma.

[9] La polisomía se define como la ocurrencia de tres o más copias de un cromosoma, en este caso, el cromosoma 17.

[10] Los anticuerpos monoclonales son proteínas que reconocen específicamente y se unen a otras proteínas únicas del organismo llamadas antígenos. Trastuzumab se une selectivamente a la proteína del HER2.

[11] La transducción de señal es el conjunto de procesos o etapas que ocurren de forma concatenada por el que una célula convierte una determinada señal o estímulo exterior, en otra señal o respuesta específica.

células que sobre-expresan HER2. En los tumores de mama HER2 positivos sería lógico esperar una buena respuesta al trastuzumab, sin embargo sólo entre un 20 % y un 30 % de estos pacientes responden al tratamiento satisfactoriamente [182]. Este hecho pudiera ser debido a la aparición de diversos mecanismos de resistencia, tales como la pérdida de expresión de HER2 en la superficie celular, activación de vías alternativas u otros mecanismos moleculares aún hoy poco conocidos [182]. No obstante, la administración de trastuzumab en el cáncer de mama metastático reduce el riesgo de muerte en un 30 % [231].

Por todo ello, la detección del estado de HER2 forma parte del diagnóstico anatomopatológico rutinario del cáncer de mama [288]. No obstante, aún hoy no existe un consenso sobre cual es el método más adecuado para determinar el estado del HER2.

Detección del estado de HER2

Existen varios procedimientos analíticos para abordar el estudio de HER2 a diferentes niveles (ADN, ARNm y proteína) aunque los más utilizados son las técnicas inmunohistoquímicas (IHC) y la hibridación in situ fluorescente (FISH). Existe un gran debate sobre cual es el mejor procedimiento para testar HER2, debido a diferencias en la fiabilidad y coste de las técnicas de IHC y FISH y a la ausencia de una referencia estándar.

La IHC mide el nivel de expresión de la proteína HER2 en tanto que FISH demuestra el grado de amplificación del gen. Generalmente, hay una relación entre sobre-expresión y amplificación aunque existen casos excepcionales que no cumplen esta regla. Se considera preferible FISH si bien, como consecuencia de la concordancia entre los resultados obtenidos de IHC y FISH y por razones prácticas, se tiende a hacer un cribado previo con IHC y sólo en los casos dudosos se lleva a cabo FISH [25, 32, 80]. No obstante, varios estudios sugieren que esta estrategia no es coste-efectiva debido a que el trastuzumab sólo es beneficioso en pacientes FISH positivos (es decir, con amplificación de HER2) [171, 222]. A continuación se incluyen unos breves comentarios sobre ambos procedimientos.

Tabla 4.1: Interpretación de los resultados de IHC.

Escala	Resultado	Visualización
0	Negativo	Tinción de cualquier intensidad en menos del 10 % de las células
1	Negativo	Tinción incompleta de la membrana en más de un 10 % de las células
2	Positivo	Tinción completa, débil o moderada de la membrana en más del 10 % de las células
3	Positivo	Tinción completa e intensa de la membrana en más de un 10 % de las células

Inmunohistoquímica (IHC)

Se denomina así a un grupo de técnicas de inmunotinción que permiten mostrar una variedad de antígenos presentes en las células o tejidos, utilizando anticuerpos específicos y sistemas de visualización no fluorescentes. Estas técnicas se basan en la capacidad de los anticuerpos de unirse específicamente a los correspondientes antígenos. Así, mediante la inmunohistoquímica de HER2, se identifican pacientes con cáncer de mama, que presentan una sobre-expresión de la proteína HER2. La visualización se consigue con reactivos unidos a un cromógeno que permite realizar posteriormente la interpretación con un microscopio óptico. Los resultados del test se presentan como 0, +1, +2 ó +3, dependiendo de la intensidad de la tinción, la proporción de membrana teñida y la proporción de células teñidas (Tabla 4.1). Los tumores con un grado de +3, según el test IHC, se consideran HER2 positivos. Los tumores de grado +2 se consideran en el límite, y los 0 y +1 negativos.

Hibridación in situ fluorescente (FISH)

Técnica que utiliza moléculas fluorescentes, fluorocromos, para localizar genes o fragmentos de ADN. Consiste en preparar cortas secuencias de ADN de una sola hebra, llamadas sondas, que son complementarias de las secuencias de ADN que se quieren estudiar. Estas sondas se marcan con fluorocromos y posteriormente se hibridan o unen al ADN complementario, permitiendo localizar las secuencias en las que se encuentran. Se aplica sobre cortes de tejido o extensiones citológicas y se observa la fluorescencia al microscopio de epifluorescencia con filtros específicos.

El método FISH detecta la amplificación del gen y no el nivel de expresión de la proteína HER2, como sucede en IHC. La anotación se realiza calculando el cociente entre el número de señales de HER2 por célula, por lo que es un procedimiento semicuantitativo. Valores mucho mayores que dos se consideran indicativos de amplificación de HER2, mientras que se recomienda repetir el ensayo cuando se obtienen cocientes de entre 1,8 y 2,2 [120].

Se ha observado que pacientes HER2 positivos (IHC+3) presentan una supervivencia menor que los 0, +1 y +2, no existiendo diferencias significativas entre los otros grupos [192]. Sin embargo, entre las mujeres IHC+2 e IHC+3 el pronóstico es peor en las FISH positivas que en las negativas. Tampoco se han encontrado diferencias en la supervivencia global al realizar FISH en pacientes IHC0 e IHC1 y clasificarlos como positivos o negativos. Estos datos muestran que los resultados obtenidos por FISH son un factor pronóstico a considerar.

Biomarcadores adicionales

Tal y como se comentó anteriormente, existen otros biomarcadores que forman parte del diagnóstico clínico rutinario del cáncer de mama. Por ejemplo, la detección del estado de los receptores hormonales (RE y RP) es fundamental, ya que aquellos tumores con ausencia de estos receptores presentan un mal pronóstico, dado que estos pacientes no pueden recibir la terapia hormonal [85]. Por tanto, dichos pacientes tienen que recibir un tratamiento tradicional de quimioterapia [105]. Otro marcador importante es el p53. La presencia de alteraciones en este gen se asocia también a un mal pronóstico.

Esto se debe a que una de las funciones principales del p53 es evitar que las células con daño en su ADN entren de nuevo en el ciclo celular. Si el gen p53 es alterado y pierde su función, las células con ADN dañado continúan reproduciéndose. Finalmente, el antígeno nuclear ki67 es un marcador de proliferación expresado por las células en todas las fases proliferativas del ciclo celular (G1, S, G2, M) [96]. Los anticuerpos ki67 son útiles para establecer la fracción de células en crecimiento del tumor. Los tumores que presentan una rápida proliferación están asociados también a un mal pronóstico [43].

El estado de todos estos marcadores se determina normalmente mediante técnicas inmunohistoquímicas. La expresión del p53 y del ki67 se evalúa de acuerdo a la proporción de núcleos teñidos en el tumor. El estado del p53 se considera negativo si esta proporción es menor que el 5 %. El umbral para el marcador ki67 se sitúa alrededor del 10 %. En el caso de los receptores hormonales, el estado se determina mediante el índice de Allred [14, 15]. Este índice evalúa la proporción de célula teñida y la intensidad de la tinción. Se calcula de acuerdo a la suma de la proporción de célula teñidas (PS, proportion score) y su intensidad (IS, Intensity score). PS oscila entre 0 (ninguna célula teñida) y 5 (entre 2/3 y el 100 % de células teñidas). IS oscila entre 0 (tinción negativa) y 3 (tinción fuerte). Valores de este índice por encima de 3 se consideran positivos.

4.3 Construcción de la tabla de datos

Los datos utilizados en este capítulo fueron proporcionados por uno de los centros españoles para la detección del estado del HER2, el Hospital Universitario Virgen de las Nieves de Granada. Se registraron datos de expresión, estadio del tumor y estado de marcadores histológicos e inmunohistoquímicos de 58 muestras de tejido. Por otra parte, se obtuvo información clínica de otros 2751 pacientes con cáncer de mama diagnosticados entre Septiembre de 2001 y Diciembre de 2007. De esta forma, se llevaron a cabo dos análisis por separado: en el primero, se realizó un análisis exploratorio de los datos clínicos obtenidos de los 2751 pacientes. En el segundo, se llevó a cabo el análisis principal de este trabajo, el cual comprende el estudio de las relacio-

nes entre datos de expresión y factores de pronóstico en las 58 muestras de tejido mencionadas anteriormente.

4.3.1 Estado del HER2

El estado del gen HER2 se evaluó mediante los análisis IHC (análisis inmunohistoquímico) y FISH (*fluorescence in situ hybridization*) de las muestras. Tal y como se explicó anteriormente, el IHC viene dado en forma de escala cualitativa de 0 a 3 que depende de la tinción de las células. La tinción del citoplasma se consideró no específica. La aparición de un patrón heterogéneo de tinción en diferentes áreas del mismo tumor se consideró asimismo como una variable adicional en el estudio.

Los resultados obtenidos mediante FISH proporcionan el cociente de señales de HER2 entre el número de señales de centrómeros de cromosoma 17 (CEP17). Aquellos tumores con una proporción de Her2:CEP17 $\geq 2:2$ se consideraron positivos para la prueba de amplificación del gen. Asimismo, el test FISH permitió determinar el número de copias del cromosoma 17, por lo que se incluyó una variable más en el estudio indicando la aparición de polisomía para el cromosoma 17.

4.3.2 Otros datos inmunohistoquímicos

El estado de los receptores de estrógenos y progesterona (RE, y RP respectivamente) se determinó mediante el índice de Allred [14, 15]. Valores de este índice mayores o iguales a 3 se consideraron positivos.

La expresión de los marcadores p53 y ki67 se evaluó de acuerdo con la proporción de núcleos teñidos en células tumorales. Respecto al p53, valores superiores al 5 % se consideraron positivos. En cuanto al ki67, se definieron dos conjuntos difusos tal y como muestra la Figura 4.2, ya que los porcentajes alrededor del 10 % se encuentran en el límite entre los casos positivos y negativos.

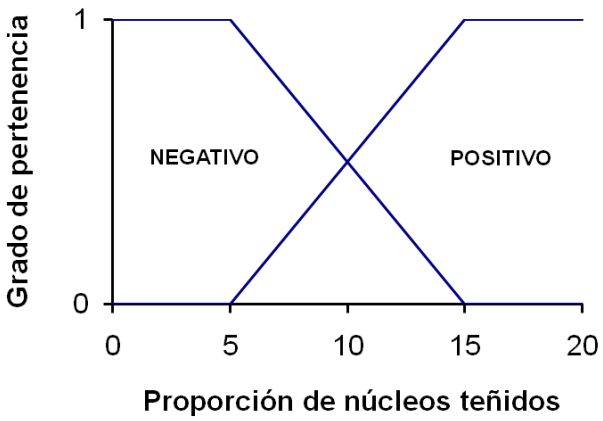


Figura 4.2: Definición de dos conjuntos difusos para el marcador ki67

4.3.3 Datos de microarrays

Se extrajo ARN de 54 tejidos tumorales y 12 tejidos de mama normal utilizando Trizol (Invitrogen, Carlsbad, CA). Tras el control de calidad utilizando Bioanalyzer (Agilent 2100), los ARNs se amplificaron, etiquetaron y finalmente se hibridaron en microarrays de genoma completo GeneChip U133 plus 2.0 de acuerdo con el protocolo marcado por Affymetrix.

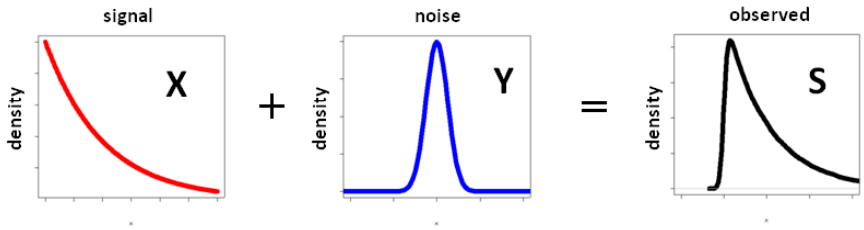
Tras este proceso se obtuvieron los valores de expresión para las 66 muestras. Tal y como se comentó en la Sección 2.2.2, todo el proceso de experimentación con los microarrays introduce de forma inevitable muchas fuentes de error. Por ello se hace imprescindible aplicar un procedimiento de normalización. Dado que no existe ningún método de normalización mejor que el resto en todas las circunstancias (Sección 2.2.2), se comprobó la bondad de la normalización obtenida mediante diversos métodos: MAS 5.0 [1], RMA [130] y GCRMA [282]. De acuerdo con las gráficas MVA, los diagramas de cajas y los histogramas, se estimó que el método de normalización RMA era el más apropiado para normalizar las intensidades (Sección 4.3.3). Seguidamente se describe brevemente dicho método.

Robust Multiarray Analysis (RMA)

Como ya se ha comentado, el proceso de normalización comprende diferentes etapas: corrección de fondo, normalización y agregación. A continuación

se describen de forma breve las estrategias llevadas a cabo en cada una de estas etapas por la metodología RMA.

Corrección de fondo Se utiliza una corrección no-lineal basada en la distribución de las intensidades de las sondas PM. La idea consiste en modelar la intensidad observada S como la suma de una intensidad verdadera X, que sigue una distribución exponencial (y por tanto positiva), y una señal de ruido aleatorio Y, que sigue una distribución normal:



De esta forma, dado un valor s de la señal S, éste es sustituido por el valor de la esperanza de X dado s, es decir, por $E(X|S = s)$.

Normalización

El método RMA lleva a cabo la normalización por cuantiles [38]. El objetivo consiste en lograr que la distribución de las intensidades de todos los arrays del experimento sea la misma. Esto se consigue mediante un sencillo procedimiento de tres pasos que se describen a continuación:

1. Dada una matriz de expresión X, de p filas (genes) y n columnas (arrays), se ordenan los elementos de cada columna. Así, en la Tabla 4.2 se recogen los datos de expresión antes de ser ordenados. La Tabla 4.3 contiene los datos ya ordenados por columnas.

Tabla 4.2: Ejemplo de datos de expresión sin normalizar.

Gen	A1	A2	A3	A4	A5	A6	A7
$Sonda1$	10,002	10,126	10,377	9,332	9,640	10,511	9,030
$Sonda2$	7,393	6,483	8,215	6,697	6,677	6,971	6,970
$Sonda3$	6,428	5,659	7,692	6,866	5,957	5,488	5,973
$Sonda4$	7,592	8,744	8,707	7,842	7,926	7,219	8,365
$Sonda5$	3,363	3,291	3,723	3,181	3,212	3,138	3,128
$Sonda6$	9,356	9,722	9,424	7,783	8,216	7,808	7,793
$Sonda7$	5,276	7,512	5,790	5,324	4,991	5,028	5,279
$Sonda8$	4,142	4,030	4,389	4,631	4,314	4,222	4,633
$Sonda9$	10,357	10,123	10,311	7,509	8,769	6,444	8,247
$Sonda10$	4,272	4,161	4,160	3,805	3,879	3,849	4,192

Tabla 4.3: Columnas de la Tabla 4.2 ordenadas.

A1	A2	A3	A4	A5	A6	A7
10,357	10,126	10,377	9,332	9,640	10,511	9,030
10,002	10,123	10,311	7,842	8,769	7,808	8,365
9,356	9,722	9,424	7,783	8,216	7,219	8,247
7,592	8,744	8,707	7,509	7,926	6,971	7,793
7,393	7,512	8,215	6,866	6,677	6,444	6,970
6,428	6,483	7,692	6,697	5,957	5,488	5,973
5,276	5,659	5,790	5,324	4,991	5,028	5,279
4,272	4,161	4,389	4,631	4,314	4,222	4,633
4,142	4,030	4,160	3,805	3,879	3,849	4,192
3,363	3,291	3,723	3,181	3,212	3,138	3,128

Tabla 4.4: Media por columnas de la Tabla 4.3.

A1	A2	A3	A4	A5	A6	A7	Media
10,357	10,126	10,377	9,332	9,640	10,511	9,030	9,910
10,002	10,123	10,311	7,842	8,769	7,808	8,365	9,032
9,356	9,722	9,424	7,783	8,216	7,219	8,247	8,567
7,592	8,744	8,707	7,509	7,926	6,971	7,793	7,892
7,393	7,512	8,215	6,866	6,677	6,444	6,970	7,154
6,428	6,483	7,692	6,697	5,957	5,488	5,973	6,388
5,276	5,659	5,790	5,324	4,991	5,028	5,279	5,335
4,272	4,161	4,389	4,631	4,314	4,222	4,633	4,375
4,142	4,030	4,160	3,805	3,879	3,849	4,192	4,008
3,363	3,291	3,723	3,181	3,212	3,138	3,128	3,291

2. Se calcula la media por filas de la nueva matriz ordenada (Tabla 4.4) y se sustituyen los valores de cada fila por la media correspondiente (Tabla 4.5).
3. Finalmente, se reordena la matriz obtenida de forma que cada valor ocupe la posición que ocupaba inicialmente (Tabla 4.6).

Agregación

En esta etapa se agregan las intensidades de las sondas de cada conjunto de sondas, de forma que se obtenga una sóla medida de expresión para cada gen. Irizarry et al. consideran que, para cada conjunto de sondas n, las intensidades de las sondas PM una vez ajustado el ruido de fondo, normalizadas y transformadas a escala logarítmica, siguen un modelo lineal aditivo:

$$Y_{ijn} = \mu_{in} + \alpha_{jn} + \varepsilon_{ijn},\ i = 1,...I,\ j = 1,...,J,\ n = 1,...,N,$$

donde:

- I es el número de arrays,
- J es el número de sondas del conjunto de sondas,

Tabla 4.5: Sustitución de los valores de cada fila por las medias de la Tabla 4.4.

A1	A2	A3	A4	A5	A6	A7
9,910	9,910	9,910	9,910	9,910	9,910	9,910
9,032	9,032	9,032	9,032	9,032	9,032	9,032
8,567	8,567	8,567	8,567	8,567	8,567	8,567
7,892	7,892	7,892	7,892	7,892	7,892	7,892
7,154	7,154	7,154	7,154	7,154	7,154	7,154
6,388	6,388	6,388	6,388	6,388	6,388	6,388
5,335	5,335	5,335	5,335	5,335	5,335	5,335
4,375	4,375	4,375	4,375	4,375	4,375	4,375
4,008	4,008	4,008	4,008	4,008	4,008	4,008
3,291	3,291	3,291	3,291	3,291	3,291	3,291

Tabla 4.6: Reordenación de los valores de la Tabla 4.5.

Gen	A1	A2	A3	A4	A5	A6	A7
*Sonda*1	9,032	9,910	9,910	9,910	9,910	9,910	9,910
*Sonda*2	7,154	6,388	7,154	6,388	7,154	9,910	7,154
*Sonda*3	6,388	5,335	6,388	7,154	6,388	6,388	6,388
*Sonda*4	7,892	7,892	7,892	9,032	7,892	8,567	9,032
*Sonda*5	3,291	3,291	3,291	3,291	3,291	3,291	3,291
*Sonda*6	8,567	8,567	8,567	8,567	8,567	9,032	7,892
*Sonda*7	5,335	7,154	5,335	5,335	5,335	5,335	5,335
*Sonda*8	4,008	4,008	4,375	4,375	4,375	4,375	4,375
*Sonda*9	9,910	9,032	9,032	7,892	9,032	7,154	8,567
*Sonda*10	4,375	4,375	4,008	4,008	4,008	4,008	4,008

- N es el número de conjuntos de sondas,
- μ_i representa una medida de la expresión del gen en el array i en escala logarítmica,
- α_j representa el efecto de la sonda concreta,
- ε_{ijn} representa una variabilidad residual, es decir, el error.

Para ajustar los parámetros de este modelo se utiliza el algoritmo *median-polish* [121], el cuál es especialmente robusto frente a valores atípicos (*outliers*). Se trata de un procedimiento iterativo que opera sobre la matriz de expresión, sustrayendo alternativamente la mediana por filas y la mediana por columnas. Las iteraciones continúan hasta que se alcanza la convergencia ó se supera un número de iteraciones máximo. Por lo general, el proceso converge rápidamente en 3 ó 4 iteraciones.

Eliminación de muestras atípicas (*outliers*)

El único procedimiento existente para confirmar la bondad de la normalización obtenida, consiste en el estudio de diferentes gráficos que muestran ciertas propiedades de las intensidades. En primer lugar, se obtuvieron las llamadas gráficas MVA para cada array (Figura 4.3). El acrónimo hace referencia a M vs A, donde M y A se definen como:

$$M = log_2(I_1) - log_2(I_2)$$

$$A = \frac{log_2(I_1) + log_2(I_2)}{2},$$

donde I_1 representa la intensidad del array estudiado, mientras que I_2 representa la intensidad de un "pseudo"-array construido como la mediana de todos los arrays. Así, se espera que los puntos de la gráfica se concentren a lo largo del eje $M = 0$, no debiendo aparecer ninguna tendencia que indique que la media de M varía en función de A. Tal y como se puede observar en la Figura 4.3, la mayor concentración de datos aparece en el eje $M = 0$, no apreciándose otra tendencia en la gran mayoría de los arrays. Sin embargo,

se observan algunos arrays que se desvían del comportamiento deseado, tales como el 13, 14, 30, 44 y otros, lo que indica que que dichos arrays son potencialmente problemáticos.

El estudio de la calidad de la normalización continuó con la obtención de un diagrama de cajas para las intensidades. Estos diagramas permiten comprobar la homogeneidad entre arrays (Figura 4.4). Si los arrays son homogéneos, las cajas deben tener aproximadamente la misma anchura y la misma posición en el eje Y. Tal y como se puede observar en la Figura 4.4, de nuevo existe bastante coherencia entre los arrays, aunque vuelven a aparecer potenciales outliers como el 13, 14, 30 y otros.

Finalmente, el estudio del histograma de intensidades confirmó la presencia de ciertos arrays cuyo comportamiento se desviaba del esperado (Figura 4.5). Por todo esto, se decidió eliminar 8 arrays (13, 14, 30, 39, 44, 46, 61 y 63) del proceso de análisis. Se realizó de nuevo el proceso de normalización descartando estas muestras, obteniéndose así las gráficas de las Figuras 4.6-4.8. Tal y como se puede observar, la mejora en la calidad de la normalización fue significativa, por lo que finalmente se consideraron 58 muestras: 12 correspondientes a tejido no tumoral y 46 correspondientes a tejido tumoral.

Etapas finales del preprocesamiento

Los valores relativos se obtuvieron dividiendo cada medida de expresión proporcionada por el método RMA por un valor de referencia, y calculando el logaritmo en base 2 del resultado tal y como se describió en la Sección 2.2.2. Los valores de referencia se obtuvieron a partir de las medidas de expresión de los 12 tejidos no tumorales. A continuación, se eliminaron aquellos genes que no modificaban su valor de expresión en más de 0.5 *fold* en al menos el 10 % de las muestras (*datos planos*). Tras este último filtrado, se obtuvo una matriz de 46 filas (una por cada muestra de tejido tumoral) y 3708 columnas (una por cada gen restante).

Finalmente, se definieron los conjuntos difusos para las etiquetas lingüísticas *sobre-expresado* y *sub-expresado* tal y como se muestra en la Figura 4.9.

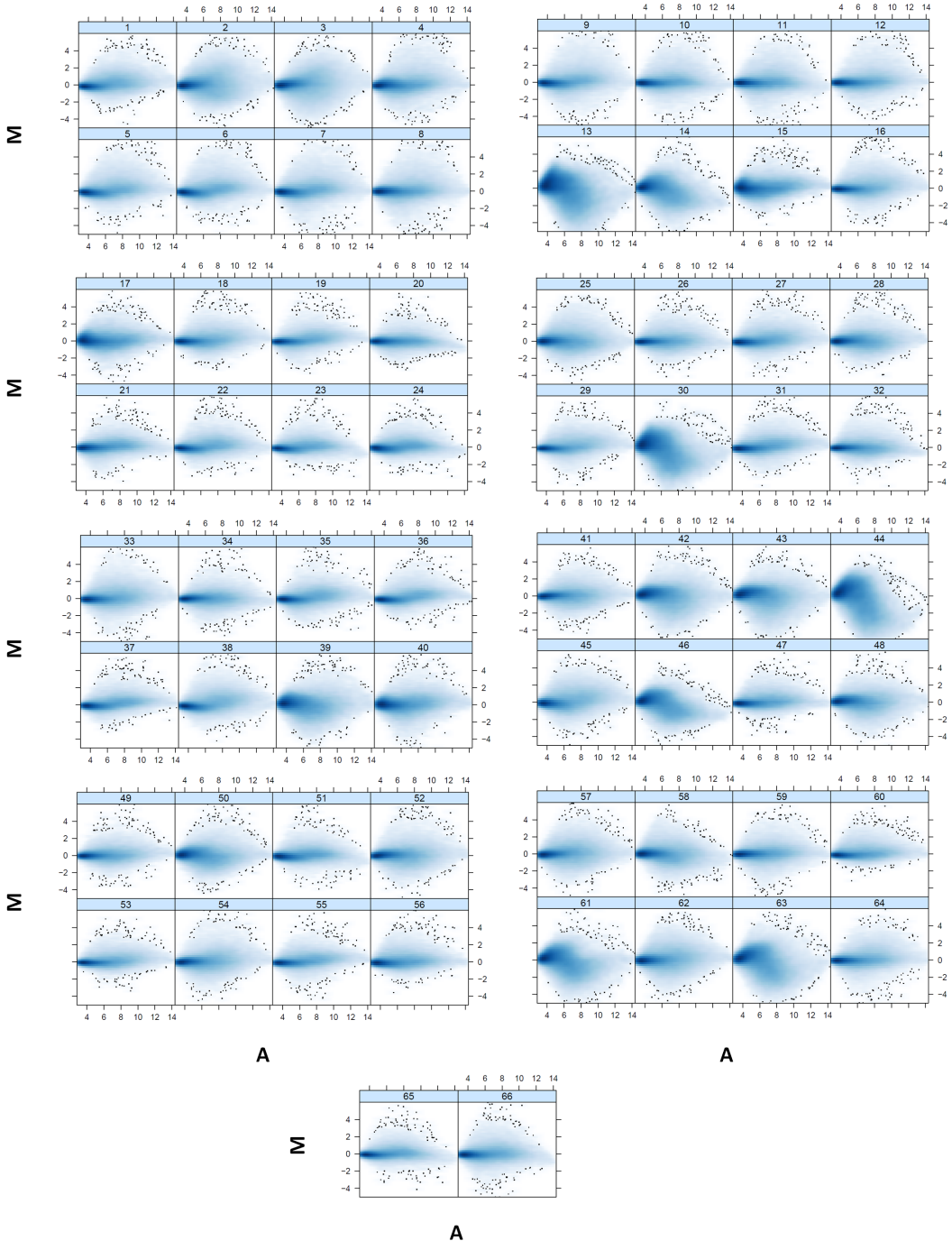


Figura 4.3: Gráficas MVA para los 66 arrays.

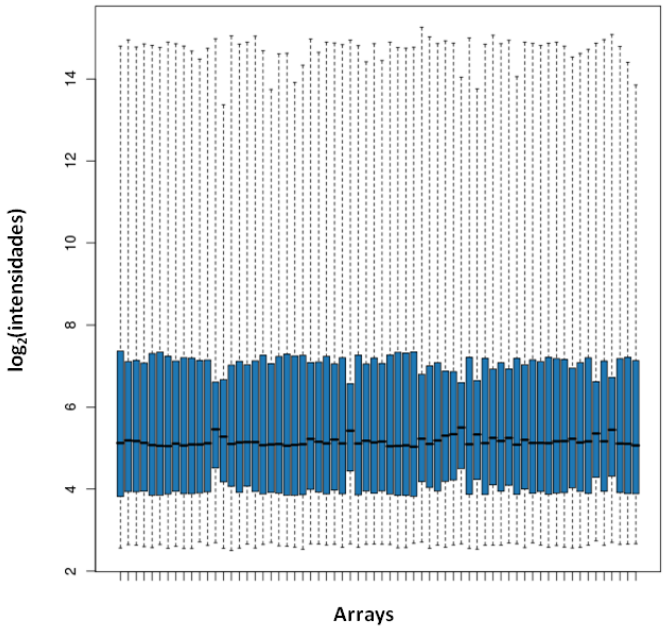


Figura 4.4: Diagrama de cajas.

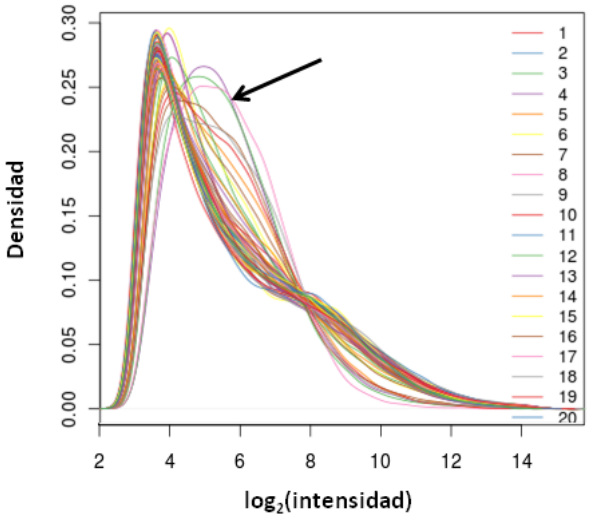


Figura 4.5: Histograma. La flecha indica la zona en la que aparecen arrays potencialmente problemáticos.

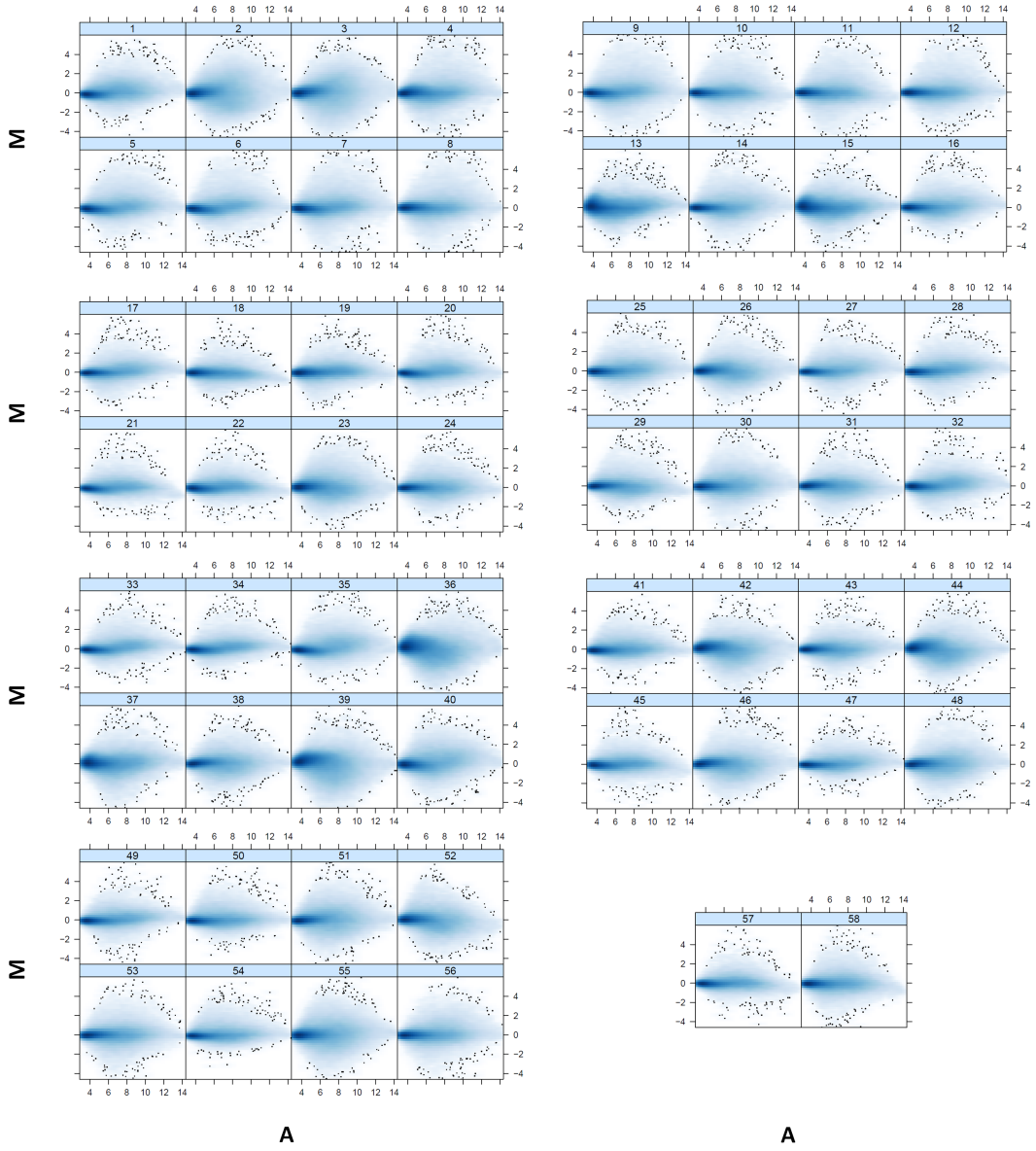


Figura 4.6: Gráficas MVA tras eliminar los outliers.

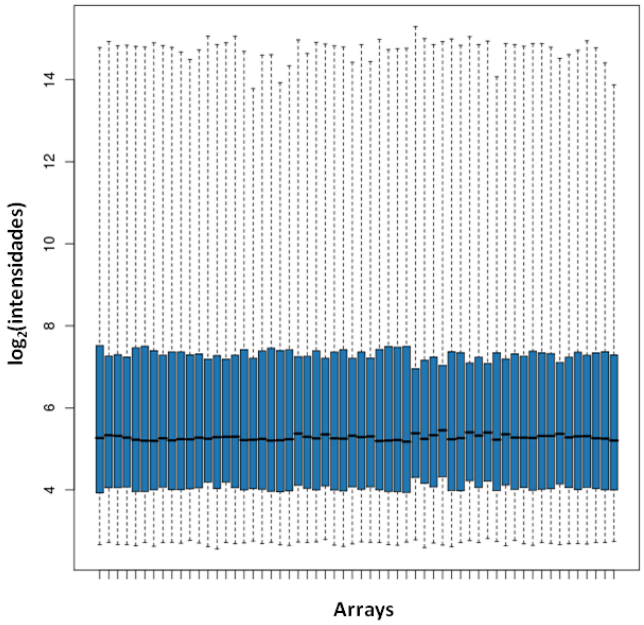


Figura 4.7: Diagrama de cajas tras eliminar los outliers.

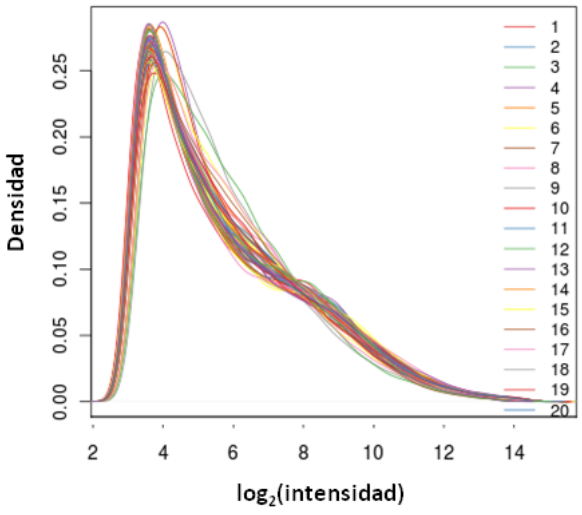


Figura 4.8: Histograma tras eliminar los outliers.

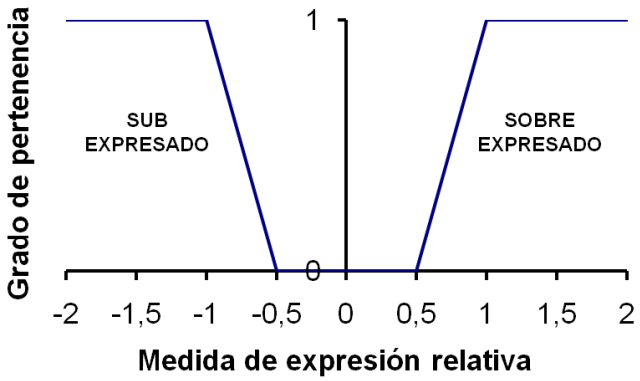


Figura 4.9: Definicion de los conjuntos difusos para la expresión.

4.4 Extracción de reglas de asociación difusas

Como resultado de los procedimientos anteriores, se habían obtenido, por una parte la matriz de expresión, y por otra una tabla de datos con la información de los factores de pronóstico. Se combinaron ambas tablas, construyendo así una tabla final de 46 filas y 3717 columnas. Esta tabla de datos se puede considerar como una tabla transaccional en la que cada tejido tumoral (fila), representa una transacción y los valores de cada variable (columna) forman los items de la transacción (Sección 2.2.3). La Tabla 4.7 muestra las variables incluidas en este conjunto de datos.

Asimismo, tal y como se comentó al inicio de la Sección 4.3, se obtuvo otra tabla adicional conteniendo datos clínicos de 2751 pacientes. La Tabla 4.8 muestra las variables incluidas en este otro conjunto de datos.

Así, el algoritmo de extracción de reglas de asociación difusas se aplicó sobre ambas tablas independientemente, con el objetivo obtener los conjuntos de relaciones potenciales entre las variables correspondientes. Al igual que en el capítulo anterior, en este trabajo se hizo uso del soporte, la confianza y el factor de certeza (FC) para evaluar la fiablidad de las reglas. En este caso sólo se tuvieron en cuenta las llamadas *reglas muy fuertes*, definidas tal y como propusieron Berzal et al. [30]. Es decir, para cada regla $X \rightarrow Y$, se requirió de forma adicional que la regla $\neg Y \rightarrow \neg X$ presentara valores de soporte, confianza y FC mayores que los umbrales especificados por el usuario. Con esta nueva consideración se pretendía hacer frente al hecho de que, tal

Tabla 4.7: Variables incluidas en el estudio de los 46 tejidos tumorales.

Variable	Valores posibles
IHC	$0, 1, 2, 3$
FISH	$+, -$
Índice Allred RE	$+, -$
Índice Allred RP	$+, -$
p53	$+, -$
ki67	$[0, 100]$
Polisomía 17	si, no
Metastasis	si, no
Estadio	I, II, III
Valores de la medida de expresión	$]-\infty, +\infty[$

Tabla 4.8: Variables incluidas en el estudio de los 2751 pacientes.

Variable	Valores posibles
IHC	$0, 1, 2, 3$
FISH	$+, -$
Índice Allred RE	$+, -$
Índice Allred RP	$+, -$
p53	$+, -$
ki67	$[0, 100]$
Polisomía 17	si, no
Heterogeneidad	si, no
Estadio	I, II, III

y como se verá posteriormente, aparecen gran cantidad de variables cuyos valores no están *balanceados*[12], lo que facilita la aparición de reglas falsas. Finalmente, al igual que se hizo en el estudio anterior, se calculó un FDR para proporcionar un valor de la calidad global del conjunto de reglas.

4.5 Resultados

En primer lugar se describe el análisis exploratorio llevado a cabo sobre los datos clínicos de los 2751 pacientes ya mencionados. Se capturan las tendencias descritas previamente entre el conjunto de factores de pronóstico, así como algunas otras asociaciones relevantes. Posteriormente, se comentan las relaciones obtenidas entre los datos de expresión y los factores de pronóstico en el conjunto de datos de las 46 muestras tumorales. De nuevo, las reglas confirman resultados previamente publicados y muestran potenciales biomarcadores en el cáncer de mama. En ambos estudios, los umbrales de FC, confianza y soporte se fijaron de acuerdo con el conocimiento de un experto, tratando al mismo tiempo de mantener el FDR en valores aceptables. En los dos siguientes apartados se muestra una selección de las reglas obtenidas. Dicha selección se obtuvo de acuerdo al conocimiento del experto y a la información obtenida de la bibliografía.

4.5.1 Análisis exploratorio de 2751 pacientes

El primer paso consistió en realizar un estudio descriptivo del conjunto de datos, con el objetivo de obtener una impresión global de la estructura y distribución de los mismos. La Tablas 4.9 a 4.17 y la Figura 4.10 muestran las frecuencias de los items que aparecen en la tabla de datos. Es importante comentar que este estudio presenta ciertas limitaciones impuestas por los datos de que se dispone: los datos clínicos se obtuvieron de muchas fuentes diferentes y falta información acerca de ciertas variables, situación que por otra parte es muy común en los análisis de datos biológicos.

[12]Con valores no *balanceados* se hace referencia a la aparición, para una misma variable, de unos valores muy frecuentes y otros muy poco frecuentes, no estando así equilibrada la proporción de sus correspondientes items en la base de datos.

Tabla 4.9: Resumen de los datos clínicos obtenidos de 2751 pacientes.

	IHC	**FISH**	**ki67**	**RE**	**RP**	**p53**	**Heter.**	**Est.**
Válidos	2680	1960	1960	711	611	755	162	886
Perdidos	71	791	791	2040	2140	1996	2589	1865

Tabla 4.10: Frecuencia de la variable IHC.

Tipo	**Valor**	**Frecuencia**	**Porcentaje**
Válidos	0	696	25.3
	1	1012	36.8
	2	530	19.3
	3	442	16.1
	Total	2680	97.4
Perdidos		71	2.6
Total		2751	100.0

Tabla 4.11: Frecuencia de la variable FISH.

Tipo	**Valor**	**Frecuencia**	**Porcentaje**
Válidos	-	1498	54.5
	+	462	16.8
	Total	1960	71.2
Perdidos		791	28.8
Total		2751	100.0

Tabla 4.12: Frecuencia de la variable Polisomía.

Tipo	Valor	Frecuencia	Porcentaje
Válidos	no	1491	54.2
	si	469	17.0
	Total	1960	71.2
Perdidos		791	28.8
Total		2751	100.0

Tabla 4.13: Frecuencia de la variable RE.

Tipo	Valor	Frecuencia	Porcentaje
Válidos	-	222	8.1
	+	489	17.8
	Total	711	25.8
Perdidos		2040	74.2
Total		2751	100.0

Tabla 4.14: Frecuencia de la variable RP.

Tipo	Valor	Frecuencia	Porcentaje
Válidos	-	363	13.2
	+	392	14.2
	Total	755	27.4
Perdidos		1996	72.6
Total		2751	100.0

Tabla 4.15: Frecuencia de la variable p53.

Tipo	Valor	Frecuencia	Porcentaje
Válidos	-	268	9.7
	+	343	12.5
	Total	611	22.2
Perdidos		2140	77.8
Total		2751	100.0

Tabla 4.16: Frecuencia de la variable Heterogeneidad.

Tipo	Valor	Frecuencia	Porcentaje
Válidos	si	162	5.9
Perdidos		2589	94.1
Total		2751	100.0

Tabla 4.17: Frecuencia de la variable Estadio.

Tipo	Valor	Frecuencia	Porcentaje
Válidos	I	177	6.4
	II	383	13.9
	III	326	11.9
	Total	886	32.2
Perdidos		1865	67.8
Total		2751	100.0

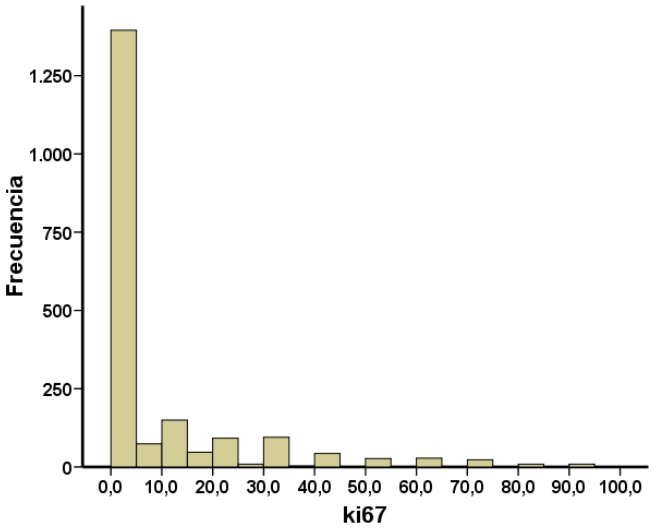


Figura 4.10: Distribución de los valores del marcador ki67.

Tal y como se puede observar (Tabla 4.16), el item que aparece con menos frecuencia en la tabla de datos es $\{Heterogeneidad = si\}$, el cual aparece en tan sólo 162 transacciones. Esto significa que si el umbral de soporte se fija en valores superiores a 0,059, no se obtendrán reglas que relacionen este item con el resto. En general, el soporte del resto de items es relativamente alto (superior a 0.1). Esto significa que hay una alta probabilidad de que dichos items sean parte de los consecuentes de las reglas simplemente por azar, dado que aparecen en muchas transacciones de la tabla de datos [76]. Por lo tanto, podrían esperarse valores relativamente bajos de los factores de certeza.

A continuación, se llevó a cabo el análisis de asociación. Los umbrales de soporte y FC se fijaron en 0.05 y 0.4 respectivamente, obteniéndose así 69 reglas. El valor de FDR calculado fue 0.00, indicando que las reglas obtenidas muy probablemente representen asociaciones reales. La Tabla 4.18 contiene una muestra de las relaciones que se obtuvieron.

Las primeras dos reglas de la Tabla 4.18 apoyan la relación previamente descrita entre el número de copias del cromosoma 17 y la amplificación de HER2 [61, 29]. Asimismo, las cuatro reglas siguientes confirman la conocida correlación entre el estado de los receptores de estrógenos y el estado

Tabla 4.18: Reglas obtenidas del conjunto de datos de 2751 pacientes.

Sop.	Conf.	FC	Regla de asociación
0.44	0.81	0.58	$Polisomia = no \rightarrow FISH = -$
0.44	0.80	0.57	$FISH = - \rightarrow Polisomia = no$
0.13	0.91	0.89	$RP = + \rightarrow RE = +$
0.13	0.73	0.68	$RE = + \rightarrow RP = +$
0.07	0.83	0.81	$RE = - \rightarrow RP = -$
0.07	0.51	0.47	$RP = - \rightarrow RE = -$
0.09	0.58	0.50	$IHC = 3 \rightarrow FISH = +$
0.09	0.56	0.47	$FISH = + \rightarrow IHC = 3$
0.26	0.71	0.35	$IHC = 1 \rightarrow FISH = -$
0.06	0.77	0.50	$ki67 = Bajo \ \& \ IHC = 2 \rightarrow FISH = -$
0.09	0.75	0.45	$Polisomia = no \ \& \ IHC = 2 \rightarrow FISH = -$

de los receptores de progesterona ($p < 0{,}0001$) [43]. Las dos reglas que aparecen a continuación muestran una buena concordancia entre los casos IHC+3 y FISH-positivos, resultado que coincide con las conclusiones de trabajos previos [71]. La no existencia de reglas relacionando los casos IHC0/1 y FISH-negativos se justifica por la presencia de un cierto sesgo en el conjunto de datos. Como ya se comentó en la Sección 4.2.1, por lo general el test FISH sólo se lleva a cabo si un análisis inmunohistoquímico previo clasificó la muestra como dudosa (IHC2) o positiva (IHC3).

Se podría esperar que existiera alguna regla que asociara el itemset $\{IHC = 3, FISH = -\}$ con alguna otra variable, dado que estudios anteriores han descrito que tan sólo los casos FISH-positivos responden al tratamiento con trastuzumab [25, 32, 80]. Sin embargo, no apareció ninguna relación de este tipo. Por ello, se comprobó el soporte de este itemset, por si el umbral fijado para el soporte pudiera estar determinando que dicho itemset no se incluyera en el análisis. Efectivamente, se observó que el soporte del itemset era 0.013. Por esto, se llevó a cabo una ejecución adicional fijando el umbral de soporte

en 0.01. Aún así, no se obtuvo ninguna regla que incluyera a este itemset, lo que indica que muy probablemente no existe una relación de dependencia entre dicho itemset y el resto de variables incluidas en el estudio.

Las dos últimas reglas que se muestran en la Tabla 4.18 son especialmente interesantes, ya que relacionan los casos IHC2 con muestras FISH-negativas. Considérese la asociación que contiene el marcador ki67. La regla $IHC = 2 \rightarrow FISH = -$ no se encontró en el conjunto de reglas, lo que significa que al añadir el itemset $ki67 = Bajo$ al antecedente de la misma, los valores de FC y confianza se incrementan hasta un 0.5 y un 0.77 respectivamente. Esto supone, en nuestra opinión, una variación significativa. En otras palabras, el 77 % de los casos IHC2 que presentan una baja tasa de proliferación ($ki67 = Bajo$) dan lugar a un resultado FISH negativo. Una situación similar ocurre con la última regla. Este tipo de relaciones podrían ser un buen punto de partida para identificar subtipos de tumores en el grupo de muestras IHC2.

Finalmente, merece la pena destacar que no se encontraron asociaciones relevantes que relacionaran el resto de variables. Esto se debe muy probablemente a la presencia de valores perdidos. Es necesario, por tanto, llevar a cabo estudios adicionales que puedan esclarecer la existencia o no de relaciones entre estas variables y el resto de factores de pronóstico.

4.5.2 Datos de expresión y factores de pronóstico

Las frecuencias de los items que aparecen en la tabla de este otro estudio se muestran en las Tablas 4.19 a 4.27 y la Figura 4.11. En este caso, el conjunto de datos presenta algunas diferencias significativas con respecto a los datos del apartado anterior. Así, los items tienen un soporte muy elevado y no se encuentran bien *balanceados*. Por ejemplo, el soporte del item $\{RE = +\}$ es 0.72, mientras que el soporte del item $\{RE = -\}$ es 0.28. Esta situación es bastante frecuente también en el conjunto de genes. Además, el número de items es enorme (7418), al contrario de lo que ocurría en los otros dos casos. Estas características facilitan la aparición de gran cantidad de reglas *falsas*. Por lo tanto, es necesario aplicar una estrategia adicional que permita reducir el FDR.

Tabla 4.19: Resumen de los datos de las 46 muestras tumorales.

	IHC	**FISH**	**Polisomia**	**RE**	**RP**	**p53**	**Est.**	**Metast.**
Válidos	45	42	46	46	46	46	39	31
Perdidos	1	4	0	46	46	0	7	15

Tabla 4.20: Frecuencia de la variable IHC.

Tipo	**Valor**	**Frecuencia**	**Porcentaje**
Válidos	0	3	6.5
	1	7	15.2
	2	22	47.8
	3	13	28.3
	Total	45	97.8
Perdidos		1	2.2
Total		46	100.0

Tabla 4.21: Frecuencia de la variable FISH.

Tipo	**Valor**	**Frecuencia**	**Porcentaje**
Válidos	-	22	47.8
	+	20	43.5
	Total	42	91.3
Perdidos		4	8.7
Total		46	100.0

Tabla 4.22: Frecuencia de la variable Polisomia.

Tipo	Valor	Frecuencia	Porcentaje
Válidos	no	32	69.6
	si	14	30.4
	Total	46	100.0

Tabla 4.23: Frecuencia de la variable RE.

Tipo	Valor	Frecuencia	Porcentaje
Válidos	-	13	28.3
	+	33	71.7
	Total	46	100.0

Tabla 4.24: Frecuencia de la variable RP.

Tipo	Valor	Frecuencia	Porcentaje
Válidos	-	18	39.1
	+	28	60.9
	Total	46	100.0

Tabla 4.25: Frecuencia de la variable p53.

Tipo	Valor	Frecuencia	Porcentaje
Válidos	-	30	65.2
	+	16	34.8
	Total	46	100.0

Tabla 4.26: Frecuencia de la variable Metastasis.

Tipo	Valor	Frecuencia	Porcentaje
Válidos	si	18	39.1
	no	13	28.3
	Total	31	67.4
Perdidos		15	32.6
Total		46	100.0

Tabla 4.27: Frecuencia de la variable Estadio.

Tipo	Valor	Frecuencia	Porcentaje
Válidos	I	3	6.5
	II	11	23.9
	III	25	54.3
	Total	39	84.8
Perdidos		7	15.2
Total		46	100.0

Inicialmente, se identificó el conjunto de variables "problemáticas". Con este objetivo, se generaron 100 conjuntos aleatorios independientes a partir del conjunto de datos original. Se ejecutó el algoritmo de extracción de reglas de asociación difusas sobre cada uno de estos conjuntos, y se contabilizó el número de veces que aparecía cada variable en una regla falsa (Figura 4.12). Se consideró que aquellas variables que aparecían en muchas reglas falsas podrían estar introduciendo demasiada incertidumbre en el conjunto de datos, por lo que se decidió eliminarlas. Así, se descartaron las variables (308) que más veces aparecían en reglas falsas. Estas 308 variables suponen el 30 % de las que formaron parte de alguna regla falsa. Algunos de los factores de pronóstico, como ki67, polisomia-17 y RE se encontraban en este 30 %. Se decidió eliminar la variable ki67, dado que la gran mayoría de las

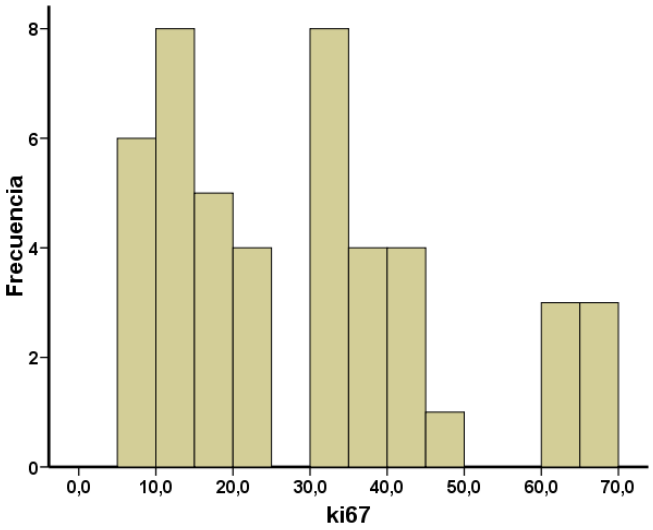


Figura 4.11: Distribución de los valores del ki67.

muestras eran positivas para este marcador (Figura 4.11), lo que significa que podía proporcionar poca información. Observando las frecuencias de las otras dos variables (Tablas 4.22 y 4.23), se puede apreciar que el problema tenía su origen en los items $\{polisomia - 17 = no\}$ y $\{RE = +\}$, que, de hecho, no representan situaciones clínicas "interesantes" (Sección 4.2.1). Por tanto, se eliminaron asimismo estos dos items de la tabla de datos. Así, el conjunto de datos final estaba formado por 3205 variables y 46 muestras de tejido tumoral.

Los umbrales de calidad escogidos fueron 0.8 y 0.1 para el FC y el soporte respectivamente. Así, se extrajeron 1377 reglas con un FDR de 0.06, lo que indica que muy pocas de las reglas se generaron aleatoriamente. No es el objetivo de este trabajo proporcionar una descripción y explicación detallada de todas ellas, sino mostrar que se pueden obtener patrones de comportamiento relevantes. Estos patrones podrían constituir una guía que los expertos podrían utilizar para dirigir sus investigaciones, permitiéndoles enunciar hipótesis que deberían ser posteriormente validadas de forma experimental.

Algunas de las reglas obtenidas vinculan los factores de pronóstico y los valores de expresión de los genes. Más concretamente, 36 de las reglas eran de este tipo. En esta sección se comentarán algunas de ellas (Tabla 4.28). Las

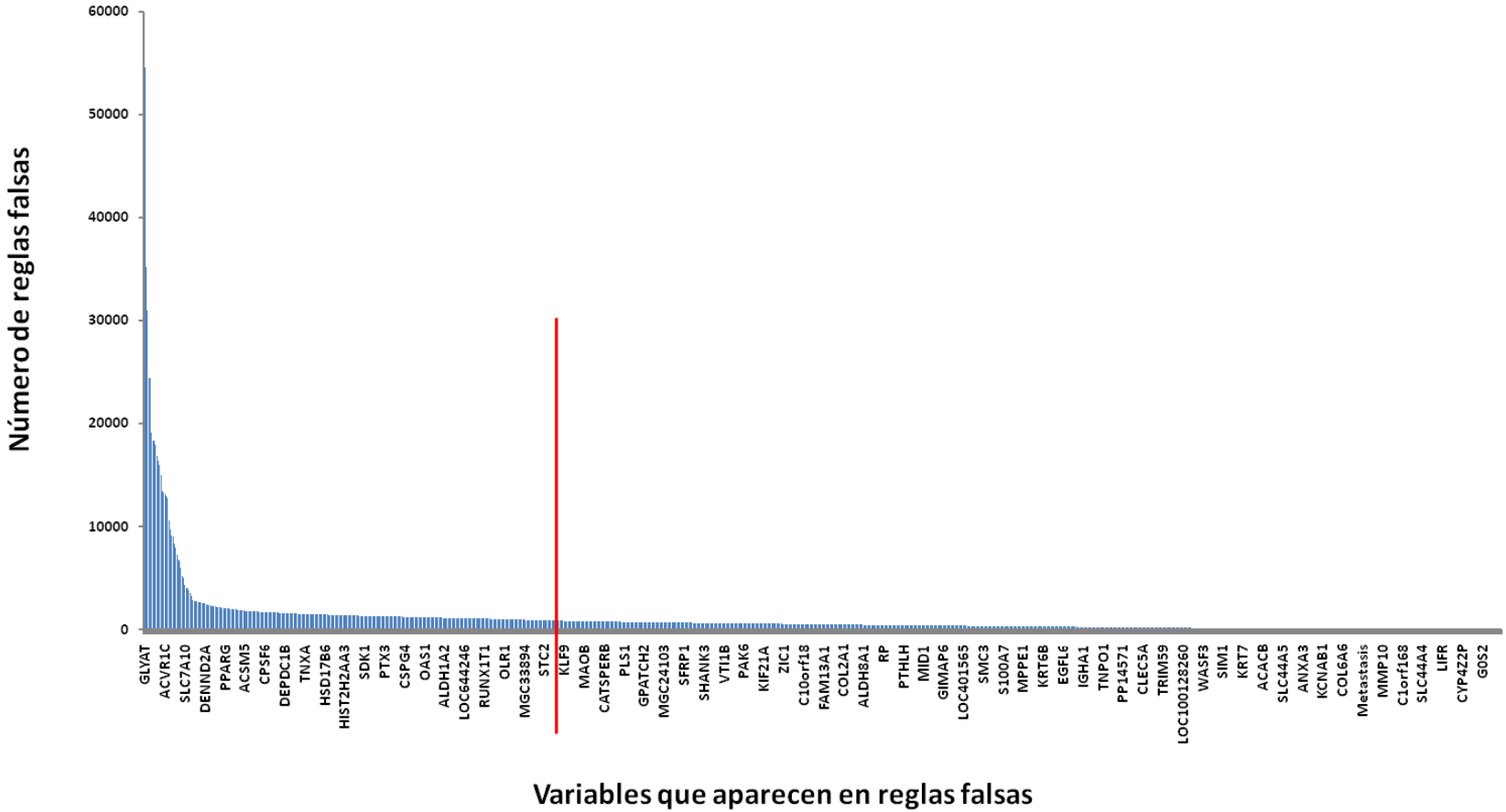


Figura 4.12: Número de reglas falsas generadas por cada variable. La línea roja indica aproximadamente el punto de corte.

reglas que se exponen fueron seleccionadas de acuerdo al conocimiento del experto y a la información obtenida de la bibliografía. Se obtuvieron también reglas con más de un item en el antecedente/consecuente que se tendrán en cuenta en trabajos futuros.

Tabla 4.28: Reglas obtenidas de los 46 tejidos tumorales.

Sop.	Conf.	FC	Regla de asociación
0.11	0.95	0.90	$CDC6 = Sobre \rightarrow FISH = +$
0.11	0.98	0.97	$ESCO1 = Sobre \rightarrow FISH = +$
0.10	1.00	1.00	$BMPR1B = Sobre \rightarrow RP = +$
0.11	0.96	0.91	$THBS1 = Sobre \rightarrow RP = +$
0.18	0.88	0.80	$SOX11 = Sobre \rightarrow RP = -$
0.14	0.97	0.91	$SCUBE2 = Sub \rightarrow RP = -$
0.26	0.93	0.81	$Metastasis = si \rightarrow LPL = Sub$
0.26	0.93	0.81	$Metastasis = si \rightarrow EFEMP1 = Sub$
0.12	1.00	1.00	$EREG = Sobre \rightarrow p53 = -$
0.12	1.00	1.00	$EREG = Sobre \rightarrow RP = +$
0.11	1.00	1.00	$PGR = Sobre \rightarrow RP = +$
0.10	0.96	0.93	$PAK6 = Sobre \rightarrow IHC = 2$
0.12	1.00	1.00	$CLDN11 = Sub \rightarrow Estadio = III$

Algunas de las reglas que se muestran relacionan las medidas de expresión con la amplificación del gen HER2. Así por ejemplo, los tumores HER2 FISH-positivos mostraron sobre-expresión del CDC6. Este gen se localiza en el cromosoma 17q21, cerca del gen HER2, lo que sugiere que ambos genes podrían estar co-expresados. La sobre-expresión del gen ESCO1 aparece también relacionada con muestras FISH-positivas. No se encontró en la bibliografía ninguna mención directa a esta asociación. Sin embargo, un estudio más profundo de la misma podría ser interesante, ya que este gen se ha asociado a distintos tipos de tumores en numerosas ocasiones [239, 40]. De hecho, el gen ESCO1 ha sido descrito por algunos autores como un candidato atractivo

para el desarrollo de una inmunoterapia en una amplia variedad de cánceres [40].

Las dos siguientes reglas de la Tabla 4.28 relacionan los genes BMPR1B y TSP1 (THBS1) con tumores RP-positivos. La sobre-expresión del gen BMPR1B se ha descrito en tumores positivos para los receptores de estrógenos [88, 116], sugiriendo un nuevo objetivo en el tratamiento del cáncer de mama [151]. Asimismo, ciertos autores han encontrado un incremento en los niveles de mARN correspondientes al TSP1 tras la incubación con progesterona [132]. Además, se ha descrito también que los carcinomas invasivos presentan un mayor contenido de TSP1 (una glicoproteína antiadhesiva y antiangiogénica) que las lesiones primarias [134].

Otra regla asocia el gen SOX11 con tumores RP-negativos. Esto la hace especialmente interesante, ya que la sobre-expresión del gen SOX11 podría formar parte de un genotipo vinculado a un mal pronóstico en el cáncer de mama [41]. Por otra parte, la siguiente regla relaciona casos negativos de receptores de estrógenos con la sub-expresión del gen SCUBE2. Los valores de expresión de estos genes podrían servir para predecir la respuesta de ciertos pacientes a la terapia hormonal. Además, estudios previos han relacionado la sub-expresión de SCUBE2 y los casos PR-negativos con un pronóstico pobre [59]. Relaciones como éstas podrían ayudar a identificar subgrupos en el cáncer de mama vinculados a un mal pronóstico.

Como ya se ha indicado, el cáncer de mama es el tipo de cáncer más común entre mujeres, siendo la metástasis la principal causa de muerte de esta enfermedad. En la Tabla 4.28 Aparecen dos reglas que vinculan dos genes (EFEMP1 y LPL) con el proceso de metastasis. El gen EFEMP1 previene la angiogénesis y la infiltración ductal [12], mientras que LPL juega un importante papel en el catabolismo de lípidos. La sub-expresión del EFEMP1 se asocia a una baja supervivencia [213]. El gen LPL podría ser considerado un candidato a gen supresor de la metástasis [254].

De especial interés son también las relaciones entre la sobre-expresión del gen de la epiregulina (EREG), y los casos p53 y PR negativos. Estudios previos han descrito la participación de la epiregulina en ciertos procesos biológicos, tales como el mantenimiento y el desarrollo del crecimiento normal

de la célula, y la progresión de los carcinomas [258]. Esto confiere sentido a la relación, ya que la presencia de la proteína p53 es uno de los mecanismos que evitan la reproducción de las células tumorales (Sección 4.2.1). Estudios en otro tipo de tumor diferente (tumor de vegija) han identificado a los miembros de la familia EGF, especialmente a la epiregulina, como potenciales marcadores en este tipo de tumores [251].

Finalmente, aparece una regla relacionando la sobre-expresión de mARN del PGR con la tinción positiva de los receptores de progesterona, lo que concuerda con el dogma central de la Biología Molecular (Sección 2.1.1). Se muestran también otras dos asociaciones que vinculan la sobre-expresión de los genes PAK6 y CLDN11 con casos IHC2 y el estadio III. El gen PAK6 se ha relacionado previamente con el cáncer de mama [155], aunque no se ha encontrado ninguna mención directa a la relación de dicho gen con casos IHC2. Asimismo, tampoco se encontró información en la bibliografía referente al posible papel del gen CLDN11 en el desarrollo del cáncer de mama. Es necesario, por tanto, llevar a cabo estudios adicionales que confirmen todas estas reglas, dado que estos genes podrían dar lugar a nuevos marcadores en el cáncer de mama.

4.6 Conclusiones

En este Capítulo se ha llevado a cabo un enfoque multidisciplinar para descubrir potenciales relaciones entre datos de expresión y factores de pronóstico en el cáncer de mama. La heterogeneidad e imprecisión de los datos, así como sus grandes dimensiones hacen que la extracción de reglas de asociación difusas sea una herramienta apropiada para el análisis.

Se obtuvieron una serie de asociaciones. El objetivo de este trabajo no consiste en descubrir y establecer el papel de todos estos genes en el cáncer de mama, dado que no se dispone de evidencia científica de estas relaciones. Por el contrario, se pretende mostrar patrones obtenidos con estas técnicas que pueden guiar a los médicos, biólogos, etc., en sus investigaciones, proporcionando relaciones que les permitan enunciar hipótesis para su posterior

validación experimental. Entre las reglas obtenidas se muestran muchas tendencias ya descritas en la bibliografía. Asimismo, aparecen otras que, aún no encontrándose una mención explícita, cobran sentido de acuerdo con las publicaciones consultadas. Finalmente, se muestran también otras relaciones de las que no se dispone evidencia bibliográfica. Los resultados obtenidos permiten, por tanto, sugerir que algunas estas relaciones podrían guiar a los expertos hacia un mejor entendimiento de esta enfermedad. De hecho, actualmente se están realizando pruebas de PCR e inmunostoquímica sobre algunos genes resultantes de las reglas de asociación obtenidas, con objeto de descubrir posibles marcadores tumorales.

CAPÍTULO 5

Estudio de la acción combinada de factores de transcripción en la levadura

5.1 Introducción

El trabajo expuesto en los capítulos anteriores se ha centrado principalmente en la integración y el análisis de datos genómicos heterogéneos. Más concretamente, una parte de dicho trabajo se ha dedicado a estudiar cómo la expresión de los genes se relaciona con otras características funcionales y estructurales del genoma, así como los vínculos existentes entre el nivel de expresión de mARN y ciertos factores de diagnóstico en el cáncer de mama. El trabajo que se presenta en este último Capítulo se dedica al estudio de las causas que originan esos niveles de expresión, en otras palabras, al estudio de los mecanismos que permiten a la célula expresar de forma selectiva un subconjunto de genes.

Tal y como se comentó en la Sección 2.1.4, todas las etapas que dan lugar a la expresión de un gen son regulables. Sin embargo, en la mayoría de los casos la iniciación de la transcripción es el punto de control más importante. En las células eucariotas, la transcripción de un gen está controlada por

combinaciones de proteínas (los factores de transcripción), que se unen de forma coordinada a las secuencias reguladoras de los genes. Estas secuencias reguladoras se encuentran dispersas por todo el ADN y se organizan en los llamados módulos de regulación (CRMs). La unión de los complejos de factores de transcripción a dichos módulos de regulación da lugar a los patrones de expresión adecuados (Sección 2.1.4).

En este Capítulo se presenta una metodología basada en la extracción de itemsets difusos para el estudio de combinaciones de factores de transcripción que participan en las redes de regulación génica. Dada la complejidad del problema, la metodología se ha aplicado en esta primera fase de la investigación sobre el genoma de la levadura. Los resultados se han validado comparándolos con los obtenidos mediante otras técnicas y mediante la herramienta STRING [138], que permite comprobar la validez de las combinaciones desde distintos puntos de vista.

Motivación

Como ya se describió en la Sección 2.2.5, se pueden distinguir dos estrategias fundamentales para reducir el espacio de búsqueda: estudiar tan sólo las regiones promotoras de los genes y centrar la búsqueda en regiones del genoma conservadas entre especies. Respecto a la primera propuesta, estos enfoques no pueden, obviamente, capturar los elementos reguladores que aparecen en los intrones o en zonas alejadas del gen. Experimentos de inmunoprecipitación de cromatina en el genoma completo, han demostrado la existencia de una proporción significativa de TFBSs lejos de la región promotora [207, 56, 129, 173, 248]. Es más, existe toda una serie de estudios, en los que se han descrito elementos reguladores involucrados en el desarrollo primario de los vertebrados y que aparecen lejos del gen que regulan [267, 158, 181, 281].

En cuanto al segundo tipo de métodos, la propuesta es útil, ya que reduce el espacio de búsqueda y mejora la significación de ciertos sitios de unión de factores de transcripción (TFBSs), pero conlleva al mismo tiempo la pérdida de una gran cantidad de potenciales sitios de unión. Las secuencias reguladoras conservadas tienen que ser lo suficientemente parecidas para poder ser

alineadas, algo que no es frecuente en especies no cercanas [267]. Es más, experimentos con especies de mamíferos y *Drosophila* han demostrado que, incluso entre especies muy cercanas, entre uno y dos tercios de los TFBSs identificados no aparecen conservados [267, 78, 68, 87].

Por otra parte, según el objetivo pretendido por los diferentes métodos, se pueden distinguir tres categorías: *a)* métodos que escanean secuencias (o genomas completos) buscando CRMs que siguen un modelo predefinido, *b)* métodos que buscan CRMs similares en un conjunto de genes relacionados (por ejemplo co-expresados o co-regulados), y *c)* métodos que escanean secuencias (o genomas completos) buscando grupos de TFBSs de cualquier combinación de factores de transcripción (Sección 2.2.5). Los métodos que se enmarcan en la primera de estas clases, se basan en la búsqueda de un modelo de CRM estrictamente definido. Requieren que el usuario especifique de forma rigurosa lo que está buscando, algo que no siempre es posible. Estos métodos suelen recibir como entrada una combinación de PWMs, la secuencia y un número variable de parámetros. Debe tenerse en cuenta que tal información tan exhaustivamente definida sólo se encuentra disponible para un número muy limitado de procesos [266]. Además, llama la atención el hecho de que uno de los principales métodos de este tipo, el *Enhancer Element Locator (EEL)* [108], imponga la restricción de que el orden de los sitios de unión funcionales debe conservarse entre especies, algo que se sabe que no se corresponde con la realidad [66, 217].

Respecto a la segunda clase de métodos, requieren especificar un conjunto de genes co-regulados o co-expresados (o sus correspondientes secuencias reguladoras), ya que tratan de detectar CRMs similares, asumiendo que patrones de expresión similares vienen dados por elementos reguladores similares. Dado que estos métodos se centran en un conjunto concreto de genes, el estudio de los elementos que los regulan requiere el análisis de sus secuencias promotoras, lo que lleva consigo las limitaciones previamente comentadas.

Finalmente, se encuentran los métodos que escanean genomas completos buscando CRMs formados por cualquier combinación de TFBSs. Estos métodos presentan la ventaja de que no necesitan realizar ningún tipo de asunción acerca del conjunto de TFs que cooperan. Se trata de métodos más generales

que los anteriores y, por tanto, no requieren ningún tipo de conocimiento previo (a excepción de una biblioteca de PWMs, igual que en los casos anteriores). Como consecuencia, estos métodos no permiten inferir la función de los CRMs que predicen ni el patrón de expresión que producen. Sin embargo, se cree que esta debilidad se puede suplir combinando los modelos CRM inferidos por esta técnica, con las otras metodologías anteriormente mencionadas. En este tipo de métodos se enmarca el trabajo de Morgan et al. [176] presentado en la Sección 2.2.5. Dicho trabajo puede parecer similar en algunos aspectos al que aquí se propone, por lo que merece la pena destacar algunas de las limitaciones que presenta, según nuestro punto de vista:

- La división del genoma en intervalos consecutivos de igual tamaño no modela adecuadamente la realidad. Ciertamente, los módulos de regulación se distribuyen a lo largo de todo el genoma, pero no tienen por qué ajustarse a esta división regular de intervalos, lo que provocará que muchos de los módulos reales queden divididos en dos o más de estos intervalos.
- La división en intervalos crisp provoca nuevamente la aparición del *sharp boundary problem*, acentuado en este caso por el hecho de que no se conoce con exactitud la posición de los TFBSs ni el tamaño exacto de los módulos de regulación.
- Tan sólo se consideran reglas de asociación de un item en el antecedente y el consecuente, lo que limita significativamente la metodología, ya que permite tan sólo estudiar combinaciones de dos TFs.
- Parece más adecuado modelar las combinaciones de TFs mediante otro tipo de expresión diferente a la regla de asociación. La unión de los TFs a los módulos no se trata tanto de una acción causa-efecto, como de la acción combinada y coordinada de un conjunto de TFs. En otras palabras, una regla $TFA \rightarrow TFB$ indica que si se produce la unión del factor de transcripción A, entonces también se produce la unión del factor de transcripción B. Esto no tiene por qué ser así en la realidad, ya que por lo general dicho factor de transcripción A podría unirse

en muchas otras ocasiones con otros factores de transcripción, lo que significa que las reglas de asociación no capturarían muchas relaciones.

- Para cada par de TFs, si dos de sus correspondientes TFBSs se solapan en un intervalo dado, ambos TFBSs son eliminados. Esto tampoco parece modelar adecuadamente la realidad, ya que podría aparecer otro TFBSs para uno de los dos TFs que no se solape con los anteriores y que, por tanto, permita la unión de ambos TFs simultáneamente (Sección 5.2.4).

Además de todo lo anterior, es importante destacar que la información que se maneja en estos estudios es imprecisa y presenta un alto grado de incertidumbre, lo que complica aún más el problema. La imprecisión surge de forma inmediata al pensar en el tamaño de los módulos, ya que se asume que comprenden varios cientos de pares de bases, pero no hay un tamaño fijo de los mismos. Otro aspecto determinante es la inexactitud en la localización de los TFBSs, especialmente si dicha información se obtiene mediante procedimientos *in silico*. Es más, el proceso de inferencia de posibles sitios de unión lleva consigo un alto número de falsos positivos provocados por:

- la complejidad probabilística del problema, ya que se trata de buscar secuencias muy pequeñas ($10-30$ pares de bases, *pb*) en cadenas de varios cientos de millones de pares de bases,
- la propia complejidad biológica del problema, ya que se pueden observar muchas subsecuencias de características idénticas a TFBSs conocidos y que sin embargo no representan sitios de unión funcionales.

En cualquier caso, éste es un problema intrínseco a todas las técnicas computacionales de detección de CRMs, ya que en algún momento del proceso todas ellas infieren la localización de potenciales TFBSs. Por tanto, la presencia de la imprecisión e incertidumbre en los datos es clara y sin embargo, son exiguas las publicaciones en las que se hace uso de técnicas difusas para modelarla.

Propuesta

En este Capítulo se presenta una metodología que solventa algunas de estas limitaciones. En primer lugar, se escaneará un genoma completo buscando

grupos de potenciales TFBSs. A continuación, se aplicara un algoritmo de extracción de itemsets difusos para obtener co-ocurrencias significativas de TFs en el genoma. El procedimiento sólo requiere como entrada la secuencia del genoma y un conjunto de PWMs previamente definidas.

Mediante el escaneo del genoma completo se evitará la pérdida de información que conlleva el centrar la atención en regiones concretas. Dicho escaneo se llevará a cabo mediante una herramienta de detección de TFBSs (Sección 2.2.4). Tal y como ya se ha comentado este proceso no está exento de limitaciones, ya que genera un alto número de falsos positivos. Sin embargo, se piensa que estos efectos negativos pueden paliarse mediante las siguientes etapas de filtrado:

1. Agrupación de los TFBSs inferidos, aplicando para esto un algoritmo de clustering jerárquico. Tal y como se indicó anteriormente (Sección 2.2.5), la aparición de grupos de TFBSs en una pequeña zona del genoma se considera un indicador fiable de su funcionalidad [267, 92, 142, 7, 224, 34, 176].
2. Búsqueda de itemsets frecuentes en los grupos obtenidos. La extracción de itemsets frecuentes se ha llevado a cabo con éxito en diferentes estrategias [241, 176, 198]. La exigencia de que las combinaciones aparezcan de forma repetida ayudará al filtrado de ocurrencias espurias.
3. Cálculo de la significación estadística de las combinaciones obtenidas, lo que proporcionará un nivel de la fiabilidad a los resultados, permitiendo eliminar las combinaciones no relevantes.

Además de lo anterior, la metodología presenta otras características interesantes. Por ejemplo, el uso del clustering jerárquico permite modelar los grupos de TFBSs de forma más adecuada a como lo hacen los intervalos consecutivos de tamaño fijo, además de evitar el *sharp boundary problem* mediante la definición de conjuntos difusos. Por otra parte, la aplicación del algoritmo de extracción de itemsets difusos permitirá obtener de forma eficiente cualquier combinación de TFs, evitando restricciones en el tamaño de las combinaciones obtenidas o en el orden de aparición de los TFBSs. Finalmente, los itemsets que pasan el filtro estadístico modelan mejor los CRMs

que las reglas de asociación, representando la co-ocurrencia significativa de los TFs correspondientes.

Como ya se ha comentado en repetidas ocasiones, la integración de información de diversas fuentes en el estudio es crucial para situar los resultados en el contexto adecuado, obtener toda la información posible del análisis y validar los resultados. Con este objetivo se hará uso de la herramienta STRING [138]. Dado un conjunto de genes/proteínas, STRING busca asociaciones entre ellos a diferentes niveles: localización cercana en el genoma, co-ocurrencia entre especies, co-expresión, interacciones proteína-proteína, bases de datos, minería de textos, etc. De esta forma, STRING proporcionará asociaciones a distintos niveles entre los TFs de la combinaciones obtenidas.

5.2 Métodos

5.2.1 Datos

Se llevaron a cabo tres estudios de forma independiente. En el primero de ellos, se consideraron únicamente los TFBSs validados experimentalmente por Harbison et al. [110]. Es decir, en este primer análisis se omitió el proceso computacional de inferencia de TFBSs. Los datos se obtuvieron de la *Saccharomyces Genome Database* (SGD) [84]. Este conjunto de datos contiene la localización de 3328 sitios de unión de 102 factores de transcripción.

En el segundo estudio, se utilizaron los TFBSs inferidos mediante una herramienta especializada (Sección 2.2.4). Para esto, se descargaron el genoma completo de la levadura de la SGD y las 177 PWMs de Jaspar [215] que corresponden a la levadura (versiones de Diciembre de 2009). Finalmente, en el tercer estudio se combinaron los datos de Harbison et al. con los TFBSs inferidos por la herramienta de detección.

5.2.2 Construcción de la base de datos transaccional difusa

En primer lugar, se aplicó una herramienta de detección de TFBSs sobre el genoma completo de la levadura. Una vez obtenidas las coordenadas de los

posibles sitios de unión, se buscaron en el genoma grupos de TFBSs cercanos. Para esto, se aplicó un algoritmo de clustering jerárquico, fijando el punto de parada del algoritmo en $300pb$. Es decir, se buscaron grupos de TFBSs que comprendieran aproximadamente $300pb$, asumiendo que los módulos de regulación abarcan por lo general unos pocos cientos de pares de bases [17].

Una vez obtenida la lista de clusters, se generó una transacción difusa por cada uno de ellos. Las funciones de pertenencia de cada transacción se definieron de la siguiente forma:

- Se calculó el centroide C de cada cluster como la mediana de los puntos que contiene.
- La región constante de la función trapezoidal se definió entre los valores $C-150$ y $C+150$, tal y como muestra la Figura 5.1.
- La región lineal creciente se definió desde $C-250$ hasta $C-150$, mientras que la región lineal decreciente se definió entre $C+150$ y $C+250$ (Figura 5.1).

Una vez definidos los grados de pertenencia se generó la base de datos transaccional difusa. La Figura 5.1 muestra gráficamente el procedimiento completo.

5.2.3 Extracción de los itemsets frecuentes difusos

Se aplicó el algoritmo de extracción de itemsets difusos desarrollado en el Capítulo 3 sobre las dos bases de datos transaccionales. Tal y como se comentó en la Sección 5.1, además del valor del soporte se calculó un p-valor para cada itemset. Con este objetivo, se adaptó al caso difuso el procedimiento descrito en el trabajo de Gallo et al. [94]. El *modelo nulo* para el cálculo de este p-valor representa la situación "no interesante" en la que no existen asociaciones entre los items, es decir, los items aparecen independientemente unos de otros en las transacciones. Así, el p-valor representa la probabilidad de que el itemset sea "sorprendente" bajo este modelo nulo. Por tanto, cuanto menor sea el p-valor, más interesante será el itemset.

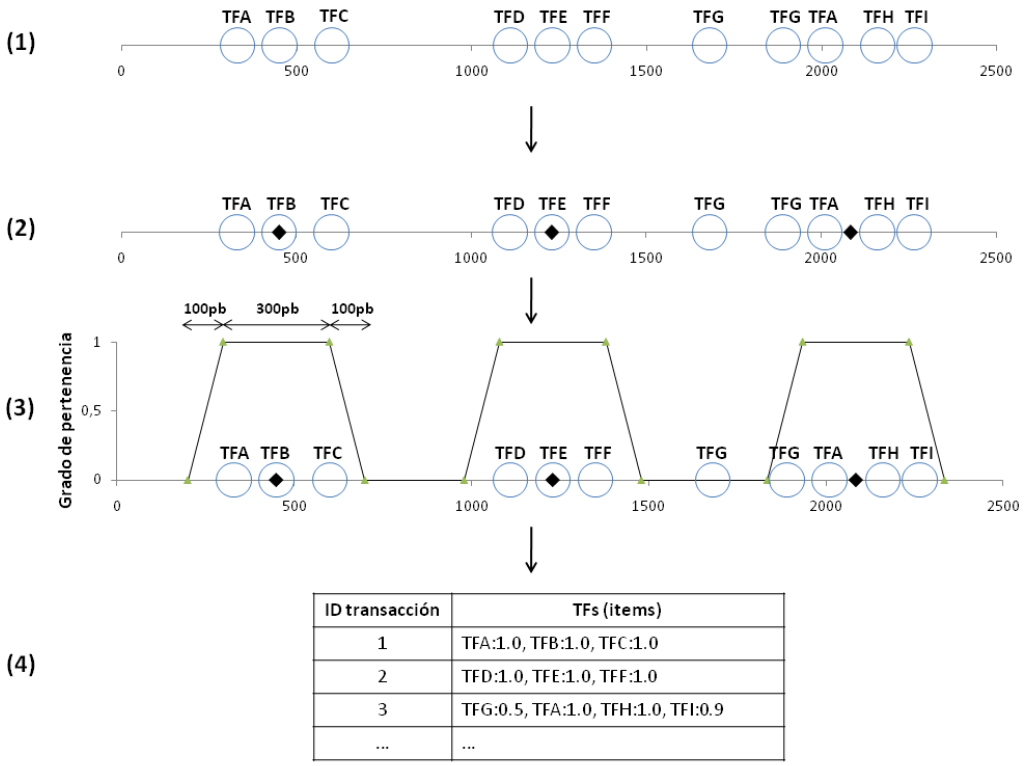
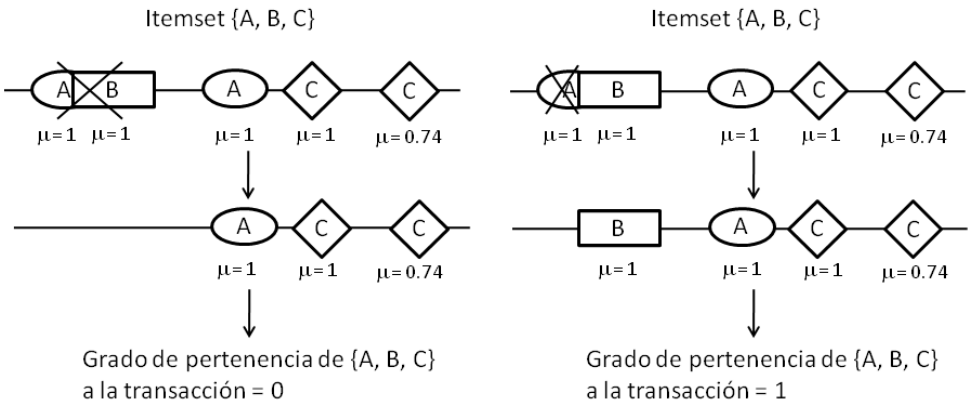


Figura 5.1: (1) Cada círculo representa un sitio de unión. Cada sitio de unión aparece etiquetado con el TF que se une a dicho sitio de unión. (2) Se obtienen 3 clusters y se calculan sus centroides. (3) Se definen conjuntos difusos para cada cluster. (4) Se genera una transacción difusa por cada conjunto difuso. El número tras los dos puntos indica el grado de pertenencia del TF correspondiente a la transacción.

5.2.4 Post-procesado del conjunto resultante

Es necesario tener en cuenta la presencia de TFBSs solapados antes de generar el conjunto resultante final. Por ejemplo, supónganse los sitios de unión de la Figura 5.2. Algunos trabajos previos [176] eliminan directamente dos sitios de unión si se solapan. Esto podría conllevar un cálculo incorrecto de las co-ocurrencias de los TFBSs, dado que es posible que apareciera una combinación de sitios de unión que permitiera la unión simultánea de ambos TFs (Figura 5.2b, el valor de μ indica el grado de pertenencia del TFBS a la transacción). Por tanto, en este trabajo se buscó la forma óptima de "enca-



(a) Eliminación de los sitios de unión solapados.

(b) Forma óptima de encajar el itemset.

Figura 5.2: Procesado de los TFBSs solapados.

jar" cada combinación de TFs (itemset) en una transacción difusa dada, de forma que se maximizara el grado de pertenencia del itemset a la transacción (Figura 5.2b).

5.3 Resultados

5.3.1 Análisis de los datos de Harbison et al.

En este primer estudio se obtuvieron 570 transacciones, con una media de 2.79 TFs diferentes por transacción y un máximo de 10 TFs. La Figura 5.3 muestra la frecuencia de aparición de los 96 TFs que quedaron capturados en alguna transacción. Los umbrales de soporte y p-valor se fijaron en 0,01, obteniéndose de este modo 32 itemsets. La obtención de p-valores muy por debajo del umbral fijado, indica que la gran mayoría de dichos itemsets representan asociaciones biológicas reales.

Tal y como se comentó en la Sección 5.1, se hizo uso de la herramienta STRING para obtener distintos tipos de evidencias que apoyaran las relaciones obtenidas. Esta herramienta proporcionó relaciones entre los factores de transcripción de 23 de los itemsets obtenidos. Es más, para 18 de estos itemsets, los grafos que representan las asociaciones entre sus TFs resultaron ser

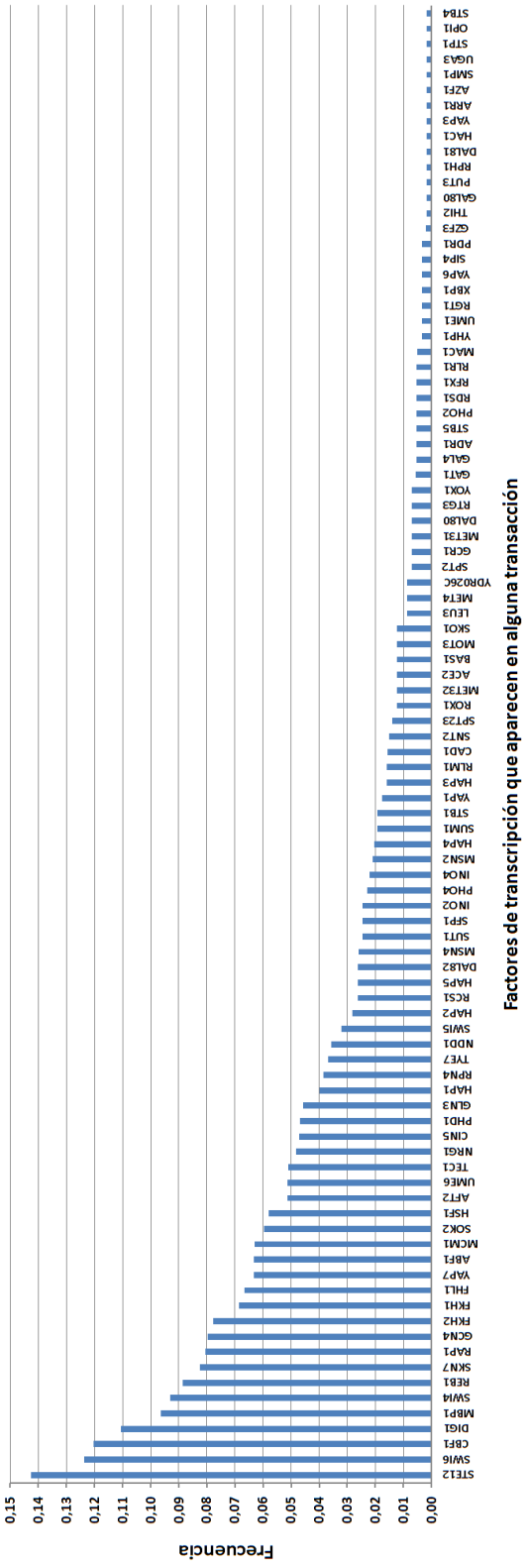


Figura 5.3: Frecuencia de los TFs en las transacciones obtenidas con el conjunto de datos de Harbison et al.

Tabla 5.1: Combinaciones de TFs

ID Itemset	TFs	p-valor	Soporte
1	$SWI6, MBP1, SWI4$	$3{,}807e-17$	0,025
2	$SKN7, SOK2, PHD1$	$2{,}661e-13$	0,014
3	$STE12, DIG1, TEC1$	$7{,}522e-12$	0,018
4	$RAP1, FHL1$	$6{,}661e-08$	0,016
5	$DIG1, SWI4, TEC1$	$5{,}129e-07$	0,011
6	$AFT2, RCS1$	$8{,}090e-07$	0,012
7	$PHD1, SUT1$	$9{,}659e-07$	0,011
8	$STE12, SWI6, SWI4$	$1{,}086e-06$	0,014
9	$FKH2, NDD1$	$2{,}925e-06$	0,016
10	$SWI6, DIG1, SWI4$	$2{,}934e-06$	0,012

grafos *conexos*, es decir, existe un camino que conecta cada par de TFs del grafo.

La Tabla 5.1 muestra una selección de las combinaciones de TFs obtenidas. Los grafos generados por STRING para dichas combinaciones se pueden encontrar en la Figura 5.4. Por ejemplo, STRING devolvió grafos conexos para las combinaciones 1, 3, 4, 6 y 9. Los factores de transcripción del itemset 1 aparecen relacionados a diferentes niveles (Figura 5.4). Esta combinación tiene sentido, ya que se sabe que SWI6 interacciona con MBP1 y SWI4 para formar dos complejos proteicos denominados MBF y SBF, los cuales a su vez cooperan y juegan un papel fundamental en la progresión de la fase G1 a la fase S del ciclo celular [150]. De hecho, merece la pena comentar que las combinaciones $\{SWI6, MBP1\}$ y $\{SWI6, SWI4\}$, se obtuvieron también por separado con p-valores $1{,}795e-11$, $2{,}245e-18$ y soportes 0,0434, 0,056 respectivamente.

Otra combinación cuyos TFS aparecen fuertemente relacionados es la que se muestra en el itemset 3. En este caso las tres proteínas forman el complejo STE12/TEC1/DIG1, el cual se sabe que se une a las zonas promotoras de genes de filamentación [64]. Por otra parte, STRING encontró evidencias a nivel experimental y a nivel de textos para los itemsets 4, 6 y 9. El itemset 4 captura una combinación interesante. La co-ocurrencia de los sitios de unión de RAP1 y FHL1 está en concordancia con resultados previos, ya que se ha

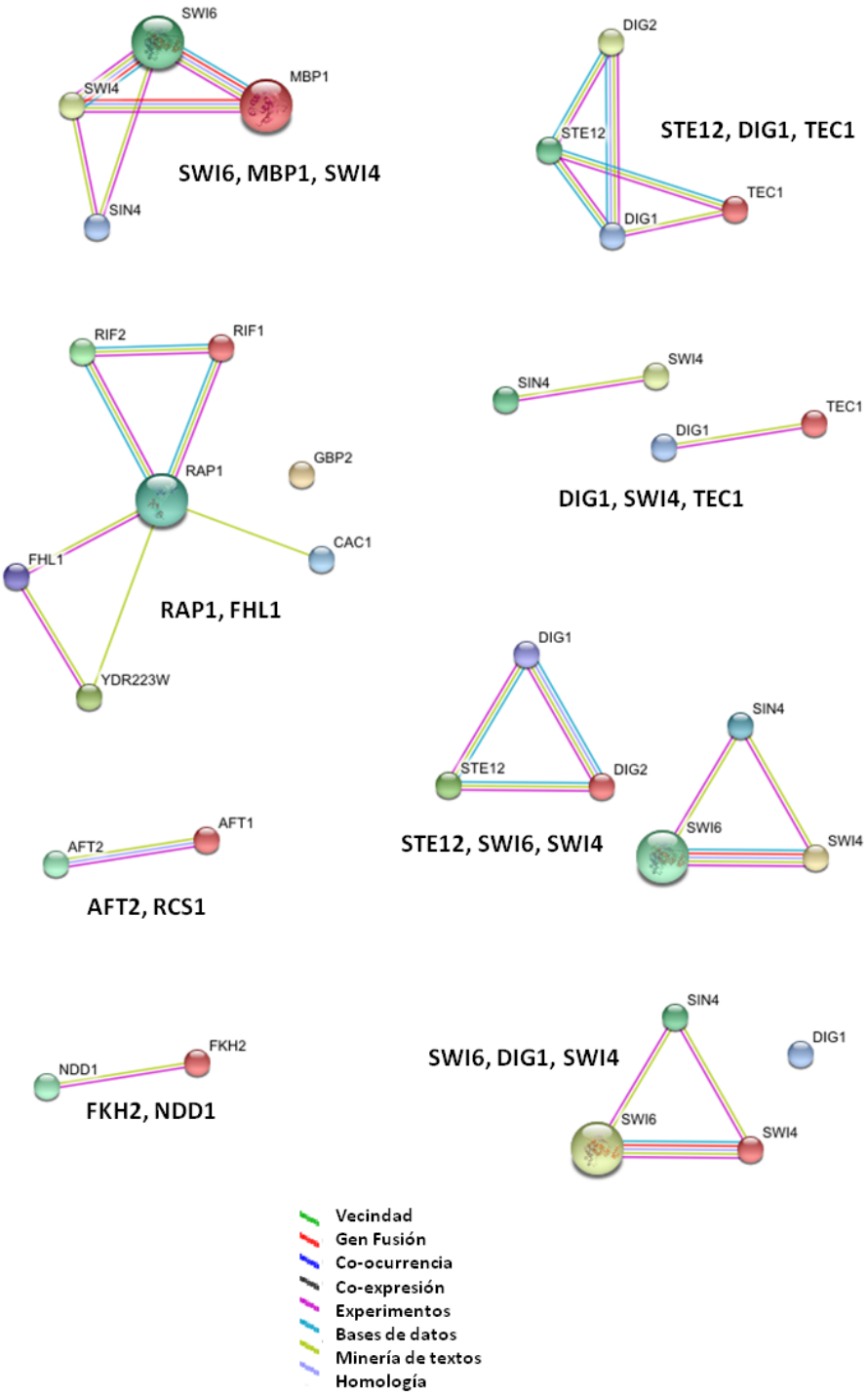


Figura 5.4: Grafos de las combinaciones obtenidas con los datos de Harbison et al..

demostrado que ambos TFs se unen a la región promotora de muchos genes codificadores de proteínas ribosómicas [219]. Otro caso interesante es el del itemset 6. La co-ocurrencia significativa de sus TFBSs aparece también apoyada por la bibliografía, dado que AFT2 y RCS1 (AFT1) son un par de activadores de la transcripción de respuesta al hierro [69]. Finalmente, la aparición de los TFs FKH2 y NDD1 en el itemset 9 se encuentra también justificada, dado que se ha demostrado que FKH2 facilita la incorporación del NDD1 a un complejo protéico que regula la expresión del cluster de genes CLB2 [200].

En cuanto a los itemsets 5, 8 y 10, STRING encontró relaciones entre algunos de los TFs de dichas combinaciones (Figura 5.4). Por ejemplo, no se encontró relación entre SWI4 y los otros dos TFs del itemset 5. Es más, se obtuvo una combinación relacionando SWI4 y TEC1 con p-valor $4{,}383e-04$ y soporte 0.018. Sin embargo, STRING tan sólo obtuvo evidencias experimentales y basadas en textos apoyando la relación entre DIG1 y TEC1 (itemset 3). Ambos genes forman parte de un complejo que regula la transcripción de genes de filamentación [64]. Una situación similar aparece para los itemsets 8 y 10, para los que no se obtuvo ninguna relación ni entre STE12 y los otros dos TFs, ni entre DIG1 y los otros factores del itemset 10.

Finalmente, no se encontró ningún tipo de evidencia que confirmara los itemsets 3 y 7. Pham et al. [198] describieron previamente alguna relación acerca de la asociación $SOK2-PHD1$ (véase el material suplementario de dicho trabajo). Por otra parte, se pueden encontrar relaciones indirectas entre algunos de los TFs del itemset 3 en el trabajo de Bar-Joseph et al. [24]. Es por tanto necesario realizar un estudio detallado y una evaluación empírica para confirmar todas estas asociaciones.

5.3.2 Detección de potenciales TFBSs

Selección de la herramienta y ajuste de sus parámetros

En este proceso se hizo uso de Patser [119], ya que se trata de una herramienta cuyo buen funcionamiento ha sido demostrado en repetidas ocasiones [178, 249, 214, 176]. Si bien es cierto que existen herramientas más

recientes que tienen en cuenta dependencias posicionales, su aplicabilidad limitada. Esto último se debe a que estas herramientas requieren la lista de secuencias a partir de las que se generó la PWM, información que ni siquiera las principales bases de datos, como Jaspar [215] o TRANSFAC [277], publican habitualmente. De hecho, Jaspar no proporciona estas secuencias para ninguno de los 177 motivos de la levadura que contiene. En el caso de TRANSFAC (versión de Marzo de 2008), tan sólo 15 de las 37 PWMs de la levadura aparecen acompañadas de sus correspondientes secuencias.

Patser requiere como entrada una PWM y una secuencia, y devuelve una medida del ajuste de la PWM en cada subsecuencia de la secuencia proporcionada (rango $[0, 15]$). Así, a fin de determinar el conjunto de subsecuencias que podrían considerarse TFBSs, es necesario fijar un umbral para la medida de ajuste. Asimismo, cada motivo presenta ciertas peculiaridades propias, lo que puede determinar que los TFBSs correspondientes a ciertos motivos sean más fácilmente detectados que otros. Por tanto, se hace necesario fijar un umbral independiente para cada motivo [259].

En primer lugar, se fijó globalmente un valor mínimo de la medida por debajo del cual no era factible situar el umbral de los motivos. Dicho mínimo global se fijó en 7,5. La selección de este valor se realizó de acuerdo a la documentación del método [5] y a la observación propia de los resultados. Así, se observó que valores por debajo de 7,5 generan un número ingente de TFBSs potenciales, lo que restaría significación a la aplicación de la metodología. Como ya se comentó anteriormente (Sección 5.1), la aparición de TFBSs espurios viene determinada por la propia complejidad del problema y se trata, en la actualidad, de un problema inherente a todas las técnicas de detección de motivos y CRMs.

Así, para cada motivo, se buscó un umbral específico por encima de 7,5, de tal forma que Patser capturara el maximo número posible de los TFBSs descritos por Harbison et al. [110]. La Figura 5.5 muestra los umbrales obtenidos para cada motivo. Además, para algunos de los motivos extraídos de Jaspar no se disponía de sus correspondientes TFBSs, ya que no aparecían en los datos de Harbison et al., por lo que hubo que seleccionar umbrales *ad hoc* para cada uno de ellos. Con este objetivo, se analizaron los umbrales ya

calculados para los motivos que sí aparecían en los datos de Harbison et al. El objetivo de este análisis consistía en buscar alguna dependencia entre el valor del umbral y ciertas propiedades del motivo, tales como la longitud del mismo, su *contenido de información* o el *contenido de información* de su *núcleo*. Encontrar dependencias podría ayudar a fijar el umbral de estos motivos de una forma más adecuada en función de dichas propiedades.

La longitud de un motivo viene dada por el número de bases de los TFBSs a los que representa. Así por ejemplo, la longitud del motivo que aparece en la Sección 2.2.4 es 10. El *contenido de información* de una posición del motivo proporciona una medida del nivel de *conservación* de esa posición, en otras palabras, da una idea de si ciertas bases aparecen claramente en dicha posición, o de si las cuatro bases pueden aparecer indistintamente. Para una posición dada, el *contenido de información* se calcula como:

$$2 + \sum_{\beta \in \{A,C,G,T\}} W_\beta log_2(W_\beta),$$

donde W_β representa la frecuencia relativa de aparición de la base β en dicha posición. Por ejemplo, el contenido de información de la primera posición del motivo de la Sección 2.2.4, vendría dado por:

$$2 + \frac{2}{10} \cdot log_2\left(\frac{2}{10}\right) + \frac{1}{10} \cdot log_2\left(\frac{1}{10}\right) + 2 + \frac{1}{10} \cdot log_2\left(\frac{1}{10}\right) + \frac{2}{10} \cdot log_2\left(\frac{2}{10}\right) = \\ = 2 + (-0{,}46) + (-0{,}33) + (-0{,}33) + (-0{,}46) = 0{,}42$$

El *núcleo* de un motivo está constituido por sus 5 posiciones consecutivas más conservadas, es decir, aquellas de mayor contenido de información. Así, se analizaron posibles dependencias entre el valor de umbral obtenido para cada motivo y su longitud, su contenido de información y el contenido de información de su núcleo. Para ello, se plasmaron en tres gráficas los valores de los umbrales ya calculados frente a estas tres propiedades (Figura 5.6). Sin embargo, tal y como se puede observar en la Figura 5.6, no aparece ninguna tendencia evidente del umbral en función de las características consideradas. Esto significa, que no se disponía de una guía clara que permitiera inferir los umbrales de los motivos sin TFBSs definidos. Por tanto, se calculó el umbral

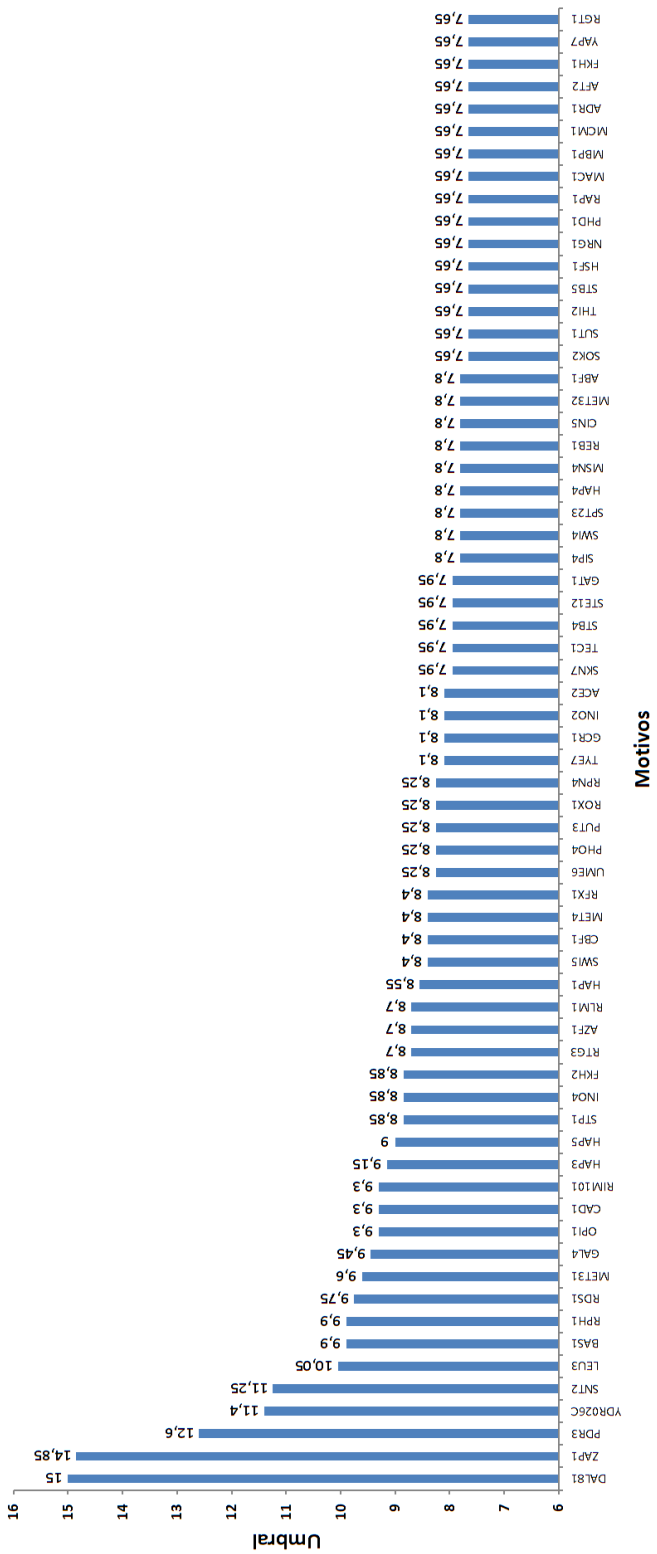


Figura 5.5: Umbrales seleccionados para cada motivo

de estos motivos como la mediana de los umbrales ya calculados, que en este caso resultó ser 8,1.

Con estos valores seleccionados, Patser detectó 77939 sitios de unión potenciales, entre los que se encuentran 1412 de los descritos por Harbison et al. Estos 1412 representan aproximadamente el 50 % del total de TFBSs descritos por estos autores. Es necesario destacar que, para un número considerable de los motivos (ARR1, ASH1, BAS1 entre otros), no era posible detectar mediante Patser ninguno de sus TFBSs conocidos, ni siquiera fijando el umbral a 0. Asimismo, aparecieron otros motivos para los que era necesario situar el umbral en valores extremadamente bajos (1,65, 2,25, 0,3 para ABF1, ADR1 y AFT2 respectivamente, entre otros) con el objetivo de capturar sus TFBSs. Todo esto sugiere la existencia de cierta incoherencia entre los datos obtenidos de Jaspar y los proporcionados por Harbison et al., lo cual podría tener su origen en outliers en los sitios de unión o incluso por incoherencias en la propia nomenclatura de los motivos. Es necesario, por tanto, llevar a cabo un estudio adicional de estos casos para determinar el origen de estas incoherencias.

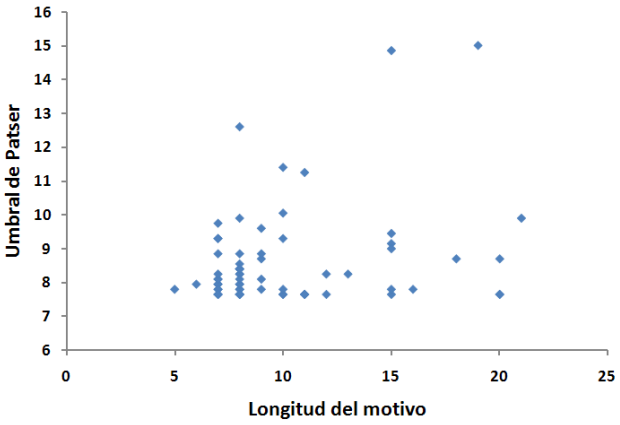
Una vez fijados los parámetros comentados, se procedió a generar la tabla transaccional difusa, obteniéndose así 8178 transacciones, con una media de 9,84 TFs diferentes por transacción y un máximo de 45 TFs. La Figura 5.7 muestra la frecuencia de aparición de los 158 TFs que quedaron capturados en alguna transacción.

Itemsets obtenidos

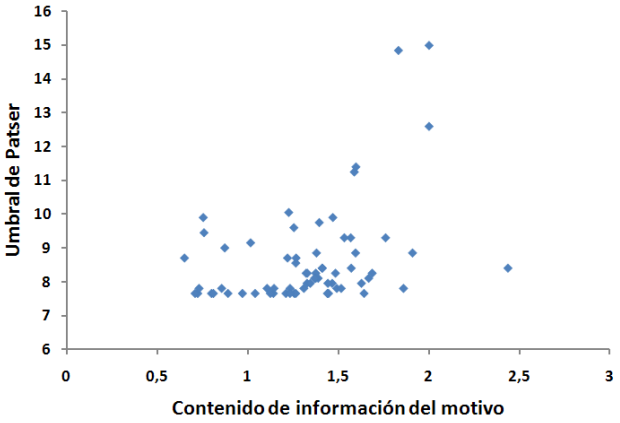
Los umbrales de soporte y p-valor se fijaron nuevamente en 0,01, obteniéndose en este caso 250 itemsets. Los p-valores calculados indican que muy probablemente muchas de las combinaciones obtenidas representen asociaciones biológicas reales entre los TFs. Además, el tamaño de las combinaciones obtenidas oscila entre 2 y 4 TFBSs, lo que de hecho coincide con los valores estimados previamente por otros autores [212, 34].

En primer lugar, se compararon los resultados obtenidos con los del apartado anterior. La comparación directa permitió observar que tan sólo el item-

(1)



(2)



(3)

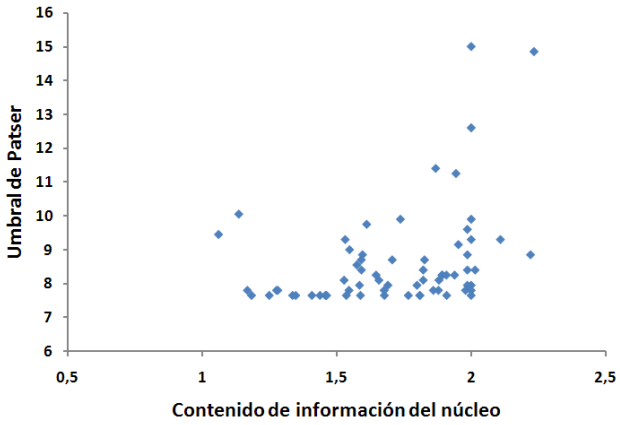


Figura 5.6: Umbrales obtenidos frente a (1) la longitud del motivo, (2) su CI y (3) el CI de su núcleo.

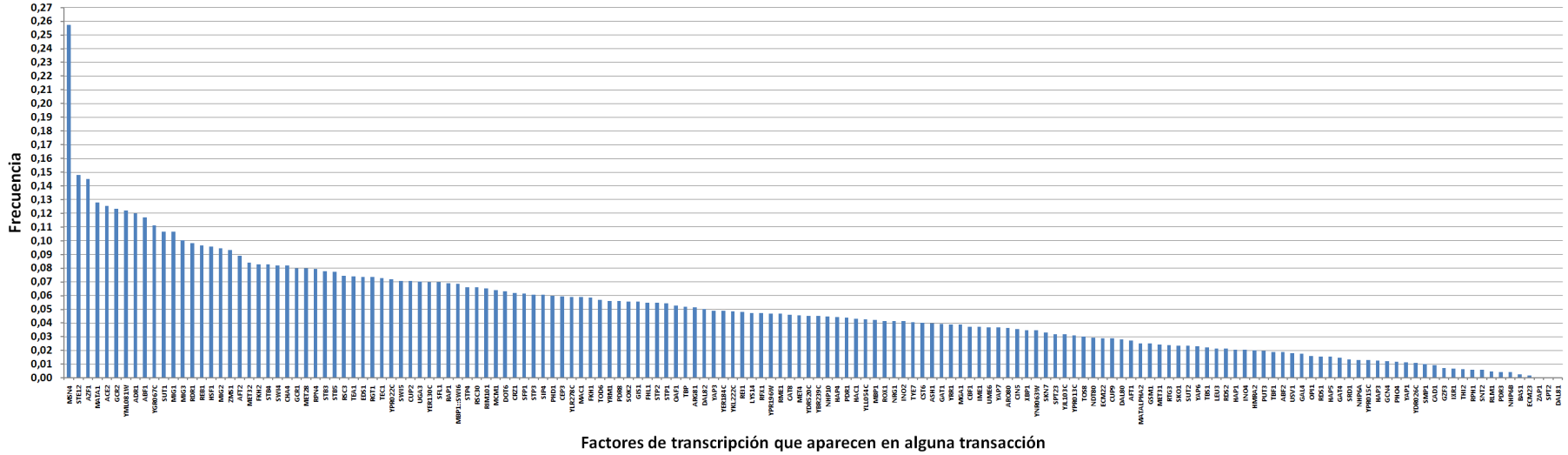


Figura 5.7: Frecuencia de aparición de los TFs en las transacciones obtenidas mediante Patser

set $\{STE12, TEC1\}$ se mantenía en un principio, por lo que se decidió comprobar si el umbral del soporte estaba determinando que no aparecieran el resto de itemsets. Efectivamente, disminuyendo el umbral del soporte se pudo comprobar que hasta 13 de los itemsets anteriores permanecían siendo significativos. Además, se observó que los TFs DIG1, NDD1, SWI6, RCS1 y STB1 no aparecían en Jaspar, por lo que se justifica así la no aparición de hasta 14 de las combinaciones anteriores. Finalmente, en otros casos los itemsets dejan de ser significativos. Esto ocurre por ejemplo con la combinación $\{SKN7, SOK2, PHD1\}$. En este caso el itemset deja de ser significativo, aunque es interesante destacar que dos de sus subitemsets se mantienen: $\{SOK2, SKN7\}$, con soporte 0,0046, p-valor $3{,}671e-07$ y $\{SOK2, PHD1\}$, con soporte 0,0028 y p-valor $8{,}129e-07$.

Además, se utilizó nuevamente la herramienta STRING, que en este caso devolvió grafos para 19 de los itemsets obtenidos, 13 de los cuales representan relaciones directas entre los TFs correspondientes. La Figura 5.8 y la Tabla 5.2 muestran algunos de estos itemsets. Por ejemplo, los TFs del itemset $\{STE12, TEC1\}$ mencionado anteriormente aparecen fuertemente relacionados (Tabla 5.2). Como ya se ha comentado (Sección 5.3.1), las proteínas STE12 y TEC1 cooperan para regular diversos procesos celulares [64, 162, 145].

Por otra parte, el factor de transcripción ADR1 aparece combinado con RAP1, RGT1 y SIP4 (itemsets 2 a 4). Distintas fuentes bibliográficas confirman estas tres asociaciones. En el caso de ADR1 y RAP1, entre otros reguladores transcripcionales, pueden actuar para definir dominios de expresión génica. Así, se ha demostrado que ciertos sitios de unión de estos factores llevan a cabo una función de "barrera", impidiendo la propagación de ciertas señales de silenciación en la levadura [289]. Asimismo, la relación entre ADR1 y RGT1 (YKL038W) se encontró previamente descrita en la bibliografía. Estos dos factores regulan la transcripción de ciertos genes de respuesta a perturbaciones en las fuentes de carbono [209]. Finalmente, tanto ADR1 como SIP4 forman parte de CRMs involucrados en el control transcripcional del metabolismo no-fermentativo de la levadura [220].

El siguiente itemset (5) que aparece en la Tabla 5.2 contiene los TFs AFT2

Tabla 5.2: Combinaciones de TFs obtenidas con Patser

ID Itemset	TFs	p-valor	Soporte
1	$STE12, TEC1$	$7{,}713e-03$	0,013
2	$ADR1, RAP1$	$4{,}625e-08$	0,014
3	$ADR1, RGT1$	$5{,}907e-05$	0,013
4	$ADR1, SIP4$	$1{,}109e-04$	0,011
5	$AFT2, RAP1$	$8{,}527e-14$	0,012
6	$MIG3, RGT1$	$3{,}087e-06$	0,012
7	$MSN4, RPN4$	$2{,}837e-03$	0,024
8	$MSN4, SKN7$	$1{,}391e-05$	0,013
9	$MSN4, GIS1$	$1{,}110e-16$	0,020
10	$MIG1, MIG2$	$1{,}110e-16$	0,016
11	$STE12, GCR2, STB5, XBP1$	$1{,}110e-16$	0,010
12	$ADR1, MIG1$	$1{,}110e-16$	0,020
13	$SUT1, MIG1$	$1{,}110e-16$	0,018

y RAP1. Se sabe que ambos factores regulan la transcripción del FRE1, cuya expresión se ve inducida con bajos niveles de hierro y cobre [6]. En la siguiente combinación de la Tabla 5.2 (itemset 6) aparece de nuevo el factor RGT1, en este caso combinado con MIG3. Hazbun et al. [115] demostraron experimentalmente que estos dos TFs se unen a la región promotora del gen SUC2, cuyo producto es una enzima involucrada en la transformación de sacarosa en glucosa y fructosa. Seguidamente se muestran tres itemsets que relacionan MSN4 con RPN4, SKN7 y GIS1 (itemsets 7 a 9), asociaciones confirmadas por la bibliografía. Así, se ha descrito que MSN4, RPN4 y SKN7 controlan la expresión de genes que se desregulan como respuesta a la presencia de ciertos herbicidas y fungicidas [247, 218]. Respecto a la relación de MSN4 con GIS1, STRING devolvió asociaciones tanto a nivel experimental, como de bases de datos y obtenidas de minería de textos. Asimismo, en la bibliografía se describe a estos factores de transcripción como mediadores del regulón Rim-15 [47]. Finalmente, los TFs MIG1 y MIG2 (itemset 10) se muestran también fuertemente relacionados a distintos niveles. Ambas proteínas se requieren para la represión por glucosa de la expresión del gen SUC2 [165]. Al final de la Tabla 5.2 (itemsets 11 a 13) se muestran algunas

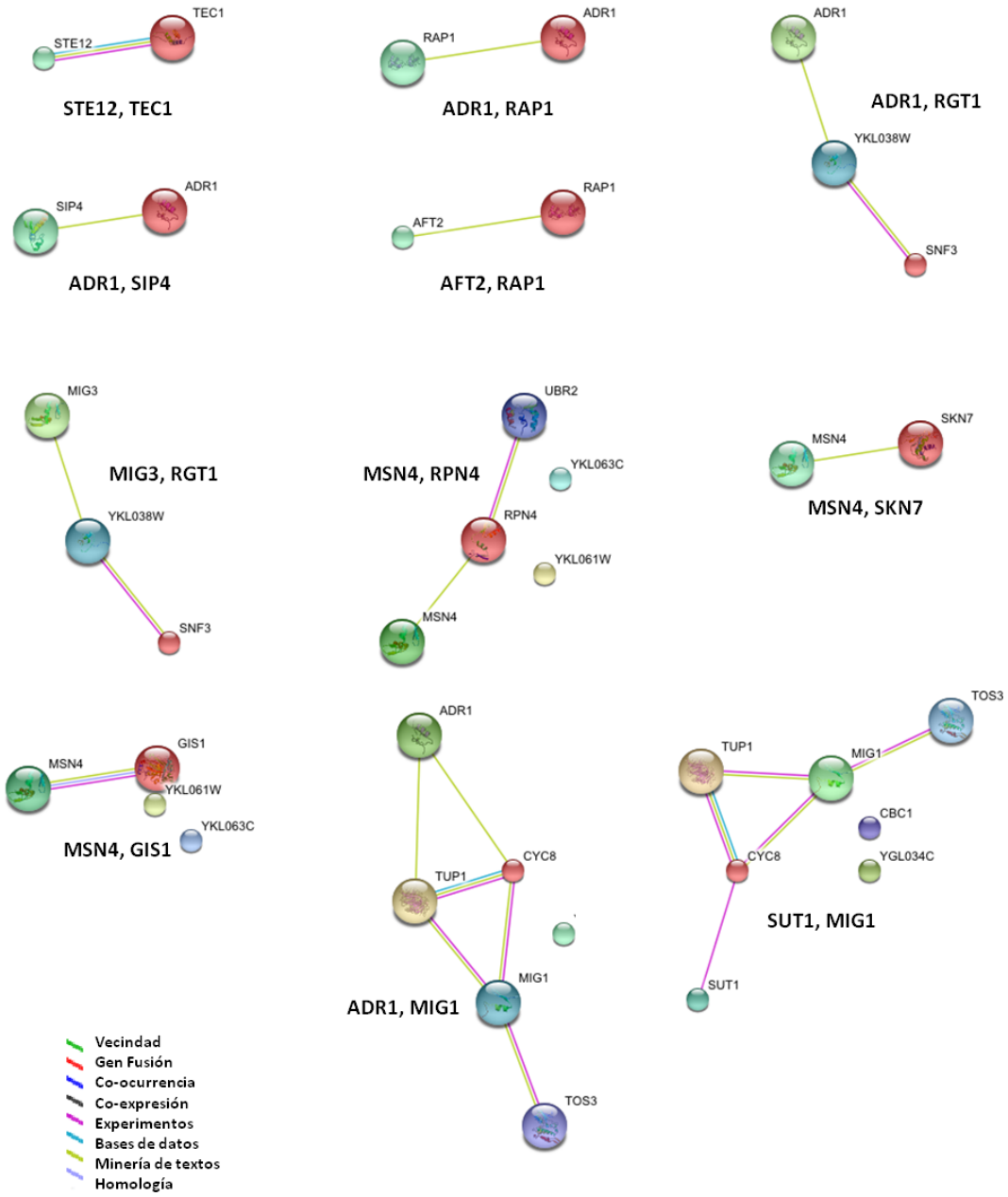


Figura 5.8: Grafos de las combinaciones obtenidas mediante Patser.

combinaciones para las que se obtuvieron relaciones indirectas, y otras para las que no se obtuvo confirmación.

Para finalizar esta sección, merece la pena comentar que se repitió todo el proceso fijando en 8 el umbral global mínimo de Patser (Sección 5.3.2). De esta forma se obtuvieron 56 itemsets. Es interesante el hecho de que este nuevo conjunto de itemsets era un subconjunto del obtenido en este apartado. Es más, los *p*-valores se mantenían prácticamente invariables con respecto a los obtenidos situando el umbral en 7,5. Este hecho es un indicador de coherencia y robustez del método, ya que si los resultados contuvieran una componente aleatoria significativa, podría esperarse que se hubieran obtenido nuevas combinaciones no consideradas anteriormente.

5.3.3 Combinación de los TFBSs de Harbison et al. y los obtenidos usando Patser

En este último apartado se incluyen los resultados otenidos al combinar los TFBSs proporcionados por Harbison et al. con los detectados utilizando Patser. De esta forma, se obtuvieron 8199 transacciones, con una media de 8.79 TFs diferentes por transacción y un máximo de 46 TFs. Al igual que en los dos casos anteriores, los umbrales de soporte y *p*-valor se fijaron en 0,01, obteniéndose así 256 itemsets.

En primer lugar se compararon los resultados obtenidos con los del apartado anterior. Se comprobó que 248 de las combinaciones coincidían con las obtenidas previamente, algo esperado ya que la gran mayoría de los TFBSs provienen de los detectados por Patser. A continuación se compararon los resultados con los de la Sección 5.3.1. En este caso 3 de los itemsets se mantenían, aún fijando el umbral del soporte en 0,01. Al disminuir dicho umbral para permitir la aparición de itemsets con menor soporte, se comprobó que 33 de los obtenidos en la Sección 5.3.1 continuaban siendo significativos.

Este conjunto de resultados supone una mejoría de los mismos como consecuencia de la introducción de los TFBSs de Harbison et al. Así, resulta que prácticamente todas las combinaciones del apartado anterior se mantienen, a la vez que se capturan mejor los itemsets obtenidos en la Sección 5.3.1.

Tabla 5.3: Itemsets obtenidos a partir de Patser y los datos de Harbison et al.

ID Itemset	TFs	p-valor	Soporte
1	$SUT1, REI1$	$1{,}644e-09$	0,010
2	$MIG3, STB4$	$7{,}273e-03$	0,011
3	$SUT1, HSF1$	$7{,}578e-03$	0,014
4	$MSN4, CHA4$	$8{,}535e-03$	0,024
5	$SUT1, SWI4$	$8{,}750e-03$	0,010

Asimismo, aparecen algunas combinaciones nuevas que podrían resultar interesantes (Tabla 5.3). La búsqueda de asociaciones en dichas combinaciones mediante STRING no devolvió ningún resultado, por lo que es necesario un estudio más detallado de las mismas para confirmar tales relaciones.

5.4 Conclusiones

En este estudio se ha presentado una nueva metodología difusa para estudiar co-ocurrencias significativas de TFBSs en un genoma. Esta metodología solventa algunas de las limitaciones de las propuestas previas. Con el objetivo de validar su utilidad se aplicó sobre el genoma de la levadura. Así, se obtuvieron toda una serie de combinaciones de factores de transcripción, un amplio número de ellas en concordancia con la información obtenida de bibliografía especializada, resultados de metodologías previas y datos proporcionados por STRING. Todo esto, junto con los bajos p-valores obtenidos, apoyan el buen funcionamiento de la metodología.

La capacidad de las herramientas de detección de TFBSs es limitada debido a la gran complejidad del problema. Asimismo, la aplicación de las técnicas más avanzadas de detección de TFBSs, se ve restringida por la falta de información en las bases de datos correspondientes. El buen funcionamiento de la técnica de detección empleada es fundamental para la obtención de CRMs fiables, por lo que los avances que se produzcan en estas técnicas mejoraran el rendimiento de la metodología. En este sentido, se piensa que la introducción de información adicional en el proceso de detección, tal como

datos estructurales de las proteínas o de la estructura de cromatina, podrían ayudar a determinar con mayor fiabilidad los posibles TFBSs.

CAPÍTULO 6

Conclusiones y trabajo futuro

6.1 Conclusiones

Este Capítulo resume las principales contribuciones de esta tesis, analizando los resultados de acuerdo con los objetivos iniciales.

En el Capítulo 3 se describe una metodología para la integración y el análisis de información biológica heterogénea. El principal aspecto de esta metodología consiste en el uso de una versión difusa del algoritmo de extracción de reglas de asociación TD-FP-Growth. La implementación de dicho algoritmo se integró en una aplicación web: BioFAR. Mediante esta aplicación se pretende extender el uso de las técnicas difusas en Bioinformática, proporcionando una herramienta que permite el análisis de grandes conjuntos de datos heterogéneos. La aplicación resultó ser muy útil en el desarrollo de este trabajo.

La metodología se aplicó posteriormente a un conjunto de datos constituido por una amplia variedad de características estructurales y funcionales del genoma de la levadura. Los resultados muestran asociaciones interesantes, muchas de ellas en concordancia con trabajos experimentales previos en este área. De esta forma se demuestra que la metodología es útil para obtener asociaciones relevantes y fiables, las cuales pueden servir como hipótesis para su posterior validación en el laboratorio. Asimismo, las reglas

de asociación difusas han mostrado ser una herramienta intuitiva para describir relaciones biológicas utilizando etiquetas lingüísticas y unos pocos parámetros fácilmente interpretables (soporte, confianza y factor de certeza). Esto las hace especialmente útiles en Bioinformática, dado que los expertos (biólogos, médicos, etc.) no suelen disponer de conocimientos informáticos avanzados y deben estudiar y validar los resultados.

Los datos evidencian la importancia de utilizar técnicas que modelen los límites de una manera más realista a como lo hacen las técnicas clásicas crisp. La definición apropiada de los conceptos introducidos en el análisis es crucial, ya que determina la interpretación que se hará de los resultados. Asimismo, la presencia de ruido e imprecisión en los datos biológicos hace más necesario aún utilizar definiciones difusas de estos conceptos y aplicar sobre ellos técnicas de *Soft Computing*.

En el Capítulo 4 se presenta un trabajo multidisciplinar para descubrir potenciales relaciones entre datos de expresión y factores de pronóstico en el cáncer de mama. La heterogeneidad e imprecisión de los datos, así como sus grandes dimensiones, hacen que la extracción de reglas de asociación difusas sea una herramienta apropiada para el análisis.

Se obtuvieron una serie de asociaciones, entre las reglas obtenidas se muestran muchas tendencias ya descritas en la bibliografía. Asimismo, aparecen otras que, aún no encontrándose una mención explícita, cobran sentido de acuerdo con las publicaciones consultadas. Finalmente, se muestran también otras relaciones de las que no se dispone evidencia bibliográfica. Los resultados obtenidos permiten, por tanto, sugerir que algunas estas relaciones podrían guiar a los expertos hacia un mejor entendimiento de esta enfermedad. De hecho, actualmente se están realizando pruebas de PCR e inmunostoquímica sobre algunos genes resultantes de las reglas de asociación obtenidas, con objeto de descubrir posibles marcadores tumorales.

En el Capítulo 5 se presenta una nueva metodología difusa para estudiar co-ocurrencias significativas de TFBSs en un genoma. Esta metodología solventa algunas de las limitaciones de las propuestas previas. Algunas de sus propiedades más interesantes de son:

- Tan sólo requiere como entrada la secuencia del genoma, una librería de PWMs y los umbrales de soporte y *p*-valor.
- Es capaz de analizar un genoma completo sin necesidad de restringir la búsqueda a regiones concretas.
- El uso de conjuntos difusos para modelar los módulos de regulación representa más adecuadamente la realidad.
- No impone ningún tipo de restricción en el tipo de CRMs que se buscan.
- Cada CRM obtenido se expresa mediante un conjunto de TFs, su frecuencia de aparición en el genoma y un *p*-valor, lo que hace que los resultados sean fácilmente interpretables.

Con el objetivo de validar su utilidad se aplicó sobre el genoma de la levadura. Así, se obtuvieron toda una serie de combinaciones de factores de transcripción, de las cuales un amplio número de ellas está en concordancia con la información obtenida de bibliografía especializada, con los resultados de metodologías previas, así como con los datos proporcionados por STRING. Todo esto, junto con los bajos *p*-valores obtenidos, apoyan el buen funcionamiento de la metodología.

La capacidad de las herramientas de detección de TFBSs es limitada debido a la gran complejidad del problema. Asimismo, la aplicación de las técnicas más avanzadas de detección de TFBSs, se ve restringida por la falta de información en las bases de datos correspondientes. El buen funcionamiento de la técnica de detección empleada es fundamental para la obtención de CRMs fiables, por lo que los avances que se produzcan en estas técnicas mejoraran el rendimiento de la metodología. En este sentido, se piensa que la introducción de información adicional en el proceso de detección, tal como datos estructurales de las proteínas o de la estructura de cromatina, podrían ayudar a determinar con mayor fiabilidad los posibles TFBSs.

Finalmente, es necesario destacar que el objetivo de este trabajo no consiste en descubrir y establecer el papel de todos los genes mostrados en los resultados, dado que no se dispone de evidencia experimental de estas relaciones. Por el contrario, se pretende mostrar patrones obtenidos con estas

técnicas que pueden guiar a los científicos del área correspondiente (médicos, biólogos, etc.), en sus investigaciones, proporcionando relaciones que les permitan enunciar hipótesis para su posterior validación experimental.

6.2 Trabajos futuros

Hay aún mucho trabajo por llevar a cabo referente al desarrollo y aplicación de técnicas difusas en Bioinformática. En este caso particular, estamos especialmente interesados en el desarrollo de nuevas estrategias de filtrado de reglas que permitan mejorar la fiablidad de los resultados. En este sentido, el desarrollo de estrategias de filtrado específicas para este dominio es probablemente la opción más apropiada, dada la gran cantidad de información disponible que podría incorporarse al proceso de búsqueda, lo que llevaría a un enfoque semisupervisado. Obviamente, las mejoras que se alcancen en la metodología serán integradas en BioFAR.

Respecto al estudio llevado a cabo sobre el genoma de la levadura, los trabajos futuros comprenderán la inclusión de atributos nuevos en el análisis. Es más, será especialmente interesante aplicar la metodología sobre genomas de otras especies y comparar los resultados. En este caso, la comparación se verá limitada por la disponibilidad de la información de otras especies y por la complejidad de los genomas correspondientes.

Además, se derivan diferentes líneas de investigación del estudio sobre el cáncer de mama. En este sentido, es altamente recomendable incluir más pacientes en el análisis, dado que esto ayudaría a capturar mejor la realidad de la población y a eliminar patrones espúrios. Asimismo, sería interesante aplicar otras técnicas de normalización de microarrays y comparar los resultados, lo que incrementaría la calidad y fiabilidad de los datos obtenidos. Además, la aplicación de otras técnicas computacionales, como por ejemplo algoritmos de biclústering, podría ayudar a mejorar los resultados.

Finalmente, surgen diferentes trabajos de la parte del estudio de los CRMs. El objetivo inmediato consiste en mejorar el proceso de detección de TFBSs, para lo que se podrían aplicar metodologías ya desarrolladas por el grupo

que capturan las dependencias interposicionales. Como ya se comentó anteriormente, su aplicación se ve limitada por falta de información en las bases de datos, lo que hace necesario investigar nuevas vías de obtención de dicha información. Posteriormente, la metodología deberá aplicarse sobre otros genomas. En este sentido, sería razonable incrementar paulatinamente la complejidad de los genomas estudiados.

6.3 Publicaciones

Relacionados con el desarrollo de esta memoria se han publicado los siguientes trabajos como autor principal:

- F J Lopez, A Blanco, F Garcia, C Cano, and A Marin. Fuzzy Association Rules for Biological Data Analysis: A Case Study on Yeast. *BMC Bioinformatics*, 9:107-124, 2008.
- F J Lopez, A Blanco, M Cuadros, and A Concha. Analysis of Breast Cancer Genomic Data by Fuzzy Association Rule Mining. Database Technology for Life Sciences and Medicine. World Scientific, London, UK, 2010. *In press*.
- F J Lopez, A Blanco, F Garcia, and C Cano. Extracting Biological Knowledge by Association Rule Mining. Data Mining in Biomedicine Using Ontologies. 2009. Artech House, Boston, MA, USA, 133-159.
- F J Lopez, A Blanco, F Garcia, and A Marin. Extracting Biological Knowledge by Fuzzy Association Rule Mining. In *Proceedings of the IEEE International Conference on Fuzzy Systems FUZZIEEE 2007: London, UK*, 583-588, 2007.
- F J Lopez, M Cuadros, A Blanco, and A Concha. Unveiling Fuzzy Associations Between Breast Cancer Prognostic Factors and Gene Expression Data. In *Proceedings of the Database and Expert Systems Applications DEXA 2009: Linz, Austria*, 338-342, 2009.
- F J Lopez, C Cano, F Garcia, and A Blanco. A Fuzzy Approach for Studying Combinatorial Regulatory Actions of Transcription Factors in

Yeast. In *Proceedings of the 10th International Conference on Data Engineering and Automated Learning IDEAL 2009: Burgos, Spain*, 477-484, 2009.

- F J Lopez, A Blanco, F Garcia, S Blanco. Aplicación de las Reglas de Asociación Difusas en Proteómica. In *Proceedings of the XIV Congreso Español Sobre Tecnologias y Logica Fuzzy ESTYLF 2008, Langreo-Mieres, España*, 599-604, 2008.

Relacionados con los temas tratados en esta memoria también se ha colaborado en las siguientes publicaciones:

- M Cuadros, P Talavera, **F J Lopez**, I Garcia-Perez, A Blanco, A Concha. Real time RT-PCR analysis for evaluating Her2/neu status in breast cancer. *Pathobiology*. 77(1), *In press*.
- F Garcia, **F J Lopez**, C Cano and A Blanco. FISim: a New Similarity Measure Between Transcription Factor Binding Sites Based on the Fuzzy Integral. *BMC Bioinformatics*. 10(1): 224-236, 2009.
- C Cano, F Garcia, **F J Lopez**, A Blanco. Intelligent System for the Analysis of Microarray Data Using Principal Components and Estimation of Distribution Algorithms. *Expert Systems with Applications*. 36(3): 4654-4663, 2009.
- C Cano, L Adarve, **F J Lopez**, and A Blanco. Possibilistic Approach for Biclustering Microarray Data. *Computers in Biology and Medicine*. 37(10): 1426-1436, 2007.
- C Cano, **F J Lopez**, F Garcia, and A Blanco. Evolutionary Algorithms for Finding Interpretable Patterns in Gene Expression Data. *IADIS International Journal on Computer Science and Information Systems*. 1(2):88-99, 2006.
- F Garcia, **F J Lopez**, C Cano, A Blanco. Study of Fuzzy Resemblance Measures for DNA Motifs. In *Proceedings of the IEEE International Conference on Fuzzy Systems FUZZIEEE 2009: Jeju Island, Korea*, 1175-1180, 2009.

- F Garcia, L Adarve, **F J Lopez**, and A Blanco. Assessment of Gene Ontology Based Recognition of Related Proteins. In *Proceedings of the International Conference on Applied Computing IADIS 2008: Algarve, Portugal*, 179-186, 2008.
- C Cano, F Garcia, **F J Lopez**, L Adarve, and A Blanco. Non-supervised Identification of Gene Regulatory Modules by Possibilistic Biclustering of Microarray Data. In *Proceedings of the 11th International Conference on Cognitive and Neural Systems: Boston, MA, USA*, 2007.
- S Blanco, A Blanco, C Cano, **F J Lopez**. SCEPGG: A Method for Biclustering Microarray Data. In *Proceedings of the International Conference on Applied Computing IADIS 2007, Algarve, Portugal*, 493-498, 2007.
- R Caliz, J Sainz, L Hassan, P Hernandez, A Blanco, F Garcia, **F J Lopez**, J Salvatierra, et al. Differences in Gene Expression Profile Between Early and Developed Rheutaoid Arthritis. In *Proceedings of the Annals of the Rheumatic Diseases, Amsterdam, The Netherlands*, 623-623, 2006.
- J Salvatierra, J Sainz, M D Collado, L Hassan, A Blanco, F Garcia, M Ferrer, **F J Lopez**, A Garcia, et al. Peripheral blood White Cells Gene Expression Profile in Early Rheumatoid Arthritis Patients by CodeLink Human Whole Genome Bioarrays. In *Proceedings of the Annals of Rheumatic Diseases, Amsterdam, the Netherlands*, 621-621, 2006.
- M Guzman, M D Collado, J Sainz, A Blanco, L Hassan, M Ferrer, F Garcia, **F J Lopez**, P Henandez, et al. Gene Expression Characterization of Peripheral Blood White Cells from Patients with Developed Rheumatoid Arthritis by Microarrays. In *Proceedins of the Annals of Rheumatic Diseases, Amsterdam, The Netherlands*, 621-621, 2006.
- F Garcia, **F J Lopez**, C Cano and A Blanco. An Ontology-Driven Similarity Providing Reliable Protein Family Recognition. In *Proceedings of the International Conference on Applied Computing IADIS 2006, San Sebastian, Spain*, 649-654, 2006.

CHAPTER 6

Conclusions and future work

6.1 Conclusions

This chapter summarizes the contributions of this thesis, analyzing the results with regards to the initial objectives.

A novel fuzzy methodology for the integration and analysis of heterogenous biological data is described in Chapter 3. The main aspect of this fuzzy methodology is a novel fuzzy association rule mining algorithm, the Fuzzy-TD-FP-Growth method. A web application was developed which runs the algorithm: BioFAR. This software aims to spread fuzzy techniques by providing a tool which helps to analyze heterogeneous and high-dimensional data. The web application has been shown to be very useful in the development of this work.

The methodology was then applied over a dataset comprising a variety of structural and functional features of the yeast genome. The results show relevant associations, many of them in agreement with previous works in this field. It demonstrates that significant biological insights can be obtained by using the proposed methodology. It also proves fuzzy association rules to be an intuitive tool to describe biological relationships by using linguistic labels and few easy-understandable parameters (support, confidence and certainty factor). Moreover, fuzzy rules are easy to understand since they are very similar to the way a person might express knowledge. This makes them

especially suitable for their application in this field in which experts must validate the results.

The data also show the importance of using techniques that can model borders in a more realistic way than classical crisp techniques do. The appropriate definition of the concepts introduced in the analysis is crucial since it determines the interpretation that one may obtain from the resultant rule set. In addition, the presence of noise in biological data makes even more necessary to use a fuzzy definition of these concepts.

An integrative and multi-disciplinary approach is presented in Chapter 4 to study potential relations between prognostic factors and whole-genome expression data in breast cancer. The heterogeneity and imprecise nature of the data along with its high dimensionality make fuzzy association rules an appropriate tool for the analysis.

A number of interesting associations have been found, many of these representing previously described trends. Likewise, some other rules have appeared which are not directly reported in the literature, but that make sense according to previously published results. Finally, a set of relations has also been obtained for which no bibliographic evidence was found. These results suggest that some of these patterns may help experts to achieve a better understanding of breast cancer. In fact, some of the obtained associations are currently being validated at the *Hospital Universitario Virgen de las Nieves* by means of PCR and IHC analysis.

Chapter 5 is devoted to a new fuzzy approach to study significant co-occurrences of closely located TFBSs in the yeast whole-genome. This methodology overcomes some of the limitations of previously proposed approaches. Some interesting properties of the proposed methodology are:

- It only requires the genome sequence, a library of PWMs and a p-value and support threshold.
- It is able of analyzing a complete genome without limiting the search procedure to specific regions.
- The use fuzzy sets allows the methodology to properly model CRMs.
- It does not impose constraints on the form of the discovered CRMs.

- Each CRM is expressed as a set of TFs, its frequency of appearance in the genome and a *p*-value, which makes the results to be easily interpretable.

In order to validate the methodology experiments were carried out over the yeast genome. A number of interesting TF combinations have been found, many of them in agreement with the literature, with the results obtained by previous approaches and with the data provided by STRING. These facts, together with the very low p-values of the reported TF combinations, support the good performance of the methodology.

The performance of TFBSs detection techniques is quite limited by the great complexity of the problem. Likewise, the use of the most advanced techniques is constrained by the lack of available information. A good performance of the TFBS detection tool is critical in order to obtain reliable CRMs. Thus, enhancements of TFBSs detection tools will clearly improve the performance of the methodology. In this respect, we believe that the integration of additional information (e.g. chromatin structure data, protein structure data, etc.) in the detection process may help to increase the reliability of the inferred TFBSs.

Finally, it is worth mentioning that it was not the aim of this work to give a comprehensive list and biological interpretation of all of the obtained biological patterns, but to show that interesting associations can be obtained following the proposed methodologies. These associations may serve to enunciate hypothesis for subsequent empirical evaluation. Therefore, a deep study and experimental evidences of these patterns are needed to confirm such associations.

6.2 Future work

There is much room for improvement regarding the development and application of fuzzy techniques, and particularly, fuzzy association rule mining techniques for the integration and analysis of biological information. In this respect, we are specially concerned with the development of new rule filtering strategies which help to increase the reliability of the results. Probably,

the definition of domain-specific rule filtering strategies is the most appropriate choice in this field, since there is much information available which could be incorporated into the discovery proccess, thus leading to a semi-supervised approach. Obviously, every enhancement of the methodology will be implemented in BioFAR.

Regarding the particular study of the yeast genome, future work comprises the inclusion of new attributes into the analysis. Furthermore, it could be very interesting to apply the methodology over information obtained from genomes of other species and to compare the results. This comparison will be limited by the availability of information from the rest of species and by the complexity of the corresponding genomes.

Several research projects also arise from the breast cancer work. Advances in this study may yield quite relevant results. For example, the inclusion of more patients in the analysis is highly recommendable, since it would help to better capture the reality of the population and to remove spurious patterns. Likewise, the use of some other microarray normalization procedures and the comparison of the results could improve the signification of the results. In addition, the application of additional computational techniques such as biclustering algorithms may help to enhance the result set.

Finally, some future works are derived from the CRMs discovery part. The next task is to improve the TFBSs detection process by applying some of the methodologies developed by our group. These methodologies take into account the positional dependencies and have been recently shown to outperform previous approaches. As previously stated, their application is limited by the available information. Therefore, additional ways of gathering the required information should be investigated. Then, the methodology shall be tested on genomes of other species. Gradually increasing the complexity of the genome is probably the most reasonable procedure.

6.3 Publications

6.3.1 Publications derived from this thesis (main author)

- F J Lopez, A Blanco, F Garcia, C Cano, and A Marin. Fuzzy Association Rules for Biological Data Analysis: A Case Study on Yeast. *BMC Bioinformatics*, 9:107-124, 2008.
- F J Lopez, A Blanco, M Cuadros, and A Concha. Analysis of Breast Cancer Genomic Data by Fuzzy Association Rule Mining. Database Technology for Life Sciences and Medicine. World Scientific, London, UK, 2010. *In press*.
- F J Lopez, A Blanco, F Garcia, and C Cano. Extracting Biological Knowledge by Association Rule Mining. Data Mining in Biomedicine Using Ontologies. 2009. Artech House, Boston, MA, USA, 133-159.
- F J Lopez, A Blanco, F Garcia, and A Marin. Extracting Biological Knowledge by Fuzzy Association Rule Mining. In *Proceedings of the IEEE International Conference on Fuzzy Systems FUZZIEEE 2007: London, UK*, 583-588, 2007.
- F J Lopez, M Cuadros, A Blanco, and A Concha. Unveiling Fuzzy Associations Between Breast Cancer Prognostic Factors and Gene Expression Data. In *Proceedings of the Database and Expert Systems Applications DEXA 2009: Linz, Austria*, 338-342, 2009.
- F J Lopez, C Cano, F Garcia, and A Blanco. A Fuzzy Approach for Studying Combinatorial Regulatory Actions of Transcription Factors in Yeast. In *Proceedings of the 10th International Conference on Data Engineering and Automated Learning IDEAL 2009: Burgos, Spain*, 477-484, 2009.
- F J Lopez, A Blanco, F Garcia, S Blanco. Aplicación de las Reglas de Asociación Difusas en Proteómica. In *Proceedings of the XIV Congreso Español Sobre Tecnologias y Logica Fuzzy ESTYLF 2008, Langreo-Mieres, España*, 599-604, 2008.

6.3.2 Publications related with this thesis (collaborator)

- M Cuadros, P Talavera, **F J Lopez**, I Garcia-Perez, A Blanco, A Concha. Real time RT-PCR analysis for evaluating Her2/neu status in breast cancer. *Pathobiology*. 77(1), *In press*.
- F Garcia, **F J Lopez**, C Cano and A Blanco. FISim: a New Similarity Measure Between Transcription Factor Binding Sites Based on the Fuzzy Integral. *BMC Bioinformatics*. 10(1): 224-236, 2009.
- C Cano, F Garcia, **F J Lopez**, A Blanco. Intelligent System for the Analysis of Microarray Data Using Principal Components and Estimation of Distribution Algorithms. *Expert Systems with Applications*. 36(3): 4654-4663, 2009.
- C Cano, L Adarve, **F J Lopez**, and A Blanco. Possibilistic Approach for Biclustering Microarray Data. *Computers in Biology and Medicine*. 37(10): 1426-1436, 2007.
- C Cano, **F J Lopez**, F Garcia, and A Blanco. Evolutionary Algorithms for Finding Interpretable Patterns in Gene Expression Data. *IADIS International Journal on Computer Science and Information Systems*. 1(2):88-99, 2006.
- F Garcia, **F J Lopez**, C Cano, A Blanco. Study of Fuzzy Resemblance Measures for DNA Motifs. In *Proceedings of the IEEE International Conference on Fuzzy Systems FUZZIEEE 2009: Jeju Island, Korea*, 1175-1180, 2009.
- F Garcia, L Adarve, **F J Lopez**, and A Blanco. Assessment of Gene Ontology Based Recognition of Related Proteins. In *Proceedings of the International Conference on Applied Computing IADIS 2008: Algarve, Portugal*, 179-186, 2008.
- C Cano, F Garcia, **F J Lopez**, L Adarve, and A Blanco. Non-supervised Identification of Gene Regulatory Modules by Possibilistic Biclustering of Microarray Data. In *Proceedings of the 11th International Conference on Cognitive and Neural Systems: Boston, MA, USA*, 2007.

- S Blanco, A Blanco, C Cano, **F J Lopez**. SCEPGG: A Method for Biclustering Microarray Data. In *Proceedings of the International Conference on Applied Computing IADIS 2007, Algarve, Portugal*, 493-498, 2007.
- R Caliz, J Sainz, L Hassan, P Hernandez, A Blanco, F Garcia, **F J Lopez**, J Salvatierra, et al. Differences in Gene Expression Profile Between Early and Developed Rheutaoid Arthritis. In *Proceedings of the Annals of the Rheumatic Diseases, Amsterdam, The Netherlands*, 623-623, 2006.
- J Salvatierra, J Sainz, M D Collado, L Hassan, A Blanco, F Garcia, M Ferrer, **F J Lopez**, A Garcia, et al. Peripheral blood White Cells Gene Expression Profile in Early Rheumatoid Arthritis Patients by CodeLink Human Whole Genome Bioarrays. In *Proceedings of the Annals of Rheumatic Diseases, Amsterdam, the Netherlands*, 621-621, 2006.
- M Guzman, M D Collado, J Sainz, A Blanco, L Hassan, M Ferrer, F Garcia, **F J Lopez**, P Henandez, et al. Gene Expression Characterization of Peripheral Blood White Cells from Patients with Developed Rheumatoid Arthritis by Microarrays. In *Proceedins of the Annals of Rheumatic Diseases, Amsterdam, The Netherlands*, 621-621, 2006.
- F Garcia, **F J Lopez**, C Cano and A Blanco. An Ontology-Driven Similarity Providing Reliable Protein Family Recognition. In *Proceedings of the International Conference on Applied Computing IADIS 2006, San Sebastian, Spain*, 649-654, 2006.

Bibliografía

[1] Affymetrix statistical algorithms descripcion document. http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf.

[2] The comprehensive yeast genome database. http://mips.gsf.de/genre/proj/yeast.

[3] The gene ontology. http://www.geneontology.org.

[4] Mesh: The medical subject headings thesaurus. http://www.nlm.nih.gov/mesh/.

[5] Rsa-tools-patser. http://rsat.ulb.ac.be/rsat/patser_form.cgi.

[6] The saccharomyces genome database. http://www.yeastgenome.org.

[7] S Aerts, P van Loo, G Thijs, Y Moreau, and B de Moor. Computational detection of cis-regulatory modules. *Bioinformatics-Oxford*, 19(2):5–14, 2003.

[8] R Agrawal, T Imielinski, and A Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD INTL Conf. on Management of Data (ACM SIGMOD 93); Washington, USA*, pages 207–216, 1993.

[9] J S Aguilar-Ruiz and F Divina. Biclustering of expression data with evolutionary computation. *IEEE Trans. Knowledge and Data Engineering*, 18(5):590–602, 200.

[10] F Al-Shahrour, P Minguez, J Tarraga, I Medina, E Alloza, D Montaner, and J Dopazo. Fatigo+: a functional profiling tool for genomic data. integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Research*, 35(Web Server issue):91–96, 2007.

[11] B Alberts, A Johnson, J Lewis, M Raff, K Roberts, and P Walter. *Molecular Biology of The Cell*. Garland Science, New York, NY, USA, 2002.

[12] A R Albig, J R Neil, and W P Schiemann. Fibulins 3 and 5 antagonize tumor angiogenesis in vivo. *Cancer research*, 66(5):2621–2629, 2006.

[13] R Alcala, J Alcala-Fdez, M J Gacto, and F Herrera. Genetic learning of membership functions for mining fuzzy association rules. In *Proceedings of the IEEE International Fuzzy Systems Conference (FUZZIEEE'07); London, UK*, pages 1538–1543, 2007.

[14] D C Allred, G M Clark, R Elledge, S A W Fuqua, R W Brown, G C Chamness, C K Osborne, and W L McGuire. Association of p53 protein expression with tumor cell proliferation rate and clinical outcome in node-negative breast cancer. *JNCI Journal of the National Cancer Institute*, 85(3):200–206, 1993.

[15] D C Allred, J M Harvey, M Berardo, and G M Clark. Prognostic and predictive factors in breast cancer by immunohistochemical analysis. *Modern pathology*, 11(2):155–168, 1998.

[16] G Alterovitz, M Xiang, M Mohan, and M F Ramoni. Go pad: the gene ontology partition database. *Nucleic Acids Research*, 35:D322–D327, 2007.

[17] M I Arnone and E H Davidson. The hardwriting of development: organization and function of genomic regulatory systems. *Development*, 124(10):1851–1864, 1997.

[18] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[19] S F Atschul, T L Madden, A A Schaffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[20] W H Au and K C C Chan. An effective algorithm for discovering fuzzy rules in relational databases. In *Proceedings of the IEEE Int. Conf. on Fuzzy Systems: 1998; Anchorage, Alaska, USA*, pages 1314–1319, 1998.

[21] W H Au and K C C Chan. Farm: A data mining system for discovering fuzzy association rules. In *Proceedings of the FUZZ-IEEE'99: 1999; Seul, South Korea*, pages 22–25, 1999.

[22] F Azuaje and J Dopazo. *Data analysis and visualization in genomics and proteomics*. Wiley, 2005.

[23] T L Bailey and C Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proc Int Conf Intell Syst Mol Biol*, volume 2(1553-0833), pages 28–36, 1994.

[24] Z Bar-Joseph, G K Gerber, T I Lee, N J Rinaldi, J Y Yoo, F Robert, D B Gordon, E Fraenkel, T S Jaakkola, R A Young, et al. Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, 21(11):1337–1342, 2003.

[25] J Barlett, E Mallon, and T Cooke. The clinical evaluation of her-2 status: which test to use? *J Pathol*, 199(4):411–147, 2003.

[26] Y Bastide, R Taouil, P Nicolas, G Stumme, and L Lakhal. Mining frequent patterns with counting inference. *SIGKDD Explorations*, 2(2):66–75, 2000.

[27] G Bebek and J Yankg. Pathfinder: mining signal transduction pathway segments from protein-protein interaction networks. *BMC Bioinformatics*, 8:335–347, 2007.

[28] C Becquet, S Blachon, B Jeudy, J F Boulicaut, and O Gandrillon. Strong-association -rule mining for large-scale gene-expression data analysis: a cas study on human sage data. *Genom Biol*, 3:1–16, 2002.

[29] V Bempt et al. Polysomy 17 in breast cancer: clinicopathologic significance and impact on HER-2 testing. *Journal of Clinical Oncology*, 26(30):4869–4874, 2008.

[30] F Berzal, I Blanco, D Sanchez, and M A Vila. Measuring the accuracy and interest of association rules: A new framework. *Intelligent Data Analysis*, 6:221–235, 2004.

[31] H Bhaskar, D Hoyle, and S Singh. Machine learning in bioinformatics: A brief survey and recommendations for practitioners. *Computers in Biology and Medicine*, 36:1104–1125, 2005.

[32] M Bilous, M Dowsett, W Hanna, J Isola, A Lebeau, A Moreno, F Penault-Llorca, J Rüschoff, G Tomasic, and M van de Vijver. Current perspectives on her2 testing: a review of national testing guidelines. *Mod Pathol*, 16(2):173–182, 2003.

[33] J A Birdsell. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol. Biol. Evol.*, 19:1181–1197, 2002.

[34] M Blanchette, A R Bataille, X Chen, C Poitras, J Laganière, C Lefèbvre, G Deblois, V Giguère, V Ferretti, D 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.

[35] M Blanchette, W J Kent, C Riemer, L Elnitski, A F A Smit, K M Roskin, R Baertsch, K Rosenbloom, H Clawson, E D Green, et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, 14(4):708–715, 2004.

[36] F Bodon. A fast apriori implementation. In *Proceedings of the 1st IEEE ICDM Workshop on Frequent Item Set Mining Implementations (FIMI 2003): 2003; Melbourne, FL, USA*, 2001.

[37] G H P M Bollen, W H Mager, L W Jenneskens, and R J Planta. Small-size mrnas code for ribosomal proteins in yeast. *Eur. J. Biochem*, 105:75–80, 1980.

[38] B M Bolstad, R A Irizarry, M Astrand, and T P Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[39] C Borgelt. Efficient implementations of apriori and eclat. In *Proceedings of the 1st IEEE ICDM Workshop on Frequent Item Set Mining Implementations (FIMI 2003); Melbourne, FL, USA*, 2001.

[40] S Bownds, P Tong-On, S A Rosenberg, and M Parkhurst. Induction of tumor-reactive cytotoxic T-lymphocytes using a peptide from NY-ESO-1 modified at the carboxy-terminus to enhance HLA-A2. 1 binding affinity and stability in solution. *Journal of Immunotherapy*, 24(1):1–9, 2001.

[41] D J Brennan, S Ek, E Doyle, T Drew, M Foley, G Flannelly, D P O'Connor, W M Gallagher, S Kilpinen, O P Kallioniemi, et al. The transcription factor Sox11 is a prognostic factor for improved recurrence-free survival in epithelial ovarian cancer. *European Journal of Cancer*, 45(8):1510–1517, 2009.

[42] S J Brill and R Sternglanz. Transcription-dependent dna supercoiling in yeast dna topoisomerase mutants. *Cell*, 54:403–411, 1988.

[43] R Burcombe, G D Wilson, M Dowsett, I Khan, P I Richman, F Daley, S Detre, and A Makris. Evaluation of Ki-67 proliferation and apoptotic index before, during and after neoadjuvant chemotherapy for primary breast cancer. *Breast Cancer Res*, 8(3):31–33, 2006.

[44] A Bykowski and C Rigotti. A condensed representation to find frequent patterns. In *Proceedings of the 20th ACM SIGMOD-SIGACT-*

SIGART Symp. on the Principles of Database Systems: 2001; Santa Barbara, California, USA, pages 267–273, 2001.

[45] C Bystroff, V Thorsson, and D Baker. Hmmstr: a hidden markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.*, 301:173–190, 2000.

[46] T Calders and B Goethals. Mining all non-derivable frequent itemsets. *Lecture Notes in Computer Science*, 2(431):74–85, 2002.

[47] E Cameroni, N Hulo, J Roosen, J Winderickx, and C de Virgilio. The novel yeast PAS kinase Rim 15 orchestrates G0-associated antioxidant defense mechanisms. *Cell Cycle*, 3(4):462–468, 2004.

[48] C Cano, L Adarve, J Lopez, and A Blanco. Possibilistic approach for biclustering microarray data. *Computers in Biology and Medicine*, 37(10):3710–3715, 2007.

[49] C Cano, F García, F J López, and A Blanco. Intelligent system for the analysis of microarray data using principal components and estimation of distribution algorithms. *Expert Systems with Applications*, 36(3P1):4654–4663, 2009.

[50] S Carbon, A Ireland, C J Mungall, S Shu, B Marshall, S Lewis, and GO Consortium. Amigo: online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289, 2009.

[51] P Carmona-Saez. Genecodis: A web-based tool for finding significant concurrent annotations in gene lists. *Genome Biology*, 8:R3, 2007.

[52] P Carmona-Saez, M Chagoyen, A Rodriguez, O Trelles, J M Carazo, and A Pascual-Montano. Integrated analysis of gene expression by association rules discovery. *BMC Bioinformatics*, 7:54–69, 2006.

[53] P Carter, L Presta, C M Gorman, J B Rdigway, D Henner, W L Wong, A M Rowland, C Kotts, M E Carver, and H M Shepard. Humanization of an anti-p185her2 antibody for human cancer therapy. *Proc Natl Acad Sci*, 89(10):4285–4289, 1992.

[54] M Caserta, G Camilloni, S Venditti, P Venditti, and E Di Mauro. Conformational information in dna: its role in the interaction with dna topoisomerase i and nucleosomes. *J Cell Biochem*, 55:93–97, 1994.

[55] J I Castrillo and S G Oliver. Yeast as a touchstone in post-genomic research: strategies for integrative analysis in functional genomics. *Journal of biochemistry and molecular biology*, 37(1):93–106, 2004.

[56] S Cawley, S Bekiranov, H H Ng, P Kapranov, E A Sekinger, D Kampa, A Piccolboni, V Sementchenko, J Cheng, A J Williams, et al. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, 116(4):499–509, 2004.

[57] A Ceglar and J F Roddick. Association mining. *ACM Computing Surveys*, 38(2), Article 5:1–42, 2006.

[58] G Chen, Q Wei, and E E Kerre. *Fuzzy Logic in Discovering Association Rules: An Overview. Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques*. Springer, New York, NY, USA, 2006.

[59] C J Cheng, Y C Lin, M T Tsai, C S Chen, M C Hsieh, C L Chen, and R B Yang. SCUBE2 suppresses breast tumor cell proliferation and confers a favorable prognosis in invasive breast cancer. *Cancer Research*, 69(8):3634–3641, 2009.

[60] Y Cheng and G Church. Biclustering of expression data. In *Proceedings of the ISMB: 93-103 2000*, pages 93–103, 2000.

[61] F Chibon, I de Mascarel, G Sierankowski, V Brouste, H Bonnefoi, M Debled, L Mauriac, and G MacGrogan. Prediction of HER2 gene status in Her2 2+ invasive breast cancer: a study of 108 cases comparing ASCO/CAP and FDA recommendations. *Modern pathology: an official journal of the United States and Canadian Academy of Pathology, Inc*, 22(3):403–409, 2009.

[62] B C Chien, Z L Lin, and T P Hong. An efficient clustering algorithm for mining fuzzy quantitative association rules. In *Proceedings of the 9th Int. Fuzzy Systems Assoc. World Congress); Vancouver, Canada*, pages 1306–1311, 2001.

[63] R Cho, M Campbell, E Winzeler, L Steinmetz, A Conway, L Wodicka, T Wolfsberg, A Gabrielian, D Landsman, and D Lockhart. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell.*, 2(1):65–73, 1998.

[64] S Chou, S Lane, and H Liu. Regulation of mating and filamentation genes by two distinct Ste12 complexes in Saccharomyces cerevisiae. *Molecular and Cellular Biology*, 26(13):4794–4805, 2006.

[65] G A Churchill. Stochastic models for heterogeneous dna sequences. *Bull Math Biol*, 51(1):79–94, 1989.

[66] N A Chuzhanova, M Krawczak, L A Nemytikova, V D Gusev, and D N Cooper. Promoter shuffling has occurred during the evolution of the vertebrate growth hormone gene. *Gene*, 254(1-2):9–18, 2000.

[67] A Coghlan and K H Wolfe. Relationship of codon bias to mrna concentration and protein length in saccharomyces cerevisiae. *Yeast*, 16:1131–1145, 2000.

[68] J Costas, F Casares, and J Vieira. Turnover of binding sites for transcription factors involved in early Drosophila development. *Gene*, 310:215–220, 2003.

[69] M Courel, S Lallet, J M Camadro, and P L Blaiseau. Direct activation of genes involved in intracellular iron use by the yeast iron-responsive transcription factor Aft2 without its paralog Aft1. *Molecular and Cellular Biology*, 25(15):6760–6771, 2005.

[70] C Creighton and S Hanash. Mining gene expression databases for association rules. *Bioinformatics*, 19(1):79–86, 2003.

[71] M Cuadros, R Villegas, et al. Systematic review of HER2 breast cancer testing. *Applied Immunohistochemistry & Molecular Morphology*, 17(1):1–7, 2009.

[72] P A Dafas and A S d'Avila. Discovering meaningful rules from gene expression data. *Current Bioinformatics*, 2:157–164, 2007.

[73] M Das and H.K. Dai. A survey of dna motif finding algorithms. *BMC Bioinformatics*, 8(Suppl 7):S21, 2007.

[74] F De Smet, J Mathys, K Marchal, G Thijs, B De Moor, and Y Moreau. Adaptive quality-based clustering of gene expression profiles. *Bioinformatics*, 18(5):735–746, 2002.

[75] M Delgado, N Marin, M J Martin-Bautista, D Sanchez, and M A Vila. Mining fuzzy association rules: an overview. In *Proceedings of the BISC International Workshop on Soft Computing for Internet and Bioinformatics; Berkeley, CA, USA*, 2003.

[76] M Delgado, N Marin, D Sanchez, and M A Vila. Fuzzy association rules: General model and applications. *IEEE Trans. Fuzzy Systems*, 11:214–225, 2003.

[77] D Dembele and D Kastner. Fuzzy c-means method for clustering microarray data. *Bioinformatics*, 19:973–980, 2001.

[78] E T Dermitzakis and A 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.

[79] X Dong, K H Zavitz, B J Thomas, M Lin, S Campbell, and S L Zipursky. Control of g1 in the developing drosophila eye: rca1 regulates cyclin a. *Genes Dev.*, 11:94–105, 1997.

[80] L G Dressler, D A Berry, Broadwater G, D Cowan, K Cox, S Griffin, A Miller, J Tse, D Novotny, D L Persons, M Barcos, I C Henderson, E T Liu, A Thor, D Budman, H Muss, L Norton, and D F Hayes. Comparison of her2 status by fluorescence in situ hybridization and

immunohistochemistry to predict benefit from dose escalation of adjuvant doxorubicin-based therapy in node-positive breast cancer patients. *J Clin Oncol*, 23(19):4287–4297, 2005.

[81] D Dubois, H Prade, and T Sudkamp. A discussion of indices for the evaluation of fuzzy associations in relational databases. In *Proceedings of the 10th Int. Fuzzy Systems Association World Congress (IFSA-03); Istambul, Turkey*, pages 111–118, 2003.

[82] B Dujon. The yeast genome project: what did we learn? *J. Biochem. Mol. Biol.*, 37:93–106, 2004.

[83] D Dumitrescu, B Lazzerini, and L C Jain. *Fuzzy Sets and Their Application to Clustering and Training*. CRC Press, Boca Raton, Florida, USA, 2000.

[84] S S Dwight, M A Harris, K. Dolinski, C A Ball, G Binkley, K R Christie, D G Fisk, L Issel-Tarver, M Schroeder, G Sherlock, A Sethuraman, S Weng, D Botstein, and J M Cherry. Saccharomyces genome database (sgd) provides secondary gene annotation using the gene ontology (go). *Nucleic Acids Res.*, 30(1):69–72, 2002.

[85] Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an oveview of the randomised trials. *Lancet*, 365(9472):1687–1717, 2005.

[86] M B Eisen, P T Spellman, P Brown, and D Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the Nat. Acad. Sci.*, 95(25):14863–14868, 1998.

[87] E Emberly, N Rajewsky, and E D Siggia. Conservation of regulatory elements between two species of Drosophila. *BMC bioinformatics*, 4(1):57–67, 2003.

[88] S Esseghir, S K Todd, T Hunt, R Poulsom, I Plaza-Menacho, J S Reis-Filho, and C M Isacke. A Role for Glial Cell Derived Neurotrophic Factor Induced Expression by Inflammatory Cytokines and RET/GFR

{alpha} 1 Receptor Up-regulation in Breast Cancer. *Cancer research*, 67(4):11732–11741, 2007.

[89] Z M Fatemeh, A Hayedeh, S Mehdei, N D Abbas, and G Bahram. New scoring schema for finding motifs in DNA Sequences. *BMC Bioinformatics*, 10:93–113, 2009.

[90] J Filipski and M Mucha. Structure, function and dna composition of saccharomyces cerevisiae chromatin loops. *Gene*, 300:63–68, 2002.

[91] A Freitas. Are we really discovering "interesting" knowledge from data? *Expert Update (the BCS-SGAI magazine)*, 9(1):41–47, 2006.

[92] M C Frith, U Hansen, and Z Weng. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, 17(10):878–889, 2001.

[93] A W Fu, M H Wong, C Sze, W C Wong, W L Wong, and W K Yu. Finding fuzzy sets for the mining of fuzzy association rules for numerical attributes. In *Proceedings of the 1st Int. Symp. on Intelligent Data Engienering and Learning (IDEAL'98): 1998; Hong Kong, China*, pages 263–268, 1998.

[94] A Gallo, T de Bie, and N Cristianini. MINI: Mining informative non-redundant itemsets. *Lecture Notes in Computer Science*, 4702:438–445, 2007.

[95] A P Gasch, P T Spellman, C M Kao, O Carmel-Harel, M B Eisen, G Storz, D Botstein, and P O Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11(12):4241–57, 2000.

[96] G Gasparini, P Boracchi, P Verderio, and P Bevilacqua. Cell kinetics in human breast cancer: comparison between the prognostic value of the cytofluorimetric S-phase fraction and that of the antibodies to Ki-67 and PCNA antigens detected by immunocytochemistry. *International Journal of Cancer*, 57(6):822–829, 1994.

[97] L Geng and H J Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys*, 38(3) Article 9:1–32, 2006.

[98] E Georgii, L Richter, U Ruckert, and S Kramer. Analyzing microarray data using quantitative association rules. *Bioinformatics*, 21(Suppl2):123–129, 2005.

[99] J L Gerton, J DeRisi, R Shroff, M Lichten, P O Brown, and T D Petes. Global mapping of meiotic recombination hotspots and coldspots in the yeast saccharomyces cerevisiae. In *Proceedings of the Natl. Acad. Sci; 1997; U. S. A*, pages 11383–11390, 1997.

[100] D H Glass. Fuzzy confirmation measures. *Fuzzy Sets and Systems*, 159:475–490, 2008.

[101] B Goethals and M J Zaki. Advances in frequent itemset mining implementations: Report on fimi'03. *SIGKDD Explorations*, 6(1):109–117, 2003.

[102] A Goffeau and et al. The yeast genome directory. *Nature*, 387:5–105, 1997.

[103] P Gönczy, B J Thomas, and S DiNardo. Routhex is a dose-dependent regulator of the second meitotic division during drosophila spermatogenesis. *Cell*, 77:1015–1025, 1994.

[104] G Grahne and J Zhu. Efficiently using prefix-trees in mining frequent itemsets. In *Proceedings of the ICDM Workshop on Frequent Itemset Mining Implementations; Melbourne, FL, USA*, 2003.

[105] V Guarneri, K Broglio, S W Kau, M Cristofanilli, A U Buzdar, V Valero, T Bucholz, F Meric, L Middleton, G N Hortogabyi, and A M Gonzalez-Angulo. Assessing genetic contributions to phenotypic differences among racial and etnic groups. *Journal of Clinical Oncology*, 24(7):1037–1044, 2006.

[106] A Gyenesey. *A Fuzzy Approach for Mining Quantitative Association Rules. TUCS Technical Report 336.* Department of Computer Science, University of Turku, Finland, 2000.

[107] M Hall and G Peters. Genetic alterations of cyclins, cyclin-dependent kinases, and cdk inhibitors in human cancer. *Adv. Cancer Res.*, 68:67–108, 1996.

[108] O Hallikas, K Palin, N Sinjushina, R Rautiainen, J Partanen, E Ukkonen, and J Taipale. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, 124(1):47–59, 2006.

[109] J Han and J Pei. Mining frequent patterns by pattern growth: Methodology and implications. *SIGKDD Explorations*, 2(2):14–20, 2000.

[110] C T Harbison, D B Gordon, T I Lee, N J Rinaldi, K D Macisaac, T W Danford, N M Hannett, J B Tagne, D B Reynolds, J Yoo, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.

[111] R C Hardison. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends in Genetics*, 16(9):369–372, 2000.

[112] J A Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1975.

[113] L H Hartwell and M B Kastan. Cell cycle control and cancer. *Science*, 266:1821–1828, 1994.

[114] T Hastie, R Tibshirani, M B Eisen, A Alizadeh, R Levy, L Staudt, W C Chan, D Botsteinm, and P Brown. Gene shaving as a method for identifying distinct sets of genes with similar expression. *Genom Biol*, 1:1–21, 2000.

[115] T R Hazbun and S Fields. A genome-wide screen for site-specific DNA-binding proteins. *Molecular & Cellular Proteomics*, 1(7):538–543, 2002.

[116] M W Helms, D Kemming, H Pospisil, U Vogt, H Buerger, E Korsching, C Liedtke, C M Schlotter, A Wang, S Y Chan, et al. Squalene epoxidase, located on chromosome 8q24. 1, is upregulated in 8q+ breast cancer

and indicates poor clinical outcome in stage I and II disease. *British Journal of Cancer*, 99(5):774–780, 2008.

[117] T P Hen, C S Kuo, and S C Chi. A fuzzy data mining algorithm for quantitative values. In *Proceedings of the 3rd Int. Conf. on Knowledge-Based Intelligent Information Engineering Systems: 1999; Adelaide, Australia*, pages 480–483, 1999.

[118] J Hermert and R Baldock. Mining spatial gene expression data for association rules. *Lecture notes in computer science*, 4414:66–76, 2007.

[119] G Z Hertz, G W Hartzell III, and G D Stormo. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Bioinformatics*, 6(2):81–92, 1990.

[120] D G Hicks and R R Tubbs. Assessment of the her2 status in breast cancer by fluorescence in situ hybridization: a technical review with interpretative guidelines. *Human Pathol*, 36(3):250–261, 2005.

[121] D Holder, R F Raubertas, V B Pikounis, V Svetnik, and K Soper. Statistical analysis of high density oligonucleotide arrays: a safer approach. In *Proceedings of the ASA Annual Meeting. GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data; Atlanta, USA*, 2001.

[122] F Hoppner, F Klawonn, R Kruse, and T Runkler. *Fuzzy Cluster Analysis, Methods for Classification, Data Analysis and Image Recognition*. Wiley, Chichester, UK, 1999.

[123] M Houtsma and A Swami. *Set Oriented Mining of Association Rules. Technical Report RJ 9567*. IBM Almaden Research Center, 1993.

[124] S Hua and Z Sun. A novel method of protein secondary estructure prediction with high segment overlap measure: support vector machine approach. *J. Mol. Biol.*, 308:397–407, 2001.

[125] J D Hughes, P W Estep, S Tavazoie, and G M Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *Journal of molecular biology*, 296(5):1205–1214, 2000.

[126] W Huh, J V Falvo, L C Gerke, A S Carroll, R W Howson, J S Weissman, and E K OŚhea. Global analysis of protein localization in budding yeast. *Nature*, 425:686–691, 2003.

[127] T Hunter and J Pines. Cyclins and cancer. ii: cyclin d and cdk inhibitors come of age. *Cell*, 79:573–582, 1994.

[128] T R Hvidsten, B Wilczynski, A Kryshtafovyc, J Tiuryn, J Komorowski, and K Fidelis. Discovering regulatory binding site modules using rule-based learning. *Genome Res.*, 15:856–866, 2005.

[129] S Impey, S R McCorkle, H Cha-Molstad, J M Dwyer, G S Yochum, J M Boss, S McWeeney, J J Dunn, G Mandel, and R H Goodman. Defining the CREB Regulon A Genome-Wide Analysis of Transcription Factor Regulatory Regions. *Cell*, 119(7):1041–1054, 2004.

[130] R A Irizarry, B M Bolstad, F Collin, L M Cope, B Hobbs, and T P Speed. Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31(4), 2003.

[131] R A Irizarry, B Hobbs, F Collin, Y Beazer-Barclay, K Antonellis, U Scherf, and T Speed. Exploration, normalization and summaries of high-density oligonucleotide array probe level data. *Biostatistics*, 4:249–264, 2003.

[132] M L Iruela-Arispe, P Porter, P Bornstein, and E H Sage. Thrombospondin-1, an inhibitor of angiogenesis, is regulated by progesterone in the human endometrium. *Journal of Clinical Investigation*, 97(2):403–412, 1996.

[133] L Iselius, J Slack, M Littler, and N E Morton. Genetic epidemiology of breast cancer in britain. *Ann Hum Genet*, 55(Pt 2):151–159, 1991.

[134] L S Ito, H Iwata, N Hamajima, T Saito, et al. Expression of interleukin-1B in human breast carcinoma. *Cancer*, 80:421–433, 1997.

[135] J K Jang, L Messina, M B Erdman, Arbel, and R S Hawley. Induction of metaphase arrest in drosophila oocytes by chiasma-based kinetochore tension. *Science*, 268:1917–1919, 1995.

[136] R Jansen and M Gerstein. Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucleic Acids Res*, 28:1481–1488, 2000.

[137] R Jansen, H Yu, D Greenbaum, Y Kluger, N J Krogan, S Chung, A Emili, M Snyder, J F Greenblatt, and M Gerstein. Bayesian networks approach for prediting protein-protein interactions from genomic data. *Science*, 302:449–453, 2003.

[138] L J Jensen, M Kuhn, M Stark, S Chaffron, C Creevey, J Muller, T Doerks, P Julien, A Roth, M Simonovic, et al. STRING 8-a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37 (Database issue):D412, 2009.

[139] D Jiang, C Tang, and A Zhang. Cluster analysis for gene expression data: A survey. *IEEE Transaction on Knowledge and Data Engineering*, 16(11):1370–1386, 2004.

[140] N Jiang, L J Leach, X Hu, E Potokina, T Jia, A Druka, R Waugh, M J Kearsey, and Z W Luo. Mehods for evaluating gene expression from affymetrix microarray datasets. *BMC Bioinformatics*, 9:284–293, 2008.

[141] H Jiawei and K Micheline. *Data Mining. Concepts and Techniques.* Morgan Kaufman, San Francisco, CA, USA, 2006.

[142] O Johansson, W Alkema, W W Wasserman, and J Lagergren. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics*, 19(1):169–176, 2003.

[143] M Kanehisa and P Bork. Bioinformatics in the post-sequence era. *Nature Genet.*, 33(Suppl):305–310, 2003.

[144] M Kanehisa and S Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nculeic Acids Research*, 28(1):27–30, 2000.

[145] T S Kim, S B Lee, and H S 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.

[146] J Klema. Constraint-based knowledge discovery from sage data. *Silico Biology*, 8:157–175, 2008.

[147] K Klepper, G K Sandve, O Abul, J Johansen, and F Drablos. Assessment of composite motif discovery methods. *BMC bioinformatics*, 9(1):123–138, 2008.

[148] W Klosgen. *Explora: A multipattern and multistrategy discovery assistant.Advances in Knowledge Discovery and Data Mining*. MIT Press, Menlo Park, CA, USA, 1996.

[149] Y Kluger, R Basri, J T Chang, and M Gerstein. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research*, 16(11):1370–1386, 2003.

[150] C Koch, T Moll, M Neuberg, H Ahorn, and K Nasmyth. A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science*, 261(5128):1551–1557, 1993.

[151] P Labhart, S Karmakar, E M Salicru, B S Egan, V Alexiadis, B W O'Malley, and C L Smith. Identification of target genes in breast cancer cells directly regulated by the SRC-3/AIB1 coactivator. *Proceedings of the National Academy of Sciences*, 102(5):1339–1344, 2005.

[152] C E Lawrence, S F Altschul, M S Boguski, J S Liu, A F Neuwald, and J C Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, 1993.

[153] L Lazzeroni and A Owen. Plaid models for gene expression data. *Statistica Sinica*, 12:61–86, 2002.

[154] M S Lee and W T Garrard. Positive dna supercoiling generates a chromatin conformation characteristic of highly active genes. In *Proceedings of the Natl Acad Sci: 1991; U. S. A.*, pages 88:9675–9679, 1991.

[155] S R Lee, S M Ramos, A Ko, D Masiello, K D Swanson, M L Lu, and S P Balk. AR and ER interaction with a p21-activated kinase (PAK6). *Molecular Endocrinology*, 16(1):85–99, 2002.

[156] T I Lee and R A Young. Transcription of eukaryotic proteincoding genes. *Ann Rev Genet*, 34:77–137, 2000.

[157] P Lenca, P Meyer, B Vaillant, and S Lallich. *A multicriteria decision aid for interestingness measure selection. Tech. Rep: LUSSI-TR-2004-EN*. LUSSI Department, GET/ENST, 2004.

[158] L A Lettice, S J H Heaney, L A Purdie, L Li, P de Beer, B A Oostra, D Goode, G Elgar, R E Hill, and E de Graaff. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, 12(14):1725–1735, 2003.

[159] C Li and W H Wong. *The analysis of gene expression data: methods and software. DNA-Chip Analyzer (dChip)*. Springer, New York, NY, USA, 2001.

[160] C Li and W H Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc Natl Acad Sci USA*, 98(1):31–36, 2001.

[161] X Liu and L Wang. Computing the maximum similarity bi-clustersof gene expression data. *Bioinformatics*, 23(1):50–56, 2007.

[162] W S Lo and A 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.

[163] F J Lopez, A Blanco, F Garcia, C Cano, and A Marin. Fuzzy association rules for biological data analysis: a case study on yeast. *BMC Bioinformatics*, 9:107–115, 2008.

[164] F J Lopez, A Blanco, F Garcia, and A Marin. Extracting biological knowledge by fuzzy association rule mining. In *Proceedings of the IEEE*

International Conference on Fuzzy Systems:2007; London, UK, pages 583–588, 2007.

[165] L L Lutfiyya, V R Iyer, J. DeRisi, M J DeVit, P O Brown, and M Johnston. Characterization of three related glucose repressors and genes they regulate in Saccharomyces cerevisiae. *Genetics*, 150(4):1377–1391, 1998.

[166] A J Mackey, T A Haystead, and W R Pearson. Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. *Molecular and Cellular Proteomics*, 1(2):139–147, 2002.

[167] S Madeira and A Olivera. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45, 2004.

[168] A Marin, M Gallardo, Y Kato, K Shirahige, G Gutierrez, K Ohta, and A Aguilera. Relationship between g+c content, orf-length and mrna concentration in saccharomyces cerevisiae. *Yeast*, 20:703–711, 2003.

[169] A Marin, M Wang, and G Gutierrez. Short-range compositional correlation in the yeast genome depends on transcriptional orientation. *Gene*, 333:151–155, 2004.

[170] R Martinez, C Pasquier, and N Pasquier. Genminer: mining informative association rules from genomic data. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine: 2007; Silicon Valley, USA*, pages 15–22, 2007.

[171] R D Mass, M F Press, S Anderson, M A Cobleigh, C L Vogel, N Dybdal, G Leiberman, and D J Slamon. Evaluation of clinical outcomes according to her2 detection by fluorescence in situ hybridization in women with metastatic breast cancer treated with trastuzumab. *Clin Breast Cancer*, 6(3):240–246, 2005.

[172] A Mateos, J Herrero, J Tamames, and J Dopazo. *Methods of Microarray Data Analysis II: Supervised neural networks for clustering conditions in*

DNA array data after reducing noise by clustering gene expression profilesFuzzy sets theory and its applications. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.

[173] A Matyash, H R Chung, and H Jäckle. Genome-wide mapping of in vivo targets of the Drosophila transcription factor Kr "uppel. *Journal of Biological Chemistry*, 279(29):30689–30696, 2004.

[174] T Mcintosh and S Chawla. Hihgh-confidence rule mining for microarray analysis. *IEEE Trans. on Computational Biology and Bioinformatics*, 4(4):611–623, 2007.

[175] R J Miller and Y Yang. Association rules over interval data. In *Proceedings of the ACM SIGMOD (SIGMOD'97); Tucson, Arizona, USA*, pages 452–461, 1997.

[176] X C Morgan, S Ni, D P Miranker, and V R Iyer. Predicting combinatorial binding of transcription factors to regulatory elements in the human genome by association rule mining. *BMC Bioinformatics*, 8:445–458, 2007.

[177] A M Moses, D Y Chiang, M Kellis, E S Lander, and M B Eisen. Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evolutionary Biology*, 3(1):19–31, 2003.

[178] S M Mount, C Burks, G Herts, G D Stormo, O White, and C Fields. Splicing signals in Drosophila: intron size, information content, and consensus sequences. *Nucleic Acids Research*, 20(16):4255, 1992.

[179] F Naef, C R Hacker, N Patil, and M Magnasco. Empirical characterization of the expression noise ratio structure in high-density oligonucleotide arrays. *Genom Biol*, 3(4):research0018, 2002.

[180] A Narayanan, E C Keedwell, and B Olsson. Artificial intelligence techniques for bioinformatics. *Appl. Bioinf.*, 1:191–222, 2002.

[181] M A Nobrega, I Ovcharenko, V Afzal, and E M Rubin. Scanning human gene deserts for long-range enhancers. *Science*, 302(5644):413, 2003.

[182] A Ocana, J J Cruz, and A Pandiella. Trastuzumab and antiestrogen therapy: focus on mechanisms of action and resistance. *Am J Clin Oncol*, 29(1):90–95, 2006.

[183] L J Oehlen and F R Cross. G1 cyclins cln1 and cln2 repress the mating factor response pathway aat start in the yeast cell cycle. *Genes Dev.*, 8:1058–1070, 1994.

[184] L J Oehlen, J D McKinney, and F R Cross. Ste12 and mcm1 regulate cell cycle-dependent transcription of far1. *Cell. Biol.*, 16:2830–2837, 1996.

[185] U Ohler, H Niemann, and G M Rubin. Joint modeling of dna sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics*, 17(S1):S199–S206, 2001.

[186] R Osada, E Zaslavsky, and M Singh. Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics-Oxford*, 20(18):3516–3525, 2004.

[187] B I Osborne and L Guarente. Transcription by rna polymerase ii induces changes of dna topology in yeast. *Genes Dev*, 2:766–772, 1988.

[188] T Oyama and K Kitano. Extraction of knowledge on protein-protein interaction by association rule discovery. *Bioinformatics*, 18(5):705–714, 2002.

[189] R B Painter and B R Youn. Radiosensitivity in ataxiatelangiectasia: a new explantation. *Cell. Biol.*, 77:7315–7317, 1980.

[190] R D Pascual-Marqui, A D Pascual-Montano, K Kochi, and J M Carazo. Smoothly distributed fuzzy c-means: a new self-organizing map. *Pattern Recognition*, 34:2395–2402, 2001.

[191] N Pasquier, Y Bastide, R Taouil, and L Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1):25–46, 1999.

[192] G Pauletti, S Dandekar, H Rong, L Ramos, H Peng, R Seshadri, and D J Slamon. Assessment of methods for tissue-based detecion of the her-2/neu alteration in human breast cancer: a direct comparison of fluorescence in situ hybridization and immunohistochemistry. *J Clin Oncol*, 18(21):3651–3664, 2000.

[193] L J Peck and J C Wang. Transcriptional block caused by a negative supercoiling induced structural change in an alternating cg sequence. *Cell*, 40:129–137, 1985.

[194] D S Pederson and R H Morse. Effect of transcription of yeast chromatin on dna topology in vivo. *EMBO Journal*, 9:1873–1881, 1990.

[195] R Peeters. The maximum edge biclique problem is np-complete. *Discrete Applied Math.*, 131(3):651–654, 2003.

[196] J Pei, J Han, and L V S Lakshamanan. Mining frequent itemsets with convertible cosntraints. In *Proceedings of the 17th Int. Conf. on Data Engineering (ICDE'01): 2001; Heidelberg, Germany*, pages 433–442, 2001.

[197] C M Perou, T Sorlie, M B Eisen, M van de Rijn, S S Jeffrey, C A Rees, J R Pollack, D T Ross, H Johnsen, L A Akslen, O Fluge, A Pergamenschikov, C Williams, S K Zhu, P E Lonning, A L Borressen-Dale, P O Brown, and D Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, 2000.

[198] T H Pham, K Satou, and T B Ho. Mining yeast transcriptional regulatory modules from factor DNA-binding sites and gene expression data. *Genome Informatics Series*, 15(2):287–295, 2004.

[199] G Piatetsky-Shapiro. *Discovery, analysis and presentation of strong rules. Knowledge Discovery in Databases*. MIT press, Cambridge, MA, USA, 1991.

[200] A Pic-Taylor, Z Darieva, B A Morgan, and A D Sharrocks. Regulation of cell cycle-specific gene expression through cyclin-dependent kinase-

mediated phosphorylation of the forkhead transcription factor Fkh2p. *Molecular and Cellular Biology*, 24(22):10036–10046, 2004.

[201] Y Pilpel, P Sudarsanam, and G M Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature genetics*, 29(2):153–159, 2001.

[202] I Ponzoni, F Azuaje, J Augusto, and D Glass. Inferring adaptive regulation thresholds and association rules from gene expression data through combinatorial optimization learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 4(4):624–634, 2007.

[203] A Preli, S Bleuler, P Zimmermann, A Wille, P Buhlmann, W Gruissem, L Hennig, Thiele, and E Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22:1122–1129, 2006.

[204] M F Press, M C Pike, V R Chazin, G Hung, J A Udove, M Markowicz, J Danyluk, W Godolphin, M Sliwkowski, R Akita, et al. Her-2/neu expression in node-negative breast cancer: direct tissue quantitation by computerized image analysis and association of overexpressio nwith increased risk of recurrence disease. *Cancer Res*, 53(20):4960–4970, 1993.

[205] Y Qu and S Xu. Supervised cluster analysis for microarray data based on multivariate gaussian mixture. *Bioinformatics*, 20:1905–1913, 2004.

[206] E A Rakha, D A El-Rehim, C Paish, A R Green, A H S Lee, J F Robertson, R W Blamey, D Macmillan, and I O Ellis. Basal phenotype identifies a poor prognosis subgroup of breast cancer clinical importance. *European Journal of Cancer*, 42(18):3149–3156, 2006.

[207] B Ren, F Robert, J J Wyrick, O Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, et al. Genome-wide location and function of DNA binding proteins. *Science's STKE*, 290(5500):2306–2309, 2000.

[208] A Rodriguez, J M Carazo, and O Trelles-Salazar. Mining association rules from biological databases. *Journal of the American Society for Information Science and Technology*, 56:493–504, 2005.

[209] M Ronen and D 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.

[210] J S Ross and J A Fletcher. The her-2/neu oncogene in breast cancer: prognostic factor, predictive factor, and target for therapy. *Stem Cells*, 16(6):413–428, 1998.

[211] B Rost and C Sander. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, 232:584–599, 1993.

[212] T Ryu, Y Kim, D W Kim, and D Lee. Computational identification of combinatorial regulation and transcription factor binding sites. *Biotechnology and Bioengineering*, 97(6):1594–1601, 2007.

[213] A Sadr-Nabavi, J Ramser, J Volkmann, J Naehrig, F Wiesmann, B Betz, H Hellebrand, S Engert, S Seitz, R Kreutzfeld, et al. Decreased expression of angiogenesis antagonist EFEMP1 in sporadic breast cancer is caused by aberrant promoter methylation and points to an impact of EFEMP1 as molecular biomarker. *International Journal of Cancer*, 124(7):1727–1735.

[214] H Salgado, S Gama-Castro, A Martinez-Antonio, E Diaz-Peredo, F Sanchez-Solano, M Peralta-Gil, D Garcia-Alonso, V Jimenez-Jacinto, A Santos-Zavaleta, C Bonavides-Martinez, et al. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. *Nucleic Acids Research*, 32(Database Issue):D303, 2004.

[215] A Sandelin, W Alkema, P Engstrom, W W Wasserman, and B Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(Database Issue):D91, 2004.

[216] A Sandelin, W W Wasserman, and B Lenhard. ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Research*, 32(Web Server Issue):W249, 2004.

[217] R Sanges, E Kalmar, P Claudiani, M D'Amato, F Muller, and E Stupka. Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage. *Genome Biology*, 7(7):R56, 2006.

[218] P M Santos, T Simões, and I Sá-Correia. Insights into yeast adaptive response to the agricultural fungicide mancozeb: A toxicoproteomics approach. *Proteomics*, 9(3):657–670, 2009.

[219] S B Schawalder, M Kabani, I Howald, U Choudhury, M Werner, and D Shore. Growth-regulated recruitment of the essential yeast ribosomal protein gene activator Ifh1. *Nature*, 432(7020):1058–1061, 2004.

[220] H J Schüller. Transcriptional control of nonfermentative metabolism in the yeast Saccharomyces cerevisiae. *Current genetics*, 43(3):139–160, 2003.

[221] R Scully, J Chen, R L Ochs, K Keegan, M Hoekstra, J Feunteun, and D M Livingston. Dynamic changes of brca1 subnuclear location and phosphorylation state are initiated by dna damage. *Cell. Biol.*, 90:425–435, 1997.

[222] A D Seidman, M N Fornier, F J Esteva, L Tan, S Kaptain, A Bach, K S Panageas, C Arroyo, V Valero, V Currie, T Gilewski, M Theodoulou, M E Monynahan, M Moasser, N Sklarin, M Dickler, G D'Andrea, M Cristofanilli, E Rivera, G N Hortobagyi, L Norton, and C A Hudis. Weekly trastuzumab and paclitaxel therapy for metastatic breast cancer with analysis of efficacy by her2 immunophenotype and gene amplification. *J Clin Oncol*, 19(10):2587–2595, 2001.

[223] R Seshadri, F A Firgaira, D J Horsfall, K McCaul, V Setlur, and P Kitchen. Clinical significance of her-2/neu oncogene amplificatio nin primary breast cancer. *J Clin Oncol*, 11(10):1936–1942, 1993.

[224] R Sharan, I Ovcharenko, A Ben-Hur, and R M Karp. CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics-Oxford*, 19(1):283–291, 2003.

[225] C J Sherr and J M Roberts. Inhibitors of mammalian g1 cyclin-dependent kinases. *Genes Dev.*, 9:1149–1163, 1995.

[226] A Siepel and D Haussler. Combining phylogenetic and hidden markov models in biosequence analysis. In *Proceedings of the 7th Annual International Conference on Research in Computational Molecular Biology: 2003; New York, NY, USA*, pages 277–286, 2003.

[227] A H Sims. Bioinformatics and breast cancer: what can high-thoroughput genomic approaches actually tell us? *J. Clin. Pathol.*, 62:879–885, 2009.

[228] D J Slamon, G M Clark, S G Wong, W J Levin, A Ullrich, and W L McGuire. Human breast cancer: correlation of relapse and survival with amplification of the her-2/neu oncogene. *Science*, 235(4785):177–182, 1987.

[229] D J Slamon, W Godolphin, L A Jones, J A Holt, S G Wong, D E Keith, W J Levin, S G Stuart, J Udove, A Ullrich, et al. Studies of the her-2/neu proto-oncogene in human breast and ovarian cancer. *Science*, 244(4905):707–712, 1989.

[230] D J Slamon, B Leyland-Jones, S Shak, H Fuchs, V Paton, A Bajamonde, T Fleming, W Eiermann, J Wolter, M Pegram, J Baselga, and L Norton. Use of chemotherapy plus a monoclonal antibody against her2 for metastatic breast cancer that overexpresses her2. *N Engl J Med*, 344(11):783–792, 2001.

[231] D J Slamon, E H Romond, and E A Perez. Advances in adjuvant therapy for breast cancer. *Clin Adv Hematol Oncol*, 4(3):suppl–9, 2006.

[232] T F Smith and M S Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[233] T Sorlie, C M Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, M B Eisen, M Rijn, S S Jeffrey, T Thorsen, H Quist, J C Matese, P O Brown, D Botstein, P E Lonning, and A L Borresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclases with clinical implications. *Proc Natl Acad Sci*, 98(19):10869–10874, 2001.

[234] T Sorlie, R Tibshirani, J Parker, T Hastie, J S Marron, A Nobel, S Deng, H Johnsen, R Pesich, S Geisler, J Demeter, C M Perou, P E Lonning, P Brown, A L Borresen-Dale, and D Botstein. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci*, 100(14):8418–8423, 2003.

[235] T Sorlie, Y Wang, C Xiao, H Johnsen, B Naume, R R Samaha, and A L Borresen-Dale. Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. *BMC Genomics*, 26(7):127–141, 2006.

[236] V Spirin and L A Mirny. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci*, 100:12123–12128, 2003.

[237] R Srikant and R Agrawal. Mining quantitative association rules in large relational databases. In *Proceedings of the ACM SIGMOD Conf. Management Data; Montreal, Canada*, pages 1–12, 1996.

[238] R Staden. Searching for patterns in protein and nucleic acid sequences. *Methods in Enzymology*, 183:193–211, 1990.

[239] E Stockert, E Jager, Y T Chen, M J Scanlan, I Gout, J Karbach, M Arand, A Knuth, and L J Old. A survey of the humoral immune response of cancer patients to a panel of human tumor antigens. *Journal of Experimental Medicine*, 187(8):1349–1354, 1998.

[240] N Sugimoto and H Iba. Inference of gene regulatory networks by means of dynamic differential bayesian networks and non-parametric regression. *Genome Informatics*, 15(2):121–130, 2004.

[241] H Sun, T de Bie, V Storms, Q Fu, T Dhollander, K Lemmens, A Verstuyf, B de Moor, and K Marchal. ModuleDigger: an itemset mining framework for the detection of cis-regulatory modules. *BMC bioinformatics*, 10(Suppl 1):S30, 2009.

[242] S Swift, A Tucker, V Vinciotti, N Martin, C Orengo, X Liu, and P Kellam. Consensus clustering and functional interpretation of gene-expression data. *Genome Biology*, 5:R94, 2004.

[243] P Tamayo, D Slonim, J Mesirov, Q Zhu, S Kitareewan, E Dmitrovsky, E S Lander, and T R Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the Nat. Acad. Sci.*, 96:2907–2912, 1999.

[244] P Tan, V Kumar, and J Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (KDD 2002); Edmonton, Canada*, pages 32–41, 2002.

[245] P N Tan, M Steinbach, and V Kumar. *An Introduction to Data Mining*. Addison-Wesley Longman Publishing, 2005.

[246] A Tanay, R Sharan, and R Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18:S136–S144, 2002.

[247] M C Teixeira, A R Fernandes, N P Mira, J D Becker, and I 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.

[248] A Testa, G Donati, P Yan, F Romani, T H M Huang, M A Vigano, and R Mantovani. Chromatin Immunoprecipitation(ChIP) on Chip Experiments Uncover a Widespread Distribution of NF-Y Binding CCAAT Sites Outside of Core Promoters. *Journal of Biological Chemistry*, 280(14):13606–13615, 2005.

[249] D Thieffry, H Salgado, A M Huerta, and J Collado-Vides. Prediction of transcriptional regulatory sites in the complete genome sequence of Escherichia coli K-12. *Bioinformatics*, 14(5):391–400, 1998.

[250] N Thierry-Mieg and L Trilling. Interdb, a prediction-oriented protein interaction database for c. elegans. *Lecture Notes in Computer Science*, 2066:144–154, 2001.

[251] V B Thogersen, B S Sorensen, S S Poulsen, T F Orntoft, H Wolf, and E Nexo. A subclass of HER1 ligands are prognostic markers for survival in bladder cancer patients. *Cancer Res*, 61:6227–6233, 2001.

[252] B J Thomas, D A Gunning, J Cho, and L Zipursky. Cell cycle progression in the developoing drosophila eye: roughex encodes a novel protein required for the establishment of g1. *Cell*, 77:1003–1014, 1994.

[253] B J Thomas, K H Zavitz, X Dong, Lane M E, K Weigmann, R L Jr Finley, R Brent, C F Lehner, and S L Zipursky. roughex down-regulates g2 cyclins in g1. *Genes Dev.*, 11:1289–1298, 1997.

[254] M Thomassen, Q Tan, and T A Kruse. Gene expression meta-analysis identifies chromosomal regions and candidate genes involved in breast cancer metastasis. *Breast Cancer Research and Treatment*, 113(2):239–249, 2009.

[255] M W Thomson, R R McInnes, and Willard H F. *Genetics in medicine*. W.B. Saunders, Filadelfia, PA, USA, 2000.

[256] I Tirosh, A Weinberger, M Carmi, and N Barkai. A genetic signature for inter-species variations in gene expression. *Nature Genetics*, 38:830–834, 2006.

[257] A Tomovic and E J Oakeley. Position dependencies in transcription factor binding sites. *Bioinformatics*, 23(8):933–941, 2007.

[258] H Toyoda, T Komurasaki, D Uchida, and S Morimoto. Distribution of mRNA for human epiregulin, a differentially expressed member of the

epidermal growth factor family. *Biochemical Journal*, 326(Pt 1):69–75, 1997.

[259] J V Turatsinze, M Thomas-Chollier, M Defrance, and J van Helden. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat Protoc*, 3(10):1578–1588, 2008.

[260] H Turner, T Bailey, and W Krzanowski. Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational Statistics & Data Analysis*, 48:235–254, 2005.

[261] H Turner, T Bailey, W Krzanowski, and C Hemingway. Biclustering models for structured microarray data. *IEEE/ACM Transcations on Computational Biology and Bioinformatics*, 2(4):316–329, 2005.

[262] V Tusher, R Tibshirani, and C Chu. Significance analysis of microarrays applied to ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98:5116–5121, 2001.

[263] A Tuzhilin and G Adomavicius. Handling very large numbers of association rules in the analysis of microarray data. In *Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining; Edmonton, Alberta, Canada*, pages 396–404, 2002.

[264] H Tveit, T Mollestad, and A Laegreid. The alignment of the medical subject headings to the gene ontology and its application in gene annotation. In *Proceedings of the 4th Int. Conf. on Rough Sets and Current Trends in Computing (RSCTC 2004):2004; Upsala, Sweden*, pages 798–804, 2004.

[265] B Vaillant, P Lenca, and S Lallic. A clustering of interestingness measures. In *Proceedings of the 7th International conference on Discovery Science (DS 2004); Padova, Italy*, pages 290–297, 2004.

[266] P van Loo and P Marynen. Computational methods for the detection of cis-regulatory modules. *Briefings in Bioinformatics*, 10(5):509–524, 2009.

[267] T Vavouri and G 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.

[268] M H Verlhac, J Z Kubiak, M Weber, G Geraud, W H Colledge, M J Evans, and B Maro. Mos is required for map kinase activation and is involved in microtubule organization during meitotic maturation in the mouse. *Development*, 122:815–822, 1996.

[269] M Vingron, A Brazma, R Coulson, J Helden, T Manke, K Palin, O Sand, and E Ukkonen. Integrating sequence, evolution and functional genomics in regulatory genomics. *Genome Biology*, 10:202–209, 2009.

[270] C L Vogel, M A Cobleigh, D Tripathy, J C Gutheil, L N Harris, L Fehrenbacher, D J Slamon, M Murphy, W F Novotny, M Burchmore, S Shak, S J Stewart, and M Press. Efficacy and safety of trastuzumab as a single agent in first-line treatment of her2-overexpressing metastatic breast cancer. *J Clin Oncol*, 20(3):719–726, 2002.

[271] K Wang, L Tang, J Han, and J Liu. Top down fp-growth for association rule mining. In *Proceedings of the 6th Pacific Area Conference on Knowledge Discovery and Data Mining; Taipei, Taiwan*, pages 334–340, 2002.

[272] T C Wang, Cardiff R D, L Zukerberg, E Lees, A Arnold, and E V Schmidt. Cyclins and cancer. ii: cyclin d and cdk inhibitors come of age. *Nature*, 369:669–671, 1994.

[273] Y H Wang and J D Griffith. The [(g/c)3nn]n motif: a common dna repeat that excludes nucleosomes. In *Proceedings of the Natl Acad Sci: 1996; U. S. A.*, pages 93:8863–8867, 1996.

[274] J Warringer and A Blomberg. Evolutionary constraints on yeast protein size. *BMC Evol Biol*, 15:6–51, 2006.

[275] T Weinert. A dna damage checkpoint meets the cell cycle engine. *Science*, 277:1450–1451, 1997.

[276] T A Weinert and L H Hartwell. The rad9 gene controls the cell cycle response to dna damage in saccharomyces cerevisae. *Science*, 241:317–322, 1988.

[277] E Wingender, X Chen, R Hehl, H Karas, I Liebich, V Matys, T Meinhardt, et al. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Research*, 28(1):316–319, 2000.

[278] C Wittenberg, Sugimoto K, and S I Reed. G1-specific cyclins of s. cerevisae: cell cycle periodicity, regulation by mating pheromone, and association with the p34cdc28 protein kinase. *Cell*, 62:225–237, 1990.

[279] J A Wohlschlegel and J R Yates. Where's waldo in yeast? *Nature*, 425(6959):671–672, 2003.

[280] T Wölfel, M Hauer, J Schenider, M Serrano, C Wölfel, E Klehman-Hieb, E De-Plaen, T Hankeln, K H Meyer-zum Büschenfelde, and D Beach. A p16ink4a-insensitive cdk4 mutant targeted by cytolytic t lymphocytes in a human melanoma. *Science*, 269:1281–1284, 1995.

[281] A Woolfe, M Goodson, D K Goode, P Snell, G K McEwen, T Vavouri, S F Smith, P North, H Callaway, K Kelly, et al. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol*, 3(1):e7, 2005.

[282] Z Wu and R A Irizarry. Preprocessing of oligonucleotide array data. *Nat Biotechnol*, 22(6):656–658, 2004.

[283] J J Wyrick, F C P Holstege, E G Jennings, H C Causton, D Shore, M Grunstein, E S Lander, and R A Young. Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. *Nature*, 402:418–421, 1999.

[284] J J Wyrick and R A Young. Deciphering gene expression regulatory networks. *Curr Opin Genet Dev*, 12:130–136, 2002.

[285] J Yang, H Wang, W Wang, and P Yu. Enhanced biclustering on expression data. In *Proceedings of the 3rd IEEE Symposium on BioInformatics*

and BioEngineering:2003; Bethesda, Maryland, USA, pages 321–327, 2003.

[286] J Yang, W Wang, H Wang, and P Yu. Clustering by pattern similarity in large data sets. In *Proceedings of the ACM SIGMOD:2002; Madison, Wisconsin, USA*, pages 394–405, 2002.

[287] J Yang, W Wang, H Wang, and P Yu. δ-cluters: capturing subspace correlation in a large data set. In *Proceedings of the 18th IEEE International Conference on Data Engineering:2002; San Jose, CA, USA*, pages 517–528, 2002.

[288] H Yaziji, L C Goldstein, T S Barry, R Werling, H Hwang, G K Ellis, J R Gralow, R B Linvingston, and A M Gown. Her-2 testing in breast cancer using parallel tissue-based methods. *JAMA*, 291(16):1972–1977, 2004.

[289] C Yu, N Zavaljevski, V Desai, S Johnson, F J Stevens, and J Reifman. The development of pipa: an integrated and automated pipeline for genome-wide protein funciton annotation. *BMC Bioinformatics*, 9:52–62, 2008.

[290] L A Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

[291] M J Zaki and C J Hsiao. CHARM: An efficient algorithm for closed itemset mining. In *Proceedings of the 2nd SIAM International Conference on Data Mining; Arlington, VA, USA*, pages 457–473, 2002.

[292] M J Zaki, S Parthasarathy, M Ogihara, and W Li. New algorithms for fast discovery of association rules. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD'97): 1997; Menlo Park, California, USA*, pages 283–296, 1997.

[293] B Zanolari and H Riezman. Quantitation of alpha-factor internalization and response during the saccharomyces cerevisae cell cycle. *Moll. Cell. Biol.*, 11:5251–5258, 1991.

[294] R Zeillinger, F Kury, K Czewenka, E Kubista, G Sliutz, W Knogler, J Huber, G Zielinski, C ad Reiner, R Jakesz, et al. Her-2 amplification,

steroid receptors and epidermal growth factor receptor in primary breast cancer. *Oncogene*, 4(1):109–114, 1989.

[295] L Zhang, M F Miles, and K D Aldape. A model of molecular interactions on short oligonucleotide microarrays. *Nat Botechnol*, 21(7):818–821, 2006.

[296] W Zhang. Mining fuzzy quantitative association rules. In *Proceedings of the 11th Int. Conf. on Tools with A.I.: 1999; Chicago, Illinois, USA*, pages 99–102, 1999.