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

Departamento de Ciencias de la Computación e I. A.

Fuzzy Technology for Gene Ontology and Transcription Factor Binding Sites

PhD Dissertation

Fernando García Alcalde

Supervisor: Dr. Armando Blanco Morón

Editor: Editorial de la Universidad de Granada
Autor: Fernando García Alcalde
D.L.: GR 2308-2010
ISBN: 978-84-693-1311-4

La memoria “Fuzzy Technology for Gene Ontology and Transcription Factor Binding Sites”, que presenta D. Fernando García Alcalde 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.

Granada, Enero de 2010.

Armando Blanco Morón

Fernando García Alcalde

A Burgos.

Contents

Cover 1

Contents i

Agradecimientos xi

Resumen xiii

Motivación xiii

Objetivos xvii

Contenidos xviii

I Introduction 1

1 Introduction 3

1.1 Antecedents 3

1.2 Objectives 6

1.3 Thesis Structure 7

II Preliminaries 9

2 Biology and Bioinformatics 11

2.1 Biology 11

2.2 Bioinformatics 21

2.3 Gene Ontology 31

3 Fuzzy theory **39**
3.1 Fuzzy sets . . . 39
3.2 Fuzzy set operators . . . 43
3.3 Fuzzy Clustering . . . 44
3.4 Fuzzy Measures . . . 47
3.5 $\lambda-$Fuzzy Measures . . . 49
3.6 Fuzzy Integral . . . 49
3.7 Other Soft Computing Methods . . . 51
3.8 Fuzzy Applications in Bioinformatics . . . 53

III Thesis Contributions **55**

4 Semantic Measures for Gene Ontology **57**
4.1 Semantic Similarity in Ontologies . . . 58
4.2 GO Crisp Semantic Measures for Related Proteins Recognition 62
4.3 Fuzzy Semantic Similarity Measure for Gene Ontology . . . 78
4.4 Protein Classification using Gene Ontology . . . 81
4.5 Concluding Remarks . . . 87

5 Comparing TFBS Motifs **89**
5.1 Background . . . 90
5.2 Motif Representations . . . 91
5.3 Probabilistic Measures . . . 93
5.4 Adapting Fuzzy Measures . . . 95
5.5 Probability and Fuzzy Measures Analysis . . . 98
5.6 Concluding Remarks . . . 104

6 Advances in Motif Measures **107**
6.1 Motivation . . . 107
6.2 New approaches . . . 109
6.3 Fuzzy Integral Similarity for Motifs (FISim) . . . 110
6.4 Novel Clustering Methodology for Motifs . . . 114
6.5 Experiments . . . 115

6.6 Analysis of the Results 127
6.7 Concluding Remarks 130

7 Scoring DNA Sequences against TFBS Motifs 131
7.1 Introduction 132
7.2 Alternative approaches 134
7.3 New Intuitionistic Approach (SC_{intuit}) 137
7.4 Comparative Study of the Performance of SC_{intuit} 140
7.5 Analysis of the Results 146
7.6 Study of SNPs in TNFR1 for the Response against *Aspergillus* . 149
7.7 Concluding Remarks 155

IV Conclusions 157

8 Conclusions and Future Work 159
8.1 Conclusions 159
8.2 Future Work 161
8.3 Publications 162

Bibliography 167

Appendix A 183

Appendix B 189

List of Figures

2.1 Cell structure . . . 12
2.2 Nucleotide . . . 13
2.3 DNA . . . 14
2.4 DNA double helix . . . 15
2.5 Human chromosomes . . . 16
2.6 From gene to protein . . . 18
2.7 Central Dogma of Molecular Biology . . . 18
2.8 Transcription factor interactions . . . 22
2.9 Growth of GenBank . . . 23
2.10 Bioinformatics as an interdisciplinary field . . . 24
2.11 GO gene products distribution . . . 33
2.12 Directed acyclic graph taken from GO . . . 34
2.13 Cellular component protein classification . . . 37

3.1 Fuzzy membership function . . . 41
3.2 Fuzzy C-Means pseudocode . . . 46

4.1 Workflow . . . 67
4.2 Pseudocode for the genetic algorithm . . . 81

5.1 Consensus sequence for TFBSs . . . 92
5.2 Motif representations. . . . 93
5.3 Recognition of random PFMs columns . . . 100
5.4 Logos for the bZIP EBP family and its corresponding FBP . . . 101
5.5 Logos for the muscle-specific and independent motifs . . . 104

6.1 FISim pseudocode . . . 112

6.2 JASPAR statistics . . . 116
6.3 Random Motifs . . . 117
6.4 Case study . . . 119
6.5 Related motifs . . . 120
6.6 ROC curves . . . 120
6.7 Pairs of seeds of related motifs . . . 121
6.8 Clusters obtained by kcmeans for the Jaspar motifs . . . 123
6.9 REL group retrieved by kcmeans . . . 125

7.1 Case Study. Motif MZF1 . . . 141
7.2 ROC curves for the the synthetic sequences . . . 143
7.3 Precision-recall and F-measure graphs for the synthetic sequences 144
7.4 ROC curves for the mutated sequences . . . 146
7.5 Precision-recall and F-measure graphs for the mutated sequences . 147
7.6 Average false-positive ratio per TF for different thresholds . . . 148
7.7 Putative binding sequences for the G allele of the TNFR1$_{\text{-609(G/T)}}$. 151
7.8 Discarded TFs . . . 153
7.9 ICSBP against TNFR1$_{\text{-609(G/T)}}$ polymorphism . . . 154

.1 Cluster 1 . . . 183
.2 Cluster 2 . . . 183
.3 Cluster 3 . . . 184
.4 Cluster 4 . . . 184
.5 Cluster 5 . . . 184
.6 Cluster 6 . . . 185
.7 Cluster 7 . . . 185
.8 Cluster 8 . . . 185
.9 Cluster 9 . . . 186
.10 Cluster 10 . . . 186
.11 Cluster 11 . . . 186
.12 Cluster 12 . . . 187
.13 Cluster 13 . . . 187
.14 Cluster 14 . . . 187
.15 Cluster 15 . . . 188

.16 MDSCAN-1 189
.17 MDSCAN-2 189
.18 MDSCAN-3 190
.19 MDSCAN-4 190
.20 MDSCAN-5 190
.21 MDSCAN-6 191
.22 MDSCAN-7 191
.23 MEME-1 191
.24 MEME-2 191
.25 MEME-3 192
.26 MEME-4 192
.27 MEME-5 192
.28 MEME-6 193
.29 Weeder-1 193
.30 Weeder-2 193
.31 Weeder-3 194
.32 Weeder-4 194
.33 Weeder-5 194

List of Tables

2.1 Types of data used in bioinformatics 28
2.2 Databases used in bioinformatics 29
2.3 GO Evidence Codes 35

4.1 Jaccard scores for *C4* 69
4.2 Jaccard scores for *M4* 70
4.3 Jaccard scores for *E10* 71
4.4 Jaccard scores for *P17* 72
4.5 Jaccard scores for *H7* 73
4.6 Best clusters for experiment *C4* 77
4.7 Proteins misplaced for experiment *C4* 78
4.8 Evidence Weights 84
4.9 Semantic similarity measures over BP and CC 85
4.10 GO Semantic Measures Performance. 86

5.1 IUPAC codes for extended consensus sequences 92
5.2 PFM toy example 98
5.3 Jaspar family distribution 101
5.4 Computed scores for Jaspar families and their FBPs 102
5.5 Measures for muscle-specific VS independent PFMs. 103

6.1 FISim example 114
6.2 Area Under the Curve scores for the related motifs experiment . . 121
6.3 JASPAR family distribution 122
6.4 Co-regulated genes 126
6.5 Best JASPAR matches for the MDSCAN algorithm 126
6.6 Best JASPAR matches for the MEME algorithm 127

6.7 Best JASPAR matches for the Weeder algorithm 128

7.1 AUC values for the synthetic and mutated sequence experiments . 144
7.2 SC_{intuit} scores for the two alleles 152

Agradecimientos

Muchas personas que se han cruzado en mi vida durante estos últimos años han contribuido de una u otra forma a la consecución de esta tesis doctoral. A todas ellas me gustaría darles las gracias, aunque hay cosas que no se pueden agradecer con unas simples líneas.

En primer lugar me gustaría agradecer a mi director de tesis, Armando Blanco, por su continuo apoyo, paciencia, y habilidad para guiarme por lo que ha sido un nuevo mundo para mí. Sin duda, ha sabido lidiar con la no siempre fácil tarea de dirigirme, y ha tenido una gran influencia en mi formación científica.

También quiero agradecer su ayuda a Carlos Cano y Francisco Javier López, mis dos compañeros de fatigas, que también presentan sus tesis en estas fechas. Parece mentira pero lo hemos conseguido. Atrás quedan esos primeros meses desamparados por aulas clandestinas, y hemos logrado objetivos que ni siquiera nos planteábamos en nuestros comienzos. A Carlos le doy las gracias por su acertadas correcciones, optimismo, y bondad que siempre consiguen hacerte ver el vaso un poco más lleno. A Javi por estar siempre ahí, desde que empezamos con la carrera, y ser bastante más que un magnífico compañero de trabajo. Muy buenos momentos hemos vivido juntos, y los que quedan. Estoy seguro de que un gran futuro os espera.

También quisiera mostrar mi agradecimiento a otras personas que me han ayudado a lo largo de estos años. A Fernando Bobillo por preocuparse por mis asuntos y mantenerme siempre informado de todos los asuntos burocráticos. A Luis Adarve por ese año que compartimos codo con codo, probablemente ha sido la persona que más me ha enseñado en mi vida. A Pedro Magaña

y Mariló Ruiz por su amistad y los innumerables almuerzos y cafés compartidos. Y también al resto de sufridos becarios del departamento.

No quisiera olvidarme de Shankar Subramaniam ni de Adrian Shepherd, que tan amablemente me acogieron en sus grupos en San Diego y en Londres. Gracias por permitirme vivir esas experiencias y por abrirme puertas que ni siquiera sabía que existían. También me gustaría darles las gracias porque se lo merecen a Pablo Bueno, Ángel Concha, Marta Cuadros, Juan Saínz, Javier García-Castro, y a muchos otros del ámbito de ciencias de la salud que tanto me han enseñado.

Desde luego, esta tesis no hubiera podido llevarse a cabo sin la ayuda que constantemente me ha brindado mi familia. Me gustaría agradecer a mis padres simplemente por ser como son, un ejemplo para mí. Todo es mucho más fácil cuando sabes que tienes detrás unos orgullosos padres que te apoyan y te apoyarán pase lo que pase. También a mi hermano, Pablo, que, aunque parezca difícil, es y será más que simplemente un hermano. Ha sido mi amigo, confidente, y muchas otras cosas más. Y que así siga. También agradecer especialmente a mi familia burgalesa por ofrecerme tantas cosas. Especialmente a mi tío, Pablo Alcalde, por sus enseñanzas y porque de él aprendí que siempre se puede un poco más.

Me gustaría agradecer a las personas que no han sido aún mencionados que me han dado la estabilidad emocional que necesita toda persona. Muy especialmente a mi novia, Ana, por estar siempre ahí apoyando, y aguantarme y quererme y hacerme feliz todo el tiempo, ya sea en la cercanía o en la distancia. También quiero recordar a mis *amigos invisibles*, gracias por todos estos años.

Resumen

Este resumen contiene una versión en español del Capítulo 1, y ha sido incluido para cumplir con los requerimientos necesarios para poder optar a la mención de *Doctorado Europeo*.

Motivación

En la última década la bioinformática se ha convertido en una parte integral de la investigación y el desarrollo en las ciencias biomédicas. La bioinformática tiene ahora un papel esencial en el desciframiento de datos genómicos, transcriptómicos y proteómicos generados por tecnologías experimentales de alto rendimiento, y en la organización de la información obtenida por la biología tradicional. Los métodos de análisis de secuencias de genes o proteínas han evolucionado y mejorado, desarrollándose nuevos métodos para el análisis de un gran número de genes o proteínas simultáneamente, así como para la identificación de grupos de genes relacionados y redes de interacción de proteínas. Con la secuenciación de los genomas de un número cada vez más alto de organismos, la bioinformática está comenzando a ofrecer tanto las bases conceptuales como los métodos prácticos para la detección de conductas funcionales sistémicas de la célula y el organismo.

La bioinformática es, por tanto, el campo de la ciencia donde la biología, la informática, y la tecnología de la información se unen para formar una única disciplina, con el objetivo de ayudar en el descubrimiento de nuevos datos biológicos. De esta forma, la comprensión de los principios biológicos que afectan a los organismos vivos es clave para el desarrollo de métodos bioinformáticos apropiados.

A lo largo de la historia de la ciencia, siempre ha existido la necesidad de modelar y gestionar la incertidumbre existente en los experimentos reales. Esto es particularmente cierto en la biología en general, y más recientemente en la bioinformática. La variabilidad exhibida por la naturaleza al estudiar el genoma y sus relaciones requieren modelos computacionales lo suficientemente flexibles para capturar lo esencial, sin tener en cuenta todas las variabilidades como algo completamente nuevo. La teoría difusa (Zadeh, 1965) es una potente herramienta que ha servido a los investigadores para el modelo de situaciones donde la principal fuente de incertidumbre es la aleatoriedad. En algunos casos, la incertidumbre puede adoptar otras formas. Al considerar una secuencia nueva de un gen, puede ser de interés conocer cómo de similar es a otra secuencia en particular. No se trata de el clásico problema *binario* de saber si dos secuencias son iguales o no, si no de saber *cuánto* se asemejan estas dos secuencias. Otras fuentes de incertidumbre incluyen: omisiones en los datos extraídos de muestras reales, la falta de expresividad o de confianza en algunas características extraídas, la falta de límites claros entre las distintas clases de proteínas, genes o productos de los genes que son miembros de más de una clase, etc.

Además, hasta la fecha casi todos los problemas bioinformáticos se han formulado de un modo determinista. La mayoría de estos problemas son definidos fijando y optimizando funciones objetivo. Sin embargo, existen diversas situaciones en las que se hace necesario considerar la vaguedad de los datos. Por ejemplo, la imprecisión que acompaña intrínsecamente a los sistemas biológicos, las múltiples funciones que una entidad biológica puede desarrollar, las descripciones difusas de algunos fenómenos biológicos, etc. Esta tesis tiene como objetivo principal resolver algunos importantes problemas bioinformáticos mediante la aplicación de nociones sobre la teoría de conjuntos difusos y lógica difusa, así como sobre otros métodos de *soft computing*.

Debido a la reciente explosión de nuevos datos biológicos procedentes de la secuenciación de genomas, los científicos se enfrentan al problema de responder a muchas preguntas básicas, tratando de extraer información de estos datos. Una de las principales tareas es la de descubrir la función a la que

los genes están asociados. Recientemente, la bio-ontologías han desempeñado un papel importante para la integración automática de conocimiento, lo que es fundamental para apoyar la generación y validación de hipótesis sobre las funciones de los productos de los genes. En este sentido, la *Gene Ontology*[1] (GO) (Ashburner et~al., 2000) se ha convertido en un estándar *de facto* para describir los productos de genes. Fue creada con el objetivo de normalizar la representación de los genes y productos de genes procedentes de distintas especies y bases de datos. Así, proporciona un vocabulario estructurado y controlado para describir las funciones de los genes y los productos de genes en cualquier organismo. Existen algunos trabajos que utilizan GO para extraer la información biológica de grupos de productos de genes potencialmente relacionados. Por lo general, estos enfoques se basan en medidas definidas para una ontología genérica que son adaptadas a las características específicas de GO (Resnik, 1995; Jiang and Conrath, 1997; Lin, 1998). Sin embargo, pocos métodos basados en tecnología difusa están disponibles actualmente. En nuestra opinión, las propiedades de la teoría de conjuntos difusos la hacen interesante para su aplicación a este problema.

Por otro lado, las células controlan la abundancia de proteínas por medio de diversos mecanismos. Uno de esos mecanismos es la regulación de la transcripción, que es un proceso continuo en el que muchos factores se combinan para garantizar una adecuada tasa de síntesis de proteínas. La comprensión de estos complejos procesos es uno de los principales objetivos de la biología computacional. Los factores de transcripción (TFs, del inglés *transcription factors*) desempeñan un papel clave en la regulación de los genes, mediante su unión a secuencias específicas, llamadas sitios de unión de factores de transcripción (TFBSs, del inglés *transcription factor binding sites*). Aquellas secuencias de ADN donde se puede producir la unión del mismo TF se se agrupan conjuntamente formando motivos, que se representan normalmente como matrices de frecuencias por posición (PFMs, del inglés *position frequency matrices*). La predicción *in silico* de la posible unión de un TF a un TFBS es un problema ampliamente estudiado por la biología computacional.

[1] Ontología de Genes en español

Un punto importante en el contexto del descubrimiento *de novo* de motivos consiste en saber si, dados los candidatos a motivos recientemente obtenidos, éstos se asemejan a otros motivos previamente descritos en las bases de datos existentes. Por ésta y otras razones se han desarrollado varias medidas de comparación entre motivos. La mayoría de los métodos propuestos están basados en técnicas estadísticas que comprueban si las diferentes columnas de los motivos pertenecen a la misma distribución (Pietrokovski, 1996; Schones et~al., 2005; Wang and Stormo, 2003). Otros trabajos más recientes proponen el uso de métodos más específicos que mejoran a los enfoques probabilísticos (Gupta et~al., 2007; Pape et~al., 2008). Sin embargo, en el contexto de la comparaciones entre motivos, la utilización de PFMs como representación de las preferencias de unión de los TFs incluye imprecisión. Además, los métodos actuales no están diseñados para tener en cuenta la mayor aportación de las posiciones mejor conservadas de los motivos a la fuerza de la unión secuencia-motivo. Por lo tanto, nuevos métodos que tengan en cuenta este tipo de problemas son necesarios.

Del mismo modo, el descubrimiento de patrones en secuencias de ADN es una de los problemas más importante de la bioinformática, con aplicaciones en la búsqueda de elementos de regulación y TFBSs. Una importante tarea en este problema es la búsqueda (o predicción) de sitios de unión conocidos en una nueva secuencia de ADN. La mayoría de las herramientas disponibles para la predicción de TFBSs asumen independencia entre las posiciones de las bases de los sitios de unión (Hertz et~al., 1990; Sandelin et~al., 2004b). Algunos trabajos recientes están empezando a considerar las dependencias entre posiciones (Tomovic and Oakeley, 2007; Zare-Mirakabad et~al., 2009). Uno de los objetivos principales en la predicción de TFBSs es reducir la tasa de falsos positivos sin comprometer la sensibilidad de los resultados. Los métodos que tienen en cuenta las dependencias entre posiciones tienden a ser significativamente más eficaces. Sin embargo, algunas cuestiones como el sobreaprendizaje de las secuencias de prueba, o la selección de un umbral arbitrario para detectar las dependencias entre posiciones, siguen abiertas y necesitan nuevos enfoques para su resolución.

Objetivos

El objetivo general de esta tesis es encontrar soluciones basadas en la tecnología difusa para algunos problemas bioinformáticos importantes, gestionando así la incertidumbre asociada a los procesos biológicos. Más concretamente, nos centramos en el estudio de las medidas de similitud semántica para GO, las comparaciones de motivos de ADN, y la cuantificación de la afinidad secuencia-motivo.

De acuerdo a esto, los objetivos específicos de esta tesis son los siguientes:

- Analizar las propiedades de GO y revisar el estado del arte de las medidas semánticas descritas sobre GO.
  - Comparar las medidas semánticas *crisp*[2] sobre GO y analizar sus limitaciones.
  - Aplicar diferentes métodos de agrupamiento y comparar su utilidad para el reconocimiento de familias de proteínas.
- Proponer una nueva medida de similitud semántica difusa para GO.
  - Incorporar a la medida los códigos de evidencia de las anotaciones para tener en cuenta la fiabilidad de la fuente de información.
  - Comparar la nueva medida con las medidas existentes en problemas de clasificación de proteínas.
- Revisar el estado del arte de las medidas de comparación entre motivos y examinar la adecuación de enfoques difusos para dicha tarea.
  - Adaptar medidas difusas clásicas al problema de la comparación de motivos.
  - Comparar las medidas difusas con otros enfoques relacionados en problemas de detección de motivos.
- Proponer una nueva medida de similitud entre motivos basada en la integral difusa, diseñada teniendo en cuenta la distinta importancia de las posiciones de los motivos en función de su contenido de información

[2] *crisp* es una palabra en inglés que significa lo contrario que difuso, se usa aquí al no existir una palabra en español para dicho concepto.

  - Revisar medidas entre motivos recientes y analizar sus inconvenientes.
  - Definir la nueva medida y demostrar su mejor funcionamiento en experimentos tanto sintéticos como reales.
- Proponer un nuevo método basado en tecnología difusa para cuantificar la afinidad secuencia-motivo.
  - Discutir los últimos avances en esta materia.
  - Mejorar la calidad de la predicción de TFs de los enfoques existentes.
  - Aplicar el nuevo método a problemas biológicos reales.

Contenidos

Esta memoria se estructura en cuatro partes bien diferenciadas, cada una de las cuales se compone de uno o más capítulos.

La Parte I contiene el Capítulo 1 que incluye una introducción en la que, partiendo de los antecedentes en el área, se motiva nuestro trabajo, se establecen los objetivos de la tesis y se describe el contenido del documento.

A continuación, la Parte II expone los conocimientos preliminares necesarios para una mejor comprensión del texto. El Capítulo 2 presenta algunos conceptos básicos de biología, así como proporciona una revisión de la bioinformática, el campo de investigación donde se enmarca esta tesis. El Capítulo 3 introduce algunas nociones básicas sobre la teoría de conjuntos difusos, la lógica difusa y otros métodos de *soft computing*.

La Parte III presenta las contribuciones de esta tesis. El Capítulo 4 define una nueva medida de similitud difusa para GO, e investiga el funcionamiento de ésta y otras medidas semánticas para GO en conjunto con distintos métodos de agrupamiento en problemas de clasificación de proteínas. El Capítulo 5 expone el problema de comparación entre motivos de TFBSs, y muestra cómo se pueden adaptar para esta tarea distintas medidas difusas clásicas. El Capítulo 6 introduce los avances más recientes de las medidas entre motivos y presenta una nueva medida de similitud para motivos de ADN basada en

la integral difusa llamada FISim. Además, propone una nueva metodología de agrupamiento basada en dicha medida y en métodos de *kernelización*. El capítulo termina con una evaluación de nuestras propuestas en comparación con los mejores métodos existentes. El Capítulo 7 aborda el problema de búsqueda (o predicción) de sitios de unión ya conocidos en una secuencia de ADN, proponiendo una nueva medida de afinidad motivo-secuencia basada en la teoría de conjuntos difusos intuicionista.

Finalmente, la Parte IV, concluye esta tesis. El Capítulo 8 resume las contribuciones de esta tesis. Los resultados se analizan de acuerdo con los objetivos anteriormente establecidos. Además, se apuntan algunas ideas para el trabajo futuro.

Part I

Introduction

CHAPTER 1

Introduction

1.1 Antecedents

In the past decade, bioinformatics has become an integral part of research and development in the biomedical sciences. Bioinformatics now has an essential role both in deciphering genomic, transcriptomic and proteomic data generated by high-throughput experimental technologies and in organizing information gathered from traditional biology. Sequence-based methods of analyzing individual genes or proteins have been elaborated and expanded, and methods have been developed for analyzing large numbers of genes or proteins simultaneously, such as in the identification of clusters of related genes and networks of interacting proteins. With the complete genome sequences for an increasing number of organisms at hand, bioinformatics is beginning to provide both conceptual bases and practical methods for detecting systemic functional behaviors of the cell and the organism. In turn, bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline, aiming to help on the discovery of new biological insights. Therefore, understanding the biological principles that affect to the living organisms is a key point for developing appropriate bioinformatics methods.

Throughout the history of science, there has always been a need to model and manage uncertainty in real experiments. This is particularly true in biol-

ogy in general, and more recently in bioinformatics. The variability exhibited in nature in studying the genome and its relationships with very different aspects require computational models to be flexible enough to capture the essential features without taking every deviation as something completely new. Fuzzy theory (Zadeh, 1965) is a powerful tool that has served researchers to model situations where the primary source of uncertainty is randomness. In some cases, uncertainty adopt other forms. In considering a new gene sequence, it may be important to know how similar it is to a particular sequence. It is not the classical *binary* problem of knowing whether or not two sequences are equal, it is a question of *how much* this particular instance of the new gene resembles a prototype. Other sources of uncertainty include incompleteness in the data extracted from actual samples, lack of expressiveness or faithfulness of some features that we extract, lack of clear boundaries between classes of proteins, genes or gene products that are member of more than one class, etc.

In addition, almost all bioinformatics problem to date are formulated in a deterministic manner. Most of these problems are defined by fixed objective functions and solve by means of optimization. However, there are several situations where fuzziness should be considered, e.g. intrinsic fuzziness in biological systems, multiple roles of a biological object, fuzzy descriptions of biological phenomena, etc. This thesis aims to solve some important bioinformatics problems applying notions on fuzzy set theory and fuzzy logic, as well as on other soft computing methods.

Due to the recent flood of new biological data from genome sequencing, scientists need to face the problem of answering many basic questions and attempting to extract information from this data. One of the main tasks is to discover to which function the genes are associated. Recently bio-ontologies have played an important role for the automatic integration of background knowledge which is fundamental to support the generation and validation of hypotheses about the function of gene products. In this sense, the Gene Ontology (GO) (Ashburner et al., 2000) has become a *de facto* standard for describing gene products in databases. It was created with the aim of standardizing the representation of gene and gene product attributes across

species and databases. Thus, it provides a structured, controlled vocabulary for describing the roles of genes and gene products in any organism. There exist some works that use GO to extract biological information from groups of potentially related gene products. Usually, these approaches are derived from general ontology measures that are adapted to the GO specific characteristics (Resnik, 1995; Jiang and Conrath, 1997; Lin, 1998). However, few fuzzy-based methods are currently available. In our opinion, the properties of fuzzy set theory make interesting its application to this problem.

On a different topic, cells control the abundance of proteins by means of diverse mechanisms. One such mechanism is the regulation of transcription, which is a continuous process whereby many factors combine to ensure appropriate rates of protein synthesis. Understanding such complex processes is one of the main objectives in computational biology. Transcription factors (TFs) play a key role in gene regulation by binding to target sequences called transcription factor binding sites (TFBSs). Related DNA sequences to which the same TFs can bind are grouped together into TFBS motifs, usually represented as position frequency matrices (PFMs). *In silico* prediction of potential binding of a TF to a binding site is a well-studied problem in computational biology.

A common question in the context of *de novo* motif discovery is whether a newly discovered, putative motif resembles any previously discovered motif in an existing database. For this matter several motif comparison measures have been proposed, most of them based on statistical techniques that test whether the different columns belong to the same distribution (Pietrokovski, 1996; Schones et al., 2005; Wang and Stormo, 2003). Other recent approaches propose more specific methods that outperforms probabilistic approaches (Gupta et al., 2007; Pape et al., 2008). However, in the context of motif comparisons, the utilization of PFMs as a representation of the binding preferences of the TFs inherently includes imprecision. In addition, existing methods are not designed to consider the higher contribution of better conserved positions to the binding affinity. Therefore, new methods that deal with these kinds of problems are needed.

Likewise, pattern discovery in DNA sequences is one of the most impor-

tant problems in bioinformatics with applications in finding regulatory elements and transcription factor binding sites. An important task in this problem is to search (or predict) known binding sites in a new DNA sequence. Most of the available tools for transcription factor binding site prediction assume sequence independence between the binding site base positions (Hertz et al., 1990; Sandelin et al., 2004b). New approaches are starting to consider position dependencies, (Tomovic and Oakeley, 2007; Zare-Mirakabad et al., 2009). One of the main goals in the prediction of TFBSs is to reduce the false positive rate without compromising sensitivity. Methods that take into account positional dependencies tend to be significantly more effective at meeting this challenge. However some issues like overlearning of the training data, or the arbitrary threshold selection for testing dependencies remain unsolved.

1.2 Objectives

The general aim of this dissertation is to find fuzzy-based solutions for important bioinformatics problems in order to manage the uncertainty associated to biological processes. More precisely, we focus on the study of semantic similarity measures for GO, DNA motifs comparisons, and quantifying sequence-motif affinity.

In accordance to this, the concretes objectives of this thesis are the following:

- To analyze GO properties and review the state of the art in GO semantic measures.
  - To compare GO crisp semantic measures and to analyze their limitations.
  - To apply different cluster methods and to compare their performance for protein family recognition.
- To propose a new fuzzy similarity measure for GO.
  - To incorporate the evidence codes of the annotations to take into account the reliability of the source of information.

  - To compare the new measure with the state of the art measures in terms of protein classification.
- To review the state of the art in motif comparison measures and to examine the adequacy of fuzzy approaches for this task.
  - To adapt classical fuzzy measures for the problem of motif comparison.
  - To compare the fuzzy measures with respect to related approaches in motif detection problems.
- To propose a novel motif similarity measure based on the fuzzy integral that outperforms the existing approaches.
  - To review recent motif measures and to analyze their drawbacks.
  - To define the new measure and to prove its superior performance in real and synthetic experiments.
- To propose a new method based on the IFS theory for the problem of scoring DNA sequences against TFBS motifs.
  - To discuss the latest advances on this topic.
  - To improve the prediction quality for TFs of the existing approaches.
  - To apply the new score to real research problems.

1.3 Thesis Structure

This dissertation is structured in four clearly defined parts, each of them being composed of one or more chapters.

The first part comprises this introduction, which has been written also in Spanish in order to fulfil the requirements to obtain the *European Doctorate* mention (page xiii).

Part II starts with some necessary preliminaries. Chapter 2 is dedicated to present some basic concepts of biology and to overview bioinformatics, the research field of this thesis. Special attention is dedicated to introduce GO. Chapter 3 reviews some basic notions on fuzzy set theory and fuzzy logic, as well as on other soft computing methods.

Part III presents the contributions of this thesis. Chapter 4 defines a novel fuzzy semantic similarity measure over GO, and investigates the performance of different ontology semantic measures over GO and clustering methods in protein classification problems. Chapter 5 introduces the problem of comparing TFBS motifs and shows how some classical fuzzy measures can be adapted for this problem. Then, Chapter 6 introduces the recent advances in motif measures and presents a new similarity measure for DNA motifs called FISim (Fuzzy Integral Similarity) together with a novel clustering methodology for motifs. An evaluation of our proposed approaches is also performed. Chapter 7 approaches the problem of searching (or predicting) known binding sites in a new DNA sequence and presents a new scoring function based on intuitionistic fuzzy set theory.

Finally, Part IV concludes with some conclusions and future work. Chapter 8 summarizes the contributions of this thesis. The results are analyzed in accordance with the objectives established in this introductory chapter (Section 1.2). In addition, some ideas for future research are finally pointed out.

Part II

Preliminaries

CHAPTER 2

Biology and Bioinformatics

All organisms consist of small cells. Each cell is a complex system consisting of many different building blocks enclosed in a membrane bag and share a common machinery for their basic functions. The exterior appearance of living organisms is infinitely diverse, however they are very similar on the inside. Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline, aiming to help on the discovery of new biological insights. Therefore, understanding the biological principles that affect to the living organisms is a key point for developing appropriate bioinformatics methods.

Thus, in this first chapter we provide in Section 2.1 a basic introduction to some biological concepts. Section 2.2 gives an overview of bioinformatics, which constitutes the research field of this thesis. Finally, Section 2.3 provides an introduction of Gene Ontology together with its main applications.

2.1 Biology

In this section the universal characteristics of all the living organisms are outlined. We briefly discuss the cellular diversity, and show how starting from a common code, shared by all the organisms, it is possible to read, measure, and disembowel the specifications of such code.

Cells, DNAs, and Chromosomes

The diversification and evolution of the cells has been taking place for more than 3.5 billion years (Berg et al., 2002). Each cell is a complex automaton capable of generating new cells which are self-sustaining and self-replicating. Figure 2.1 shows the structure of the cell. All cells with no known exceptions, store their hereditary information in DNA (Deoxyribonucleic acid) molecules.

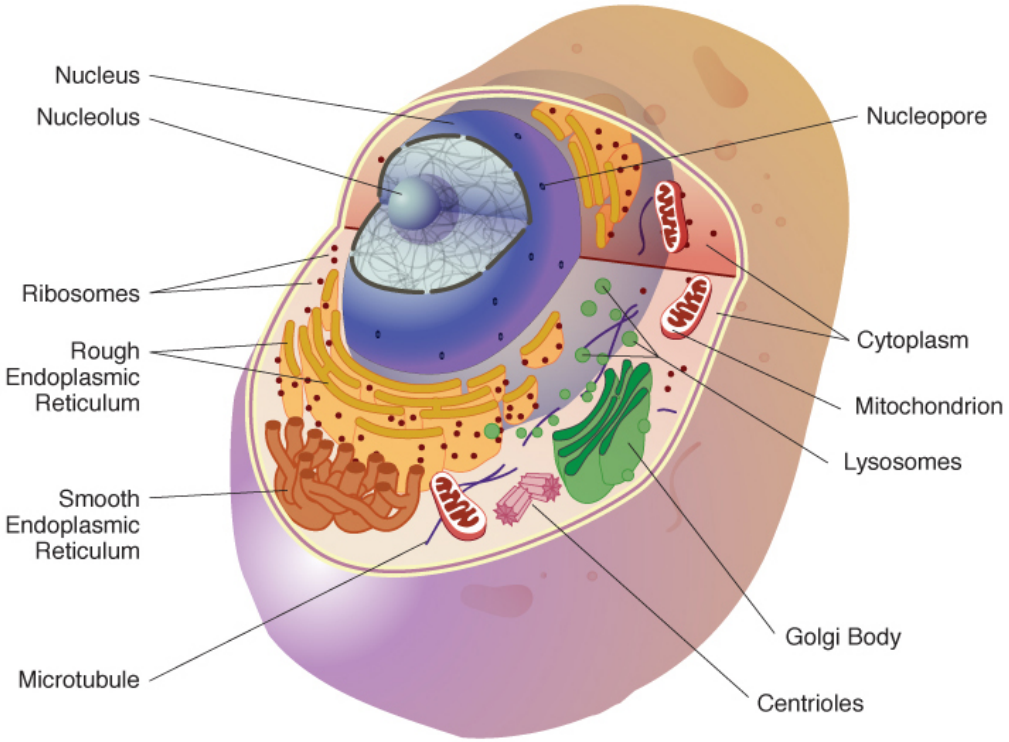

Figure 2.1: ***Cell structure.*** *(National Human Genome Research Institute.)*

Chemically, a DNA molecule is a long polymer of simple units called nucleotides, with backbones made of sugars and phosphate groups joined by ester bonds. In living organisms, DNA does not usually exist as a single molecule, but instead as a pair of molecules that are held tightly together. These two long strands entwine, in the shape of a double helix, which was first described by Watson and Crick (1953).

In order to understand the biological mechanisms, we first need to know the double helix structure of the DNA molecules. The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). As an example, human DNA consists of about 3 billion bases. Each base is also attached to a sugar molecule and a phosphate molecule. Together, a base, sugar, and phosphate are called a nucleotide (Figure 2.2). Nucleotides are arranged in two long strands that form a spiral called a double helix. DNA bases pair up with each other, A with T and C with G, to form units called base pairs, and the paired bases are said to be complementary (Figure 2.3). The structure of the double helix is somewhat like a ladder, with the base pairs forming the ladder's rungs and the sugar and phosphate molecules forming the vertical sidepieces of the ladder (Figure 2.4).

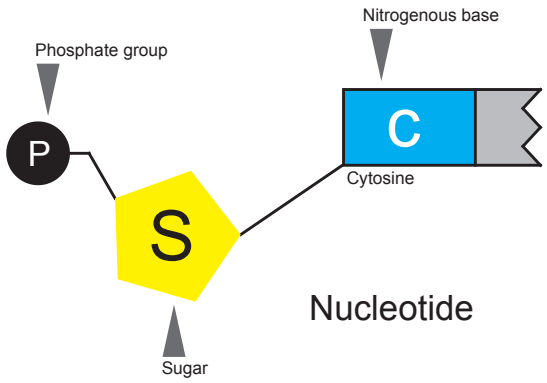

Figure 2.2: ***Nucleotide.***

This model states that the two polynucleotide chains run in opposite directions (antiparallel). The bases lie on the inside. They are flat structures, lying in pairs perpendicular to the axis of the helix. Each base pair is rotated $\sim 36^{\circ}$ around the axis of the helix relative to the next base pair. So ~ 10 base

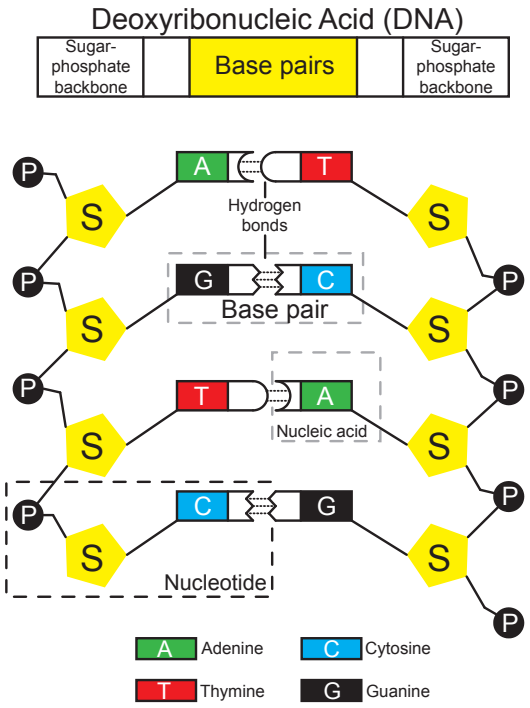

Figure 2.3: ***DNA.***

pairs make a complete tum of 360°. The twisting of the two strands around one another forms a double helix with a minor groove (∼ 12 Å across) and a major groove (∼ 22 Å across). These features represent the accepted model for what is known as the *B-form* of DNA.

It is important to realize that the B-form represents an average, not a precisely specified structure. DNA structure can change locally. If it has more base pairs per turn it is said to be overwound; if it has fewer base pairs per turn it is underwound. Local winding can be affected by the overall conformation of the DNA double helix in space or by the binding of proteins to specific sites.

Typically, the DNA in a cell is found not in one but in several physically separate molecules called *chromosomes*. While different species may have different number of chromosomes, the specific arrangement among all member in the same species is always consistent, and is called karyotype. Any

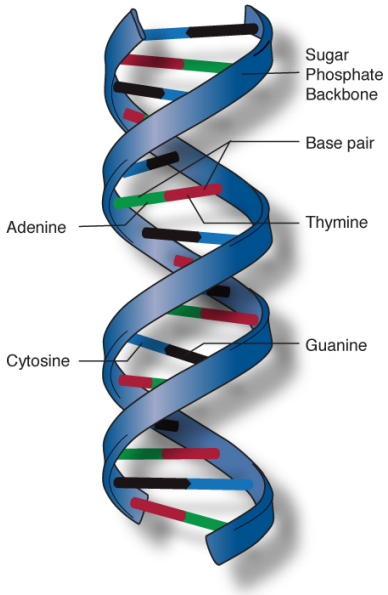

Figure 2.4: ***DNA double helix.*** *(National Human Genome Research Institute.)*

aberration from the default chromosomal arrangement is often lethal or lead to serious genetic disorders. A well-known chromosomal disease in humans is the Down's Syndrome, in which an extra copy of one of the chromosomes causes mental retardation and other associated problems.

The chromosomes are usually organized in homologous pairs, each chromosome pair containing one chromosome from each parent. In humans there are 23 pairs of homologous chromosomes ranging in length from about 50 million to 250 million base pairs. Figure 2.5 shows a photograph of the 23 pairs of human chromosomes. Collectively, the genetic information in the chromosomes are called the *genome*. As each cell divides, the entire genome in the DNA is copied exactly in the new cells. Therefore, in theory, any of the cell on our body possesses the necessary information for building a complex living organism as ourselves.

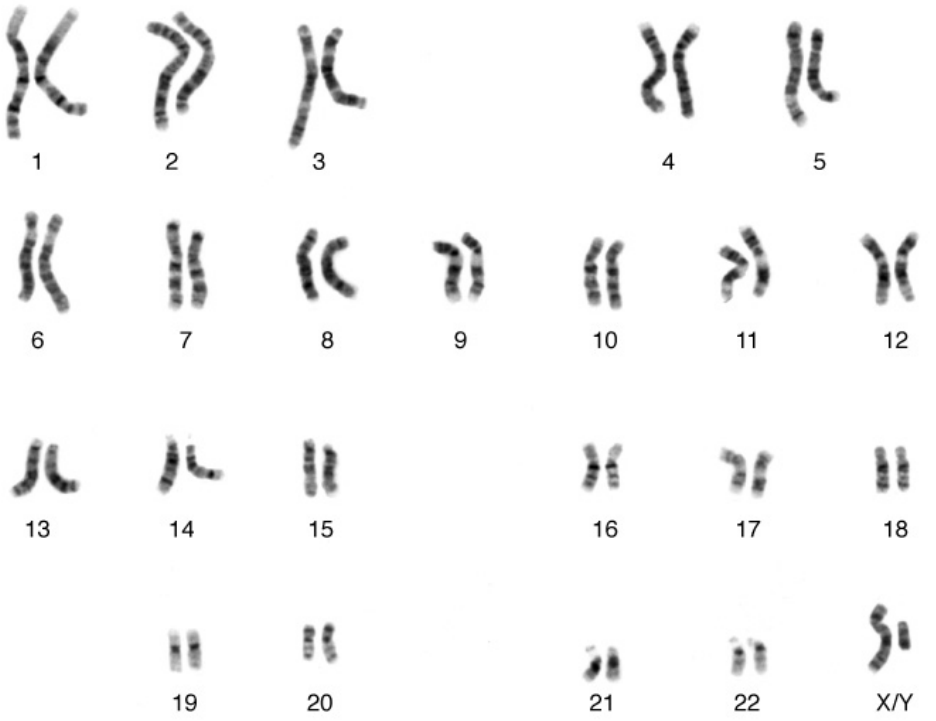

Figure 2.5: ***Human chromosomes.*** *(National Human Genome Research Institute.)*

From Gene to Protein: The Central Dogma

DNA contains the genetic recipes for making proteins, which are the actual workhorses that perform most most life functions. In this section we introduce how a DNA sequence turns into a chain of amino acids and forms a protein in the cell.

For a cell to make a protein, the information from a gene recipe is first copied (base by base) from a strand of DNA in the cell's nucleus into a strand of messenger RNA (mRNA). Chemically, the RNA, or ribonucleic acid, and the DNA are very similar. RNA molecules are also made up of four different nucleotides (A,C,G,U), the nucleotide U (uracil) replaces the T (thymine) in DNA. Like thymine, the uracil also base-pairs to adenine.

After copying the genetic recipes on the DNA in the nucleus, the mRNA molecules then travel out into the cytoplasm and becomes accessible to the

cell organelles called ribosomes. Here, each ribosome molecule reads the specific genetic code on an mRNA, and translates the genetic code into the corresponding amino acid sequence based on a genetic coding scheme. With the help of transfer RNA (tRNA), molecules that transport different amino acids in the cell to the ribosome molecule as needed, the prescribed protein molecule is assembled (amino acid by amino acid) as instructed by the genetic recipe. Figure 2.6 illustrates how information stored in DNA is ultimately transferred to protein in the cell.

Figure 2.7 provides a schematic view of the relationship between DNA, RNA and protein in terms of three major processes:

- **Replication.** Process by which the information in the DNA molecule in one cell is passed on to new cells as the cell divides and the organism grows. Entire genetic blueprint can be passed on from cell to cell through DNA replication. In this way, virtually, all cells in our body have the full set of recipes for making all the proteins necessary to sustain life's many different functions.
- **Transcription.** Process by which the relevant information encoded in DNA is transferred into the copies of messenger RNA molecules during the synthesis of the messenger RNA molecules. Transcription allows the amount of the corresponding proteins synthesized by the protein factories (the ribosomes) to be regulated by the rate at which the respective mRNAs are synthesized in the nucleus. The study of the regulation of the transcription is one of the main goals in this thesis, therefore, a deeper discussion of this process is provided in the following section.
- **Translation.** Process by which genetic information on the mRNA is transferred into actual proteins. Protein synthesis is carried out by the ribosomes, and it involves translating the genetic code transcribed on the mRNA into a corresponding amino acid string which can then fold into the functional protein.

This multi-step process of transferring genetic information from DNA to RNA to protein is known as the *Central Dogma of Molecular Biology*.

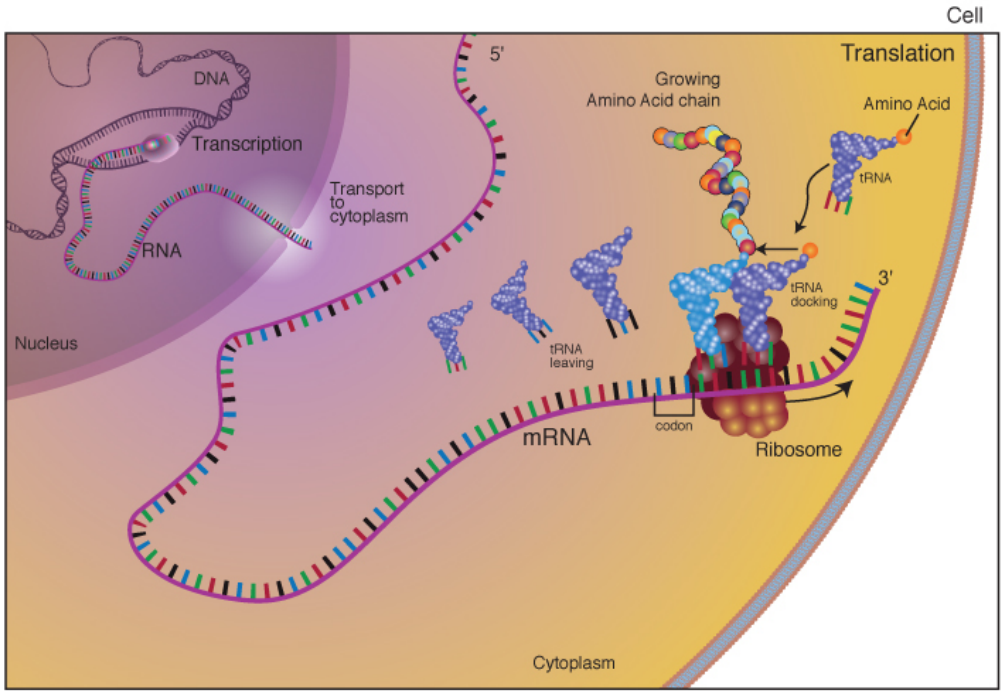

Figure 2.6: ***From gene to protein.*** *(National Human Genome Research Institute.)*

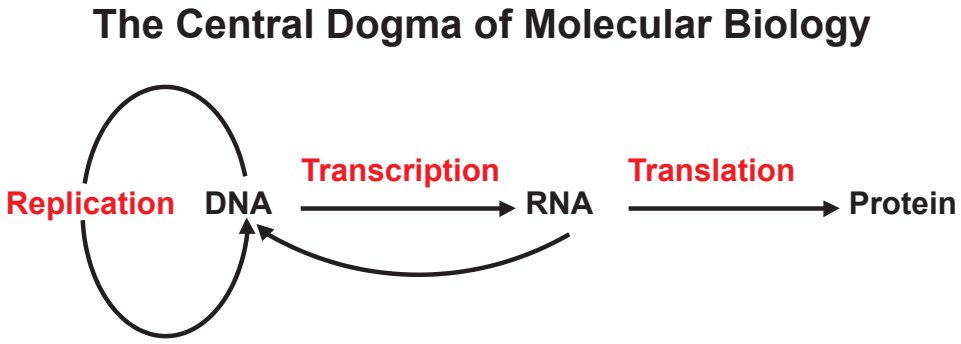

Figure 2.7: ***The Central Dogma of Molecular Biology.***

Transcription

As outlined above, all cells within one organism share the same code for creating them and making them live. This is the general rule for all living

organisms ranging from those composed of just a single cell (e.g., yeast) to more complex creatures such as ourselves. Given that the DNA sequence is known to us, the first level of understanding is where the functional units, called genes, are coded in it. Assuming this is also roughly known via various experimental and computational analysis performed, we focus in this work on a certain type of programs regulating the production of the genes at the transcription level, i.e. when DNA is transcribed into mRNA. This is a necessary step in gene production and much of the living cell's control on its spatial content is believed to be conducted via regulatory programs at this transcriptional stage, hence their importance to us.

Experimental methods developed in recent years have enabled us to simultaneously measure the mRNA levels of thousands of genes under diverse conditions. Other new methods help us by testing thousands of genes to see whether certain proteins called transcription factors (TFs) are involved in regulating their production. These high-throughput experiments, combined with the huge amounts of DNA coding sequence, are actually responsible for developing our field of research, that of computational biology. They pose us the challenge of turning these vast amounts of data into valid biological hypotheses that can later be tested in a lab. We need to develop computational tools that are able to cope with large amount of data that is usually very noisy and partial.

We now give a brief description of the biology behind genetic regulation, followed by some of the experimental methods used to produce high-throughput data.

Transcription involves synthesis of an RNA chain representing strand if a DNA sequence. RNA synthesis is catalyzed by the enzyme *RNA polymerase*. Transcription starts when RNA polymerase binds to a special region, the *promoter*, at the start of the gene. Promoter surrounds the first base pair that is transcribed into RNA (the so-called *startpoint*), and from there, the RNA polymerase moves along the template, synthesizing RNA, until it reaches a terminator sequence. In order for the RNA polymerase to initiate the transcription, it separates the two strands of DNA and uses one of them as a

template to directly synthesize a complementary RNA sequence. The transcription reaction can be divided into four stages:

1. **Template recognition.** It begins with the binding of RNA polymerase to the double-stranded DNA at a promoter. Then the strands are unwound locally to to allow the template strand to be available for base pairing with ribo-nucleotides.
2. **Initiation.** It describes the synthesis of the first nucleotide bonds in RNA. The enzyme remains at the promoter while it synthesizes the first ~ 9 nucleotide bonds. The initiation phase ends when the enzyme succeeds in extending the chain and clears the promoter.
3. **Elongation.** The enzyme moves along the DNA and extends the growing RNA chain. As the enzyme moves, it unwinds the DNA helix to expose a new segment of the template in single-stranded condition.
4. **Termination.** The enzyme recognizes the point at which no further bases should be added to the chain. When the last base is added to the RNA chain, the RNA polymerase is separated from the DNA, and the DNA reforms in duplex state.

The transcription process can be regulated at multiple levels by diverse mechanisms. We can distinguish at least five potential control points:

1. Activation of gene structure.
2. Imitation of transcription.
3. Processing of the transcript.
4. Transport to the cytoplasm.
5. Translation of mRNA.

As it was stated above, in this thesis we focus in the study of the regulation of the transcription in the control point 2 (Initiation of transcription). In fact, this point may be the most important one in the regulation of the production of many genes in eukaryotes (Lewin, 2004).

Initiation of transcription involves many protein-protein and protein-DNA interactions among transcription factors and bound at the promoter or at an enhancer. Figure 2.8 summarizes their properties. The factors required for transcription can be divided into several classes.

- **Basal factors**, together with RNA polymerase, bind at the startpoint.
- **Activators**, transcription factors that recognize specific short consensus elements. They bind to sites in the promoter or in enhancers. They act by increasing the efficiency with which the basal apparatus binds to the promoter. They therefore increase the frequency of transcription, and are required for a promoter to function at an adequate level. These factors are therefore responsible for the control of transcription patterns in time and space.
- **Coactivators**, which provide a connection between activators and the basal apparatus without binding themselves DNA. They work by protein-protein interactions, forming bridges between activators and the basal transcription apparatus.
- **Repressors**, transcription factors that inhibit basal apparatus function. Repression is usually achieved by affecting chromatin structure, but there are repressors that act by binding to specific DNA locations.

2.2 Bioinformatics

What is Bioinformatics?

The beginning of bioinformatics can be traced back to Margaret Dayhoff in 1968 and her collection of protein sequences known as the Atlas of Protein Sequence and Structure (Dayhoff, 1969). Since then, major advances in the field of molecular biology, together with advances in genomic technologies, have led to an exponential growth in the biological information generated by the scientific community (see Figure 2.9 as an example). The scientific community has applied these advances to achieve new goals in diverse research projects such as the well-known Human Genome Project, which aimed to identify all the approximately 20,000-25,000 genes in human DNA and determine the sequences of the 3 billion chemical base pairs that make up human DNA. This torrent of genomic information has, in turn, led to an absolute requirement for computerized databases to store, organize, and index

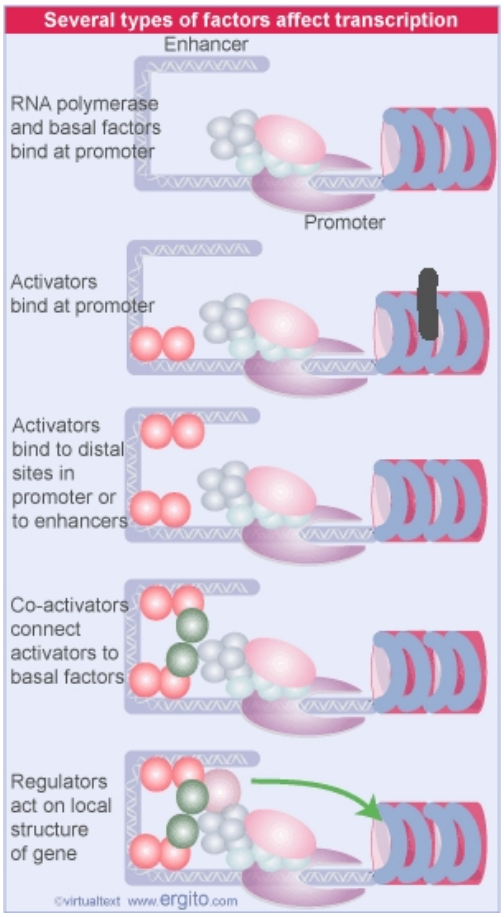

Figure 2.8: ***Transcription factor interactions.***

the data and for specialized tools to view and analyze the data. In this context, bioinformatics emerged as a new discipline combining many scientific fields including computational biology, statistics, mathematics, molecular biology, and genetics.

Definition

The term *bioinformatics* has been recently created. Although, as outlined above, the application of information technologies to biomedical sciences began many years before, the word *bioinformatics* did not start to be used until the last decade. Bioinformatics, as any interdisciplinary field, is open to

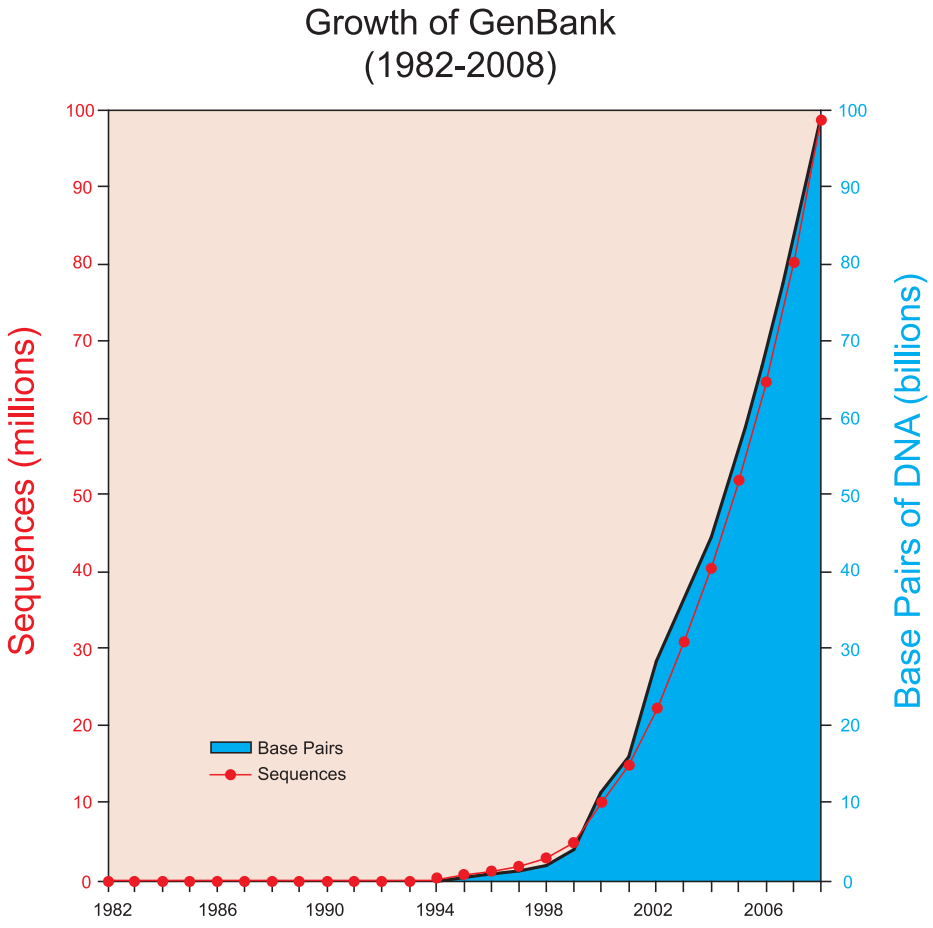

Figure 2.9: ***Growth of GenBank.***

multiple definitions. However, they share a common feature: Bioinformatics is the link between maths, informatics, and biology. Although there is not a standard definition for bioinformatics to date, a commonly accepted the one given by the National Institutes of Health (NIH):

> Bioinformatics is the research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store,organize, archive, analyze, or visualize such data.

In addition, some scientists provide different definitions for *bioinformatics* and *computational biology*. For example the NIH states:

> Computational biology is the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques o the study of biological, behavioral, and social systems.

As can be seen, although the two concepts refer to different disciplines, the line between them is not very clear and they overlap in many aspects. A graphical representation of the context of the bioinformatics discipline can be seen in Figure 2.10.

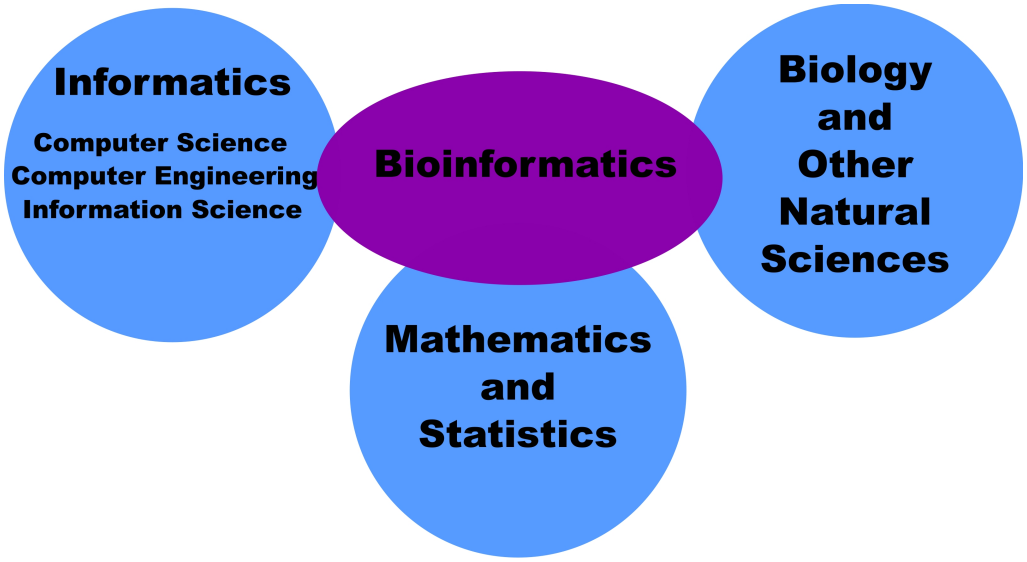

Figure 2.10: ***Bioinformatics as an interdisciplinary field.***

Goals of Bioinformatics

The goals of bioinformatics can be divided in three different groups:

1. Organise data in a way that allows researchers to access existing information and to submit new entries as they are produced, e.g. the Gene Ontology Consortium (Ashburner et al., 2000), where the scientific community annotates their findings related to gene products (more details will be given in Section 2.3). Data-curation was one of

the first duties of bioinformatics and it is still an essential task. However, the information stored in these databases is essentially useless until analysed, mainly due to its enormous size. Thus the purpose of bioinformatics was extended much further.

2. Develop tools and resources for the analysis of biomedical data. For example, having sequenced a genome for a particular organism, it is of interest to compare its genes with those from previously characterized organisms. This needs more than just a simple text-based search and programs such as BLAST (Altschul et al., 1997) must consider what comprises a biologically significant match. For the development of such resources expertise in computational theory as well as a thorough understanding of biology are necessary.

3. Use these tools to analyse the data and provide biologically meaningful interpretations of the results. Traditionally, biological studies examined individual systems in detail, and comparing them with a few that are related afterwards. In bioinformatics, we can conduct global analyses of all the available data with the aim of uncovering common principles that apply across many systems and highlight novel features.

Bioinformatics Data Sources

One of the main features of modern molecular biology is the generation (usually automatized) of extensive amount of data. As it was previously commented, one of the major issues in bioinformatics is to find ways to collect the data and organise them in a meaningful manner. Table 2.1 lists the types of data that are analysed in bioinformatics and the range of topics where they fall within the field. These sources of information can be divided into DNA sequences, protein sequences, macromolecular structures, genome sequences, and other whole genome data. On the other hand, Table 2.2 lists the main bioinformatics databases. Two type of databases can be distinguished: i) primary databases, containing among others DNA and protein sequences, protein structures, and gene and protein expression profiles; ii) secondary databases, containing the data obtained the analysis of primary

databases, such as protein families, regulatory motifs, mutation, polymorphisms, etc. In addition, there exist other specialized databases that provide different information about the biomedical data, e.g. PubMed (NIH, 2009) for bibliographic information and OMIM (McKusick and Antonarakis, 1998) for genetic diseases. In this section we present the bioinformatics data types and their corresponding databases that provide access to the principal sources of information.

The DNA sequence is the classic data type of the molecular biology. In order to decipher the nucleotide sequence of a DNA string a sequencing process needs to be carried out. The sequencing process is nowadays so automatized that sequencing the complete genome of a given organism has become a routine task. The GenBank database (Benson et al., 1999) of nucleic acid sequences currently contains a total of 99.1 billion bases in 98.8 million sequences (data from February 2009, see Figure 2.9).

At the next level are protein sequences comprising strings of 20 amino acid-letters. There are over 1 million known protein sequences, with a typical bacterial protein containing approximately 300 amino acids. Macromolecular structural data represents a more complex form of information. There are currently 61.695 entries in the Protein Data Bank (PDB) (Berman et al., 2002) most of which are protein structures. A PDB record for a medium-sized protein typically contains the x,y,z, coordinates of approximately 2.000 atoms.

Recently, the impact of the genome projects drastically increased the amount of sequence data. A genome sequence for a given organism presents the complete set of genes and their precise locations in the chromosome. As with the DNA sequences, genomes consist of strings of bases, ranging from 1.6 million bases in Haemophilus influenzae to 3 billion in humans. Automatic sequencing has had an enormous impact as it has been at the starting point of the high-throughput generation of various biological data like sequence tags (ESTs) and single-nucleotide polymorphisms (SNPs) among others.

DNA microarrays systematically analyze gene expression profiles. There exist some public resources where the scientific community can store and

query microarray data (GEO, 2009; Demeter et al., 2006). One should note that, due to the variety of platforms and preprocessing methods that are available, these databases are not as standardized as, for example, sequence databases. Other genomic scale data include biochemical information on metabolic pathways, regulatory networks, protein-protein interaction data from two-hybrid experiments, and systematic knockouts of individual genes to test the viability of an organism.

Table 2.1: *Types of data used in bioinformatics.*

Data source	*Data Size*	*Bioinformatics topics*
DNA sequence	98.8 million sequences (99.1 billion bases)	Separating coding and non-coding regions Identification of introns and exons Gene product prediction Forensic analysis
Protein sequence	$>$ 1 million sequences ($\sim$ 300 amino acids each)	Sequence comparison algorithms Multiple sequence alignments algorithms Identification of conserved sequence motifs
Macromolecular structure	61000 structures ($\sim$ 1.000 atomic coordinates each)	Secondary, tertiary structure prediction 3D structural alignment algorithms Protein geometry measurements Surface and volume shape calculations Intermolecular interactions
Genomes	189 complete genomes (1.6 million - 3 billion bases each)	Characterisation of repeats Structural assignments to genes Phylogenetic analysis Genomic-scale censuses Relating specific genes to diseases
Gene expression	$\sim$ 50000 probes per chip	Correlating expression patterns Mapping expression data to sequence, structural and biochemical data
Literature	11 million citations	Text mining Digital libraries
Metabolic pathways		Pathway simulations

Table 2.2: ***Databases used in bioinformatics.***

Knowledge	*Database*	*URL*
DNA sequences	DDBJ	www.ddbj.nig.ac.jp
	EMBL	www.ebi.ac.uk/embl
	GenBank	www.ncbi.nlh.nih.gov
Protein functions	PROSITE	www.expasy.ch/prosite
	BLOCKS	www.blocks.fhcrc.org
	SMART	http://smart.embl-heidelberg.de
	TIGRFAMs	www.tigr.org/TIGRFAMs/
	PRINTS	www.bioinf.man.ac.uk/dbbrowser/PRINTS/
	Pfam	www.sanger.ac.uk/Pfam/
Protein 3D folds	SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
	CATH	www.biochem.ucl.ac.uk/bsm/cath_new/
Gene expression	ArrayExpress	www.ebi.ac.uk/arrayexpress
	GEO	www.ncbi.nlm.nih.gov/geo
	SMD	http://smd.stanford.edu
Transcription factors	TRANSFAC	http://transfac.gbf.de/TRANSFAC/
	JASPAR	http://jaspar.cgb.ki.se/
Protein interactions	BIND	www.bind.ca/
	DIP	http://dip.doe-mbi.ucla.edu/
Protein pathways	KEGG	www.genome.ad.jp/kegg/
	EcoCyc	www.ecocyc.org/
Ortholog groups	COG	www.ncbi.nlm.nih.gov/COG/
Controlled vocabulary	GO	www.geneontology.org/

Computer Science in Bioinformatics

A major role of bioinformatics is to provide insights about gene function from existing data. Unfortunately, data is usually incomplete, noisy, and covers different organisms that may not share most of their features. Therefore, it is necessary to constantly make use of the biological principles to obtain meaningful information. Based on the availability of the data and goals described above, we now present the different computer science tasks that lead to a better understanding of gene function. They can be summarized as follows:

- **Comparing Sequences.** Given the increasing number of sequences available, there has been a need to develop algorithms to deal with comparisons of large numbers of long sequences. These algorithms take into account the possibility of deletion, insertion, and replacements of symbols representing the sequences, as might occur in nature.
- **Constructing Evolutionary Trees.** These trees are often also known as phylogenetic trees. They are usually built from the comparison of the sequences belonging to different organisms, grouping the sequences according to their degree of similarity. They shed light in the problem of infer how the sequences have been transformed through evolution.
- **Detecting Sequence Patterns.** This task involves, for example, one of the first problems that the bioinformatics community tried to solve: the detection of genes in a DNA sequence. Another example is the detection of common short sequences in the promoter regions of related genes (the so-called motifs). There are several ways to perform these tasks. Many of them are based on machine learning and include probabilistic grammars, or neural networks.
- **Determining 3D Structures.** These tasks intrinsically needs a high computational effort. The determination of RNA shape from sequences requires algorithms of cubic complexity. On the other hand, the inference of structures of proteins from their amino acid sequences remains an unsolved problem.
- **Inferring Cell Regulation.** The role of a gene or protein in a metabolic or signaling pathway provides a good idea about its function. As we

discussed in previous sections, genes interact with each other, while proteins can also prevent or assist in the production of other genes or proteins. In this sense, microarray technology helps to understand how some genes are co-regulated under similar circumstances.

- **Determining Protein Function and Metabolic Pathways.** The objective here is to interpret protein functions from human annotations usually derived from specific experiments, and also to provide databases representing graphs where one can query for the existence of nodes (reactions) and paths (sequences of reactions).
- **Assembling DNA Fragments.** Fragments provided by sequencing machines are assembled using computers. There has been an increasing interest in this field with the advent of the Next-Generation Sequencing (NGS). The main problems are: i) very short fragments to assemble; ii) big data, e.g. a microbial genome will yield about 200 Mbp; iii) new technologies are continuously appearing.

2.3 Gene Ontology

The Gene Ontology (GO) Consortium (Ashburner et al., 2000) has already become a *de facto* standard for describing gene products in databases. It was created with the aim of standardizing the representation of gene and gene product attributes across species and databases. Thus, it provides a structured, controlled vocabulary for describing the roles of genes and gene products in any organism.

Structure

GO organizes the information by means of three different ontologies:

- **Cellular component.** It describes locations, at the levels of subcellular structures and macromolecular complexes. Examples of cellular components include *nuclear inner membrane* and *cytoplasmic vesicle*.

- **Biological process.** It represents recognized series of events or molecular functions. A process is a collection of molecular events with a defined beginning and end. Examples of biological process terms are *cellular physiological process* or *signal transduction*.
- **Molecular function.** It describes activities that occur at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place. Examples of molecular function terms are *catalytic activity* or *binding*.

Terms within each of these ontologies are independent of each other, i.e. a term does not belong to more than one ontology. Figure 2.3 shows the distribution of the annotated gene products in the different GO ontologies.

The ontologies of GO are structured as a graph, with terms as nodes in the graph and the relations between the terms as arcs. In addition the relations between GO terms are also categorized and defined. These comprise *is a* (is a subtype of); *part of*; and *regulates*, *negatively regulates* and *positively regulates*. The GO Consortium (Ashburner et al., 2000) uses the following conventions for the relations:

- Where it is appropriate to talk about a parent-child relationship between nodes, parent refers to the node closer to the root(s) of the graph, and child to that closer to the leaf nodes; the parent would be a broader GO term, and the child would be a more specific term.
- The arrowhead indicates the direction of the relationship.

Annotations

Thus, the terms (nodes) in the GO database form a *Directed Acyclic Graph* (DAG), in which terms are children of one or several more general terms. This implies that the closer a term is to the root, the more general it is, and the closer a term is to the leaf, the more specific it is. The terms themselves do not describe specific genes or gene products. Genes and gene products are annotated by collaborating databases in one or more terms at the most

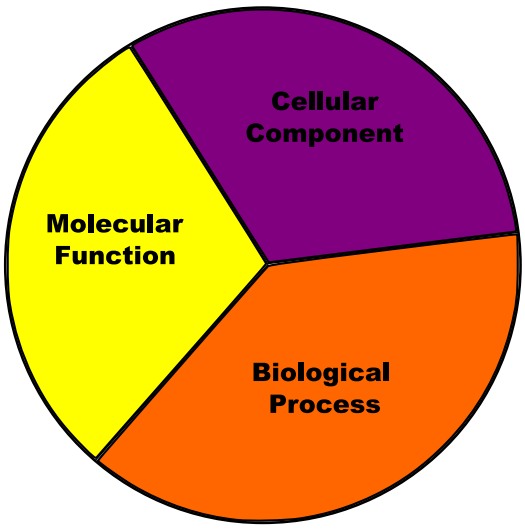

Figure 2.11: ***GO gene products distribution.*** *GO has 377451 annotated gene products as of December 2009.* Biological process *has 270918 (71.7 %).* Molecular function *has 289839 (76.7 %).* Cellular component *has 258598 (68.5 %).*

specific level possible, but are considered to share the attributes of all the parent nodes (Dwight et al., 2002). In Figure 2.3 can be seen an example of the DAG structure of GO.

Evidences

The annotations include not only the source's attribution, but also an indication of the evidence on which the annotation is based. A simple controlled vocabulary, which is provided in Table 2.3, is used to describe the evidence supporting the attribution e.g. *Traceable Author Statement* (TAS). Referencing each annotation with both experimental method and citation is intended to help researchers evaluate their reliability and is critically important to the

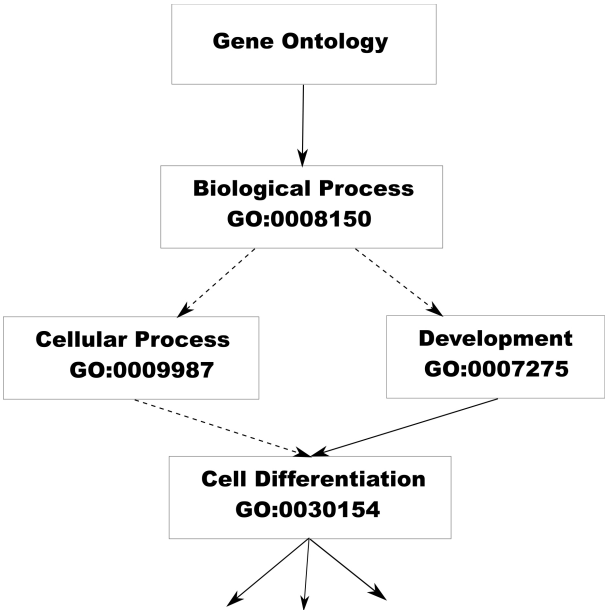

Figure 2.12: ***Directed acyclic graph taken from GO.*** *The solid arrows indicate the GO 'part-of' link and the dashed arrows the GO 'is-a' link. The GO unique identifiers (IDs) are printed below each term. The term 'Cell Differentiation' has two parents (Cellular Process and Development), which in turn link back to the same antecedent 'Biological Process' which is part-of the Gene Ontology. The unterminated arrows leading from Cell Differentiation indicate that it has a number of offspring terms.*

evaluation and use of these annotations. One may have greater confidence in an assignment based on direct experimental evidence than one based solely on a computational method such as sequence similarity. In GO there exist four types of evidence codes:

- **Experimental evidence codes**, which indicate that the cited paper displayed results from a physical characterization of a gene/gene product that has supported the association of a GO term.
- **Computational analysis evidence codes**, which indicate that the annotation is based on an *in silico* analysis of the gene sequence and/or other data as described in the cited reference.
- **Author statement codes**, which indicate that the annotation was made on the basis of a statement made by the author in the cited reference.

- **Curatorial statement evidence codes**, which indicate an annotation made on the basis of a curatorial judgment that does not fit into one of the other evidence code classifications.

Table 2.3: ***GO Evidence Codes.***

Code	Name	Type
EXP	Inferred from Experiment	Experimental
IDA	Inferred from Direct Assay	Experimental
IPI	Inferred from Physical Interaction	Experimental
IMP	Inferred from Mutant Phenotype	Experimental
IGI	Inferred from Genetic Interaction	Experimental
IEP	Inferred from Expression Pattern	Experimental
ISS	Inferred from Sequence or structural Similarity	Computational
ISO	Inferred from Sequence Orthology	Computational
ISA	Inferred from Sequence	Computational
ISM	Inferred from Sequence Model	Computational
IGC	Inferred from Genomic Context	Computational
RCA	Inferred from Reviewed Computational Analysis	Computational
TAS	Traceable Author Statement	Author Statement
NAS	Non-traceable Author Statement	Author Statement
IC	Inferred by Curator	Curatorial
ND	No biological Data available	Curatorial
IEA	Inferred from Electronic Annotation	Automatic

GO Applications

GO can be applied to a number of different tasks and many public GO-based tools have been developed. In this section a summary of the most important applications is presented:

- **Obtain the information of a gene product.** This is the first and most direct application of GO. From a given gene product it is possible to know its molecular functions, the biological processes where the gene product take place, and where in the cell is acting. This information is very useful in a diversity of biological experiment since it gives an idea of the behaviour of the gene product before it is exposed to the

experimental conditions. AmiGO (Carbon et al., 2009), provided by the GO Consortium, is the standard tool for this matter.

- **Obtain the information of a term.** Unlike the previous application, the information of interest is now the terms themselves. From them it is possible to extract their annotations and those from their offsprings, obtaining gene products that are related to each other. Here again, the AmiGO tool is usually used for this task. This information is used, for example in a study by Zhou et al. (2002) where they perform a microarray experiment analysis. In this work, the authors previously separated the genes according to their cellular component annotations in order to take into account the fact that a given metabolic pathway will have its corresponding genes activated or not regarding where in the cell the pathway is taking place at a determined moment.
- **Make a connection between biological knowledge and gene expression data.** From the combination of microarray data and GO information it is possible to obtain group of co-regulated genes biologically meaningful (see Figure 2.13). As exposed above, there are a number of tools to do this and scientifics continuously apply them in their research. For example, West et al. (2008) used the GO annotations to find overrepresentation of gene ontology groups among the significantly expressed genes.
- **Predict new associations between terms.** Some studies have focused in the prediction of new GO terms or relations between them taking into account the existing terms and/or associations. For this matter, King et al. (2003) used decision trees and Bayesian networks, and Lægreid et al. (2003) applied supervised learning techniques to face this problem.
- **Study semantics similarities between gene products.** Given two or more gene products it is possible to measure their semantic similarity studying their GO annotations. This similarity can be used to perform analysis of group of genes, to compare and validate experimental results, to group gene products for further studies, etc. For this task, sev-

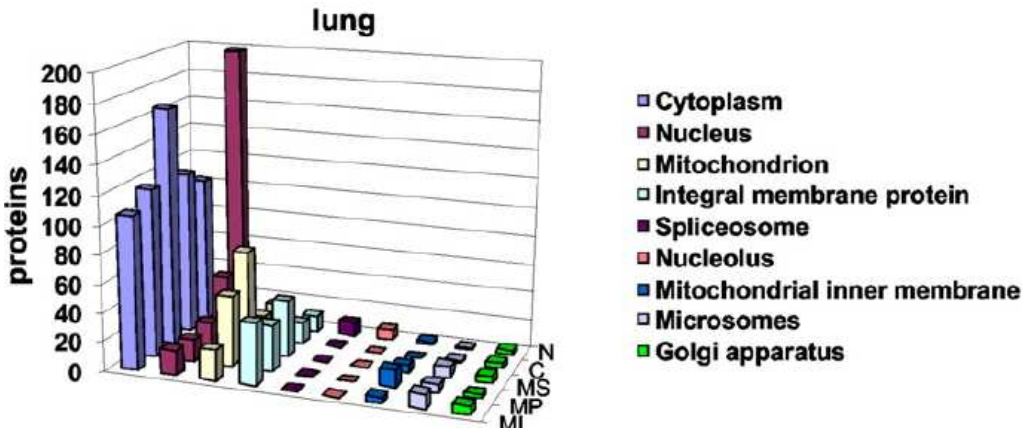

Figure 2.13: ***Cellular component protein classification.*** *Schematic representation of the subcellular fractionation procedure (Kislinger et al., 2003)*

eral measures have been previously defined (Resnik, 1995; Jiang and Conrath, 1997; Lin, 1998; Popescu et al., 2006). As it was stated above, this chapter focuses precisely on this topic. Thus, in the following sections we describe the problem and present our proposed solutions.

CHAPTER 3

Fuzzy theory

This chapter introduces some basic notions on fuzzy set theory and fuzzy logic, as well as on other soft computing methods which will be necessary to understand the rest of this document. Further details about these topics can be found in Dubois et al. (2007), Hájek (2005), and Klir and Yuan (1995).

Section 3.1 begins with the definition of a fuzzy set. Section 3.2 describes some basic operations with fuzzy sets. Fuzzy clustering methods are introduced in Section 3.3. Then, Section 3.4 and Section 3.5 present the concept of fuzzy measure and the define the λ-fuzzy measures respectively. Section 3.6 outlines the fuzzy integrals. Section 3.7 gives a brief introduction on other soft computing methods used for the purpose of this work (kernel methods and genetic algorithms). Finally, a summary of some of the most important applications of fuzzy theory on bioinformatics is given in Section 3.8.

3.1 Fuzzy sets

The classical notion of set is deeply related to the fulfillment of a given property which is satisfied by all the members of the set. We may think of a property as a function defined over a set of objects U (which is referred as the referential set or domain of discourse) relating each of these objects to a

element of the set $\{0,1\}$. A particular element belongs to the set if the function assigns 1 to it; otherwise (if the function assigns 0 to it), the element does not belong to the set. These sets are called crisp or classical.

According to this, any property P determines a set S_P which is composed by the following elements:

$$S_P = \{u \in U : P(u) = 1\}.$$

In the same way, any subset $S \subseteq U$ induces a property P_S which is determined by the following expression:

$$P_S(u) = 1 \text{ if and only if } u \in S.$$

Fuzzy set theory, originally proposed by (Zadeh, 1965) generalizes this classical notion of set, having into account that the properties which define a set are defined over the referential U, but now using as an image the real interval $[0,1]$. Any property satisfying these characteristic is said to be a fuzzy property, and the set that it determines is given by the following expression:

$$S_P = \{< u, \alpha >: P(x) = \alpha, u \in U, \alpha \in [0,1]\}.$$

Definition 1 Fuzzy set. *Let U be a referential set. A fuzzy subset A of U is every set of the form $A = \{(u,\alpha), u \in U, \alpha \in [0,1])\}$, that is, every set formed by the objects from U, having associated each of them some membership degree, defined in the interval $[0,1]$, to A.*

Consequently, a fuzzy set A defined over the domain of discourse U is univocally characterized using a *membership function* $\mu_A(u)$, or simply $A(u)$, which assigns any $u \in U$ to a value in the interval of real numbers between 0 and 1, representing the membership degree of the element u to A. As in the classical case, 0 means no-membership and 1 full membership, but now a value between 0 and 1 represents the extent to which u can be considered

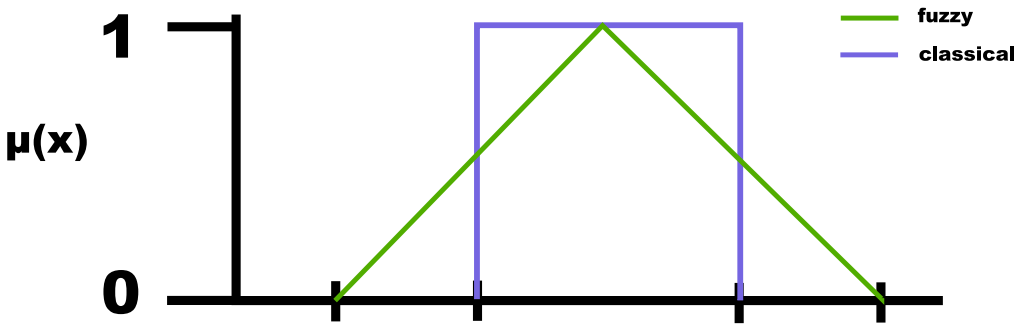

Figure 3.1: ***Fuzzy membership function.***

as an element of A. For example, in Figure 3.1 the elements in $[a,b]$ fully belong to A, whereas the elements in (b,c) partially belong to A.

If the domain of discourse U is discrete ($U = \{u_1, u_2, ..., u_n\}$), the fuzzy set is usually expressed using the following notation:

$$A = \mu_A(u_1)/u_1 + \mu_A(u_2)/u_2 + \ldots + \mu_A(u_n)/u_n.$$

When U is continuous, the fuzzy set is denoted by:

$$A = \int_{u \in U} \mu_A(u)/u.$$

The set of all fuzzy subsets which can be defined over a domain of discourse U is called $\tilde{\wp}(U)$. Classical sets are a special case of fuzzy sets and hence $\wp(U) \subseteq \tilde{\wp}(U)$.

Level Cuts

Definition 2 ***α-cut.*** *For each $\alpha \in [0,1]$ and each fuzzy set A, the α-cut of A is defined as the set of all elements of the domain of universe which have a membership degree to A which is greater or equal than α, that is:*

$$A_{\geq\alpha} = u \in U : \mu_A(U) \geq \alpha.$$

The different α-cuts of a fuzzy set have an inclusion relation between them which is determined by the following property:

$$(\alpha > \beta) \Rightarrow (A_{\geq\alpha} \subseteq A_{\geq\beta}).$$

Definition 3 ***Strict*** α***-cut.*** *Analogously, for each* $\alpha \in [0,1]$ *and each fuzzy set A, the strict* α*-cut of A is defined as the set of all elements of the domain of universe which have a membership degree to A which is strictly greater than* α*, that is:*

$$A_{>\alpha} = u \in U : \mu_A(U) > \alpha.$$

Obviously, strict α-cuts are contained in α-cuts:

$$A_{>\alpha} \subseteq A_{\geq\alpha}.$$

Among the crisp sets which can be defined from a fuzzy set, there are two of special significance: the support and the core.

Definition 4 ***Support.*** *The support of a fuzzy set A defined over a domain of discourse U is the set of elements of U which have a membership degree strictly greater than 0, that is:*

$$supp(A) = \{u \in U : \mu_A(u) > 0\}.$$

Definition 5 ***Core.*** *The core of a fuzzy set A defined over a domain of discourse U is the set of elements of U which have a membership degree equal to 1, that is:*

$$core(A) = \{u \in U : \mu_A(u) = 1\}.$$

Finally, Zadeh's Resolution's Identity (Zadeh, 1965) shows that a fuzzy set A can be univocally represented from its decomposition in α-cuts in the following way:

$$\wedge(A) = \{\alpha : \mu_A(u) = \alpha \text{ for some } u \in U\}.$$

3.2 Fuzzy set operators

Zadeh (1965) provided the basic definitions for the generate the *fuzzy set theory*, i.e. *union* of two fuzzy sets, *intersection* of two fuzzy sets and *complement* of a fuzzy set. Let $A : X \rightarrow [0,1]$ be a fuzzy subset of X. The complement, A^c, of A is defined as:

$$A^c(x) = 1 - A(x). \tag{3.1}$$

In addition, if $B : X \rightarrow [0,1]$ is another fuzzy subset of X, Zadeh defined:

$$(A \cup B) = \max\{A(x), B(x)\} = A(x) \vee B(x),$$

and

$$(A \cap B) = \min\{A(x), B(x)\} = A(x) \wedge B(x).$$

These standard definitions are just one of the infinite number of ways to define complement, union and intersection (Klir and Yuan, 1995). For example, Yager (1980) proposed a family of operators very useful for multicriteria decision making with the complement, union and intersection are given by:

$$\begin{aligned} A^c(x) &= (1 - A(x)^w)^{1/x}, && (3.2)\\ (A \cap_w B)(x) &= \min\{1, A(x)^w + B(x)^w)^{1/w}\}, \text{ and} && (3.3)\\ (A \cup_w B)(x) &= 1 - \min\{1, ((1 - A(x))^w + (1 - B(x))^w)^{1/w}\}, && (3.4) \end{aligned}$$

where $w \in (0, \infty)$. Another fuzzy set theory used for fuzzy logic inference is generated by the operators:

$$\begin{aligned} (A \cup_b B)(x) &= 1 \wedge (A(x) + B(x)), && (3.5)\\ (A \cap_b B)(x) &= 0 \vee (1 - (A(x) + B(x))), && (3.6) \end{aligned}$$

together with the standard complement defined in equation 3.1.

3.3 Fuzzy Clustering

One of the main tools to mine and analyze unlabeled data is clustering, which is a division of data into groups of similar objects. Thus, each cluster is a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. From the machine learning perspective, clustering can be seen as an unsupervised learning of concepts. In order to describe some fuzzy approaches for this task, we first introduce the problem of clustering.

Let X be a data set consisting of data points x_k $(1 \leq k \leq n)$. All clustering is based on the concept of a C-partition of the set X defined by the a partition matrix $U = \{u_{ik}\}$ $(1 \leq k \leq C$ and $1 \leq i \leq n)$, where u_{ik} is the membership degree of x_k to the cluster A_i, fulfilling the following property:

$$\sum_{i=1}^{C} u_{ik} = 1 \text{ for all } k. \tag{3.7}$$

In the crisp case, each x_k is assigned to a single cluster A_i. In other words, for each $x_k \in X$, $u_{ik} = 1$ for some cluster i between 1 and C, and $u_{ik} = 0$ for all other clusters. These conditions are relaxed in the fuzzy case as we will show next.

C-Means

C-Means clustering (Hartigan, 1975) is a maximization of expectation algorithm that minimize the following cost function:

$$c - means_{cost} = \sum_{i=1}^{c} \frac{\sum_{j=1}^{n} \sum_{k=1}^{n} M_{ji} M_{ki} D_{jk}}{\sum_{l=1}^{n} M_{li}}, \tag{3.8}$$

where c is the number of clusters, n is the number of objects to cluster, D is the pairwise distance matrix, and M is a binary stochastic matrix $M \in \{0,1\}^{n \times k}$ where $M_{ji} = 1$ if object j is in cluster i.

Fuzzy C-Means

The Fuzzy C-Means (FCM) algorithm (Bezdek, 1981) aims to partition a set of data into a given number of clusters considering the uncertainty of cluster assignment. Likewise, it allows for sharing objects between clusters. This method represents each cluster by a prototype (or cluster center). Let v_i be the prototype of cluster A_i and let V be the set of all C cluster prototypes. The objective of FCM is to minimize:

$$J(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{C}(u_{ik})^m d^2(x_k, v_i), \tag{3.9}$$

where d^2 is a distance function and the condition $\sum_{i=i}^{C} u_{ij} = 1$ for all k is satisfied. In this equation, the parameter m is called the fuzzifier. Larger values of m benefits more *fuzzy* partitions. In order to perform this minimization the following two equations need to be solved.

Prototypes must have the following form:

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m}. \tag{3.10}$$

In addition, the necessary condition on the membership values is:

$$u_{ik} = \frac{\left(\frac{1}{d(x_k, v_i)}\right)^{\frac{2}{m-1}}}{\sum_{j=1}^{C}\left(\frac{1}{d(x_k, v_i)}\right)^{\frac{2}{m-1}}}. \tag{3.11}$$

The FCM algorithm iteratively updates cluster memberships and cluster prototypes in each iteration. The pseudocode for FCM can be found in Figure 3.3. The inputs for FCM are the data set X, the number of clusters C, the fuzzifier parameter m, and the stop criterion ε.

FUZZY C-MEANS(X, C, m, ε)

$V^{(0)} \leftarrow \{v_1^{(0)}, \ldots, v_C^{(0)}\}$ ▷ randomly
$t \leftarrow 0$
repeat
  for $k \leftarrow 1$ **to** $length[X]$
    do
      if $d(x_k, v_i) = 0$ for some i
        then $u_{ik}^{(t)} \leftarrow 1$ and $u_{jk}^{(t)} = 0$ for $j \neq i$
        else Compute $u_{ik}^{(t)}$ applying Equation 3.11
  $t \leftarrow t+1$
  Compute $V^{(t)}$ applying Equation 3.10 using $U^{(t-1)}$
until $\sum_{i=1}^{C} \|v_i^{(t)} - v_i^{(t-1)}\| < \varepsilon$ ▷ $\|*\|$ is any vector norm

Figure 3.2: ***Fuzzy C-Means pseudocode.***

Possibilistic C-Means

The FCM algorithm sheds light into the problem of crisp grouping when the features possess ambiguity. FCM has offered promising results in different fields of application, it suffers from some problems specially when the object present high similarities with elements of two different clusters, i.e can be a member of two different classes at the same time. In reality these is no reason that memberships of a given feature sum to one (Equation 3.7). FCM is subject to this constraint in order to avoid the trivial solution (all memberships equal zero) in minimizing the criterion function (Equation 3.9). Also, it is not uncommon that some of the features extracted are outlier that really do not belong to any cluster.

Krishnapuram and Keller (1993) proposed a new clustering method to overcome the drawbacks of FCM by relaxing the sum constraint while avoiding the trivial solution. They defined a new criterion function resulting a the new algorithm called Possibilistic C-Means (PCM). The criterion function is:

$$J(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{C}(u_{ik})^m d^2(x_k, v_i) + \sum_{i=1}^{C}\eta_i \sum_{k=1}^{n}(1-u_{ik})^m, \tag{3.12}$$

where the η_i are appropriately chosen or estimated values, and can be interpreted as the radio of the corresponding cluster (Krishnapuram and Keller, 1996). Here, the necessary conditions to minimize Equation 3.12 are:

$$u_{ik} = \frac{1}{1 + \left(\frac{d(v_i, x_k)^2}{\eta_i} \right)^m},$$

and the condition on the cluster prototypes is identical to Equation 3.10.

3.4 Fuzzy Measures

Let $X = \{x_1, x_2 ..., x_n\}$ be a finite set, let $A, B \subseteq X$, and let $\wp(X)$ the power set of X. A fuzzy measure, μ, is a real valued function $\mu : \wp(X) \rightarrow [0, 1]$, satisfying the following properties:

1. $\mu(\emptyset) = 0$ and $\mu(X) = 1$. (3.13)
2. $\mu(A) \leq \mu(B)$ if $A \subseteq B$. (3.14)

The reader should note that the additivity condition of probability theory is relaxed in property 2 to the condition of monotonicity.

For a fuzzy measure μ, let $\mu(\{x_i\}) = \mu^i$. The mapping $x_i \rightarrow \mu^i$ is known as *fuzzy density function*. The fuzzy density of a single element $x_i \in X$, μ^i, can be interpreted as the importance of x_i in determining the set X.

Fuzzy measures can be separated into different classes according to the strategies used to evaluate the similarity. Next, we provide a summary of those that have been used for the purpose of this thesis.

Set-theoretic measures: Jaccard coefficient

Set-theoretic measures can be considered generalizations of the classical set-theoretic similarity functions. The set-theoretic operations on fuzzy sets are used to define various measures. Among them, we selected the well-known Jaccard coefficient for some problems addressed in this work. The Jaccard

coefficient is an unparameterized ratio model of similarity (Jaccard, 1908). It is also known as index of communality. The Jaccard coefficient of two fuzzy sets A and B over a finite universe of discourse $U = \{u_1, u_2, \cdots, u_n\}$ is:

$$S_J(A,B) = \sum_{i=i}^{n} \frac{\mid \mu_A(u_i) - \mu_B(u_i) \mid}{\max\left(\mu_A(u_i), \mu_B(u_i)\right)}. \tag{3.15}$$

Angular coefficient-based: Bhattacharyya distance

Bhattacharyya distance measures the cosine of the angle between two vectors when the values in each vector are standardized as deviates from the mean of the membership function (Bhattacharyya, 1946). This cosine is taken as the corresponding similarity measure

$$S_B(A,B) = \frac{\sum_{i=1}^{n}\left(\mu_A(u_i)\cdot\mu_B(u_i)\right)}{\left(\sum_{i=1}^{n}\left(\mu_A(u_i)\right)^2\right)^{\frac{1}{2}}\cdot\left(\sum_{i=1}^{n}\left(\mu_B(u_i)\right)^2\right)^{\frac{1}{2}}}, \tag{3.16}$$

where notation from previous section holds.

Proximity-based measure: Minkowsky r-metric

The distance between the partial membership functions of fuzzy sets A and B over a finite universe of discourse $U = \{u_1, u_2, ..., u_n\}$ may be measured using a Minkowsky r-metric (Zwick et al., 1987). A fuzzy set A is represented by a point $[\mu_A(u_1), ..., \mu_A(u_n)]$ in the n-dimensional space:

$$d_r(A,B) = \left(\sum_{i=1}^{n} \mid \mu_A(u_i) - \mu_B(u_i) \mid^r\right)^{\frac{1}{r}}, r \geq 1. \tag{3.17}$$

As the name implies, the Minkowsky r-metric is a metric. With $r = 1$ d_r becomes the *city-block* model or Hamming distance; $r = 2$, the Euclidean distance; and with $r = \infty$, the dominance metric. More on this topic can be found in Cross and Sudkamp (2002).

3.5 λ−Fuzzy Measures

Due to the nature of the definition of a fuzzy measure μ, the measure of the union of two disjoints subsets cannot be directly computed from the component measures. In other words, the fuzzy measure value of a subset is not just the sum of the measures of its elements. Therefore, in order to define a fuzzy measure one needs to know not only the individual fuzzy densities of the elements of the measured set, but also the measure for each combination thereof. This information can be supplied by an expert or extracted from the problem definition. However, when dealing with sets of numerous elements this task might become noisy, tedious or even unfeasible. A possible solution for this problem is the use of λ-fuzzy measures

λ-fuzzy measures (Sugeno, 1977) satisfy the properties of fuzzy measures plus the following additional property: for all $A, B \subset X$ and $A \cap B = \emptyset$,

$$\mu(A \cup B) = \mu(A) + \mu(B) + \lambda\mu(A)\mu(B), \text{ for some } \lambda > -1. \tag{3.18}$$

Furthermore it can be proved that λ can be obtained by solving:

$$\lambda + 1 = \prod_{i=1}^{n}(1 + \lambda\mu^i). \tag{3.19}$$

Therefore, applying Equation 3.18 and 3.19 one will only need to know the individual fuzzy densities of the elements, μ^i, $(i = 1, \ldots, n)$, in order to construct the fuzzy measure.

3.6 Fuzzy Integral

Sugeno Fuzzy Integral

Let X be a set and let $h : X \to [0,1]$ represent a function that matches each element of X to its evidence, and let $\mu : \wp(X) \to [0,1]$ be a fuzzy measure. Then the Sugeno fuzzy integral is defined by:

$$\int h(x) \circ \mu = \sup_{E \subseteq X} [\min(\min_{x \in E}(h(x), \mu(E)))] = \sup_{\alpha \in [0,1]} [\min(\alpha, \mu(A_\alpha))],$$

where $A_\alpha = \{x | h(x) \geq \alpha\}$.

For the finite case, let $X = \{x_1, \ldots, x_n\}$ be a finite set representing a set of n information sources. Let's suppose that $h(x_1) \geq h(x_2) \geq \cdots \geq h(x_n)$, if it is not the case for any element, then reorder X so that the relation holds. Then the Sugeno fuzzy integral of h with respect to the fuzzy measure μ is:

$$S_\mu(h) = \max_{i=1}^{n} [\min(h(x_i), \mu(A_i))], \tag{3.20}$$

where $A_i = \{x_1, \ldots, x_i\}$. The reader should note that if μ is a λ-fuzzy measure, then $\mu(A_i)$ can be obtained applying equation (3.18).

The fuzzy integral considers the evidence supplied by each element of a given set and the worth of each subset of elements (by means of a fuzzy measure) in its decision making process. This combination of the importance of the sources and the information provided makes the fuzzy integral appropriate for information fusion. Due to its ability to deal with uncertainties associated with the data extracting and processing procedures, it has been widely applied in pattern recognition and classification (Keller et al., 2000; Sugeno, 1977).

Choquet Fuzzy Integral

Murofushi and Sugeno proposed the Choquet fuzzy integral, referring to a function defined by Choquet in a different context (Murofushi and Sugeno, 1989). The Choquet fuzzy integral is a fuzzy integral based on λ-fuzzy measure that provides alternative computational scheme for aggregating information. Let $h(x_1), \ldots, h(x_n)$ be a collection of input sources of h and let μ be a λ-fuzzy measure, then the Choquet fuzzy integral can be defined as

$$\int_X h(x) \circ \mu = \int_0^1 \mu(A_\alpha) d\alpha, \text{ where } A_\alpha = \{x | h(x) \geq \alpha\}. \tag{3.21}$$

For the finite case, the Choquet fuzzy integral is defined by

$$C_\mu(h) = \sum_{i=1}^{n} [h(x_i) - h(x_{i+1})] \cdot \mu(A_i), \tag{3.22}$$

where $h(x_1) \geq h(x_2) \geq \cdots \geq h(x_n)$, $h(x_{n+1}) = 0$, and $A_i = \{x_1, \cdots, x_i\}$.

3.7 Other Soft Computing Methods

Kernel Methods

Given a space X of objects we want to classify, cluster, rank, etc., we can define a function $\phi : X \to F$, where F is a feature space that eases X classification, clustering, ranking, etc. For example, objects could be more separable in F than in X. Imagine we have a real-valued function $k : X \times X \to \Re$ and for each $x, y \in X$, $k(x, y)$ tells us how similar x and y are in F. k is called a kernel function and can be defined as the inner product in F: $k(x, y) = \phi(x) \cdot \phi(y)$. In fact, most of the times F is hard or impossible to compute, *e.g.* it could be infinite dimensional. A learning method that uses k to avoid F computation is called a kernel method. More on this topic can be found in Schölkopf et al. (2004).

Let us call $P = \{x_1, x_2, ..., x_n\}$ the set of objects to be analyzed. We can construct a kernel matrix $K_{i,j} = k(x_i, x_j)$, $x_i, x_j \in X$. K can be thought as a similarity matrix in F and it is the only way kernel methods access data. For K to be a kernel, it must be semidefinite positive, i.e. all its eigenvalues must be non-negative.

Any learning algorithm that can be formulated in terms of inner products can be interpreted as a kernel method if we replace the inner product with a kernel function. This is known as the kernel trick Schölkopf et al. (2004) and allows us to convey kernel ideas to clustering, as can be seen in the original paper.

Genetic Algorithms

Genetic algorithms, developed in the mid-1960s, are inspired by Darwin's theory about evolution. In this section we provide an introduction to the schema of the genetic algorithms. In the work of (Holland, 1992) more details can be found.

Possible solutions to a problem are called chromosomes, and a diverse set of chromosomes is grouped into a gene pool. The relative quality of these answers is determined using a fitness function. This quality is used to determine whether or not the chromosomes will be used in producing the next generation of chromosomes. The contents of high quality chromosomes are more likely to continue into the next generation. The next generation is generally formed via the processes of crossover, i.e. combining elements of two chromosomes from the gene pool, and mutation, i.e. randomly altering elements of a chromosome. A large number of strategies have been proposed for determining the contents of a new generation of chromosomes. Two classical approaches are discussed here.

The first approach was described by Holland (1992). Each solution in the gene pool is evaluated by the fitness function. A probability of being selected as a member of the next pool is assigned to each chromosome, according to the quality of the actual solution. Therefore, those with better quality are more likely to be chosen. A new pool is then constructed by randomly selecting solutions following the probability distribution generated. The new generation is then created by mixing the chromosomes of the new pool chosen at random. This is generally called crossover.

A second approach that may be used was proposed by Bean (1994). This strategy as elitist, since some percentage of the solutions with the best fitness function values are copied directly into the the next generation. In addition some random chromosomes are used to perform crossovers. The next generation is then formed by some of the best solutions and some of those generated from such crossover. The choice here is done entirely at random; no weighting based on the quality of the solutions is performed. Bean (1994) shows that this strategy leads to stable results.

Genetic algorithms are not sensitive to the presence of local minima since they work on a large number of points in the problem space simultaneously. Discussion of the benefits gained by using genetic algorithms instead of exhaustive search for different optimization problems can be found in Painton and Campbell (1995).

3.8 Fuzzy Applications in Bioinformatics

Fuzzy theory has been successfully applied to diverse practical areas. In bioinformatics, it has been used to develop systems and methods for a variety of problems. Some of the most important applications are:

- Gene functions prediction (Tari et al., 2009).
- Study of differences between polynucleotides (Torres and Nieto, 2003).
- Alignment of DNA sequences based on the characteristics of DNA fragments and a fuzzy logic system (Kim et al., 2008a).
- DNA sequencing using genetic fuzzy systems (Cordon et al., 2004).
- Clustering genes from microarray expression data (Mukhopadhyay and Maulik, 2009)
- Spot segmentation and quantification of gene expression level from microarray images (Wang et al., 2008).
- Measuring the significance of gene pathways in a particular disease (Liang et al., 2008).
- Deciphering genetic networks (Ressom et al., 2003).
- Classification of amino acid sequences (Bandyopadhyay, 2005).
- Biological knowledge extraction (Lopez et al., 2008).

Part III

Thesis Contributions

CHAPTER 4

Semantic Measures for Gene Ontology

Due to the recent flood of new biological data from genome sequencing, scientists are struggling to answer many basic questions and attempting to extract information from this data. One of the main tasks is to discover to which function the genes are associated. Ontology has long been the preserve of philosophers and logicians. Recently, ideas from this field have been applied by computer scientists as a basis for encoding knowledge and with the hope of achieving interoperability and intelligent system behavior. In bioinformatics, ontologies ease query and data-mining activities. Recently bio-ontologies have played an important role for the automatic integration of background knowledge which is fundamental to support the generation and validation of hypotheses about the function of gene products.

In this chapter we introduce the Gene Ontology (GO) as a tool to extract biological information from groups of potentially related gene products. Likewise, we investigate the performance of different ontology semantic measures and clustering methods in protein family recognition tasks. Finally, we propose a novel fuzzy semantic similarity measure over GO and we show its effectiveness in protein classification problems, comparing our proposed measure with existing approaches.

This chapter is organized as follows. Section 4.1 reviews the principles of semantic similarity in ontologies. Section 4.2 discusses the crisp approaches to the problem and presents a novel methodology for evaluating these measures together with different cluster methods for protein family recognition. The novel fuzzy similarity measure is presented in Section 4.3. In Section 4.4 we evaluate the performance of GO semantic similarity measures based on the effectiveness in recognizing protein families. Section 4.5 concludes this chapter.

4.1 Semantic Similarity in Ontologies

Before we discuss the GO semantic similarity, we will outline first how the semantic similarity for two elements in a general ontology can be defined. The definition of semantic similarity measures for elements of an ontology has been intensively studied in diverse fields such as artificial intelligence, psychology, linguistic, etc. In particular, the data mining and information retrieval literature presents a number of research works focused on documents similarity (Lee et al., 2007; Janowicz et al., 2008).

Edge-based Methods

The traditional method for measuring the semantic distance between two elements of an ontology is to measure the distance between the two nodes where the elements are annotated (Lee et al., 1993). In other words, given two elements c_1 and c_2, their semantic distance is computed as a function of the distance between the nodes where they are contained. There exist a number of proposed approaches that makes use of this idea. The most widely extended are:

$$sim(c_1, c2) = \min(dist(C_1, C_2)) \tag{4.1}$$

$$sim(c_1, c2) = \frac{\sum dist(C_1, C_2))}{k} \tag{4.2}$$

$$sim(c_1, c2) = \min(\text{dist}(C_1, \text{NCA}(C_1, C_2)), \text{dist}(C_2, \text{NCA}(C_1, C_2))), \tag{4.3}$$

where C_1 and C_2 are the nodes where the elements c_1 and c_2 are contained, $dist(c_1, c_2)$ is the edge distance between the elements c_1 and c_2, min is the usual minimum function, k is the number of paths between c_1 and c_2, and $NCA(c_1, c_2)$ is the nearest common ancestor of c_1 and c_2.

In equation 4.1 the semantic distance of c_1 and c_2 is defined as the length of the minimum path between C_1 and C_2.

In equation 4.2 the semantic distance of c_1 and c_2 is defined as the average of the length of the paths between C_1 and C_2.

In equation 4.3 the semantic distance of c_1 and c_2 is defined as the minimum distance between each node and the nearest common antecesor of them.

One disadvantage of these techniques is that the distance between nodes of the higher levels of the ontology (non-specific nodes) might be similar to those obtained between nodes in low levels (very specific). This fact lead to spurious similarities as it was shown in a work from Richardson and Smeaton (1995).

A main drawback derived from the use of these methods is that they need a uniform distribution of the nodes and their relations across the ontology in order to provide proper results. These conditions are not always fulfilled, specially when the size of ontology increases, as is usually the case when dealing with ontologies defined for real problems like GO.

Node-based Methods

There exist a variety of techniques that make use of the properties of the nodes to calculate the semantic similarity between two elements of an ontology. Some of these techniques were studied by Bernstein et al. (2005). Among all the approaches, the Jaccard measure (Manning and Schutze, 2002), based on the set theory, was one of the most popular for the analysis of the similarity of ontology nodes:

$$s(c_1, c_2) = \frac{|c_1 \cap c_2|}{|c_1 \cup c_2|}, \tag{4.4}$$

where c_1 and c_2 are the ontology nodes being compared.

More recently, node-based methods that use the information content of the nodes of the ontology in order to provide a semantic similarity/distance for two elements have emerged as a good alternative. The main idea is to measure the semantic similarity/distance of two concepts as a function of the degree of information that the two concepts share in common. It is assumed that the more information the nodes share, the more similar they are. This is precisely the approach that fits the best with the features of GO. Therefore, in this section we focus on those models that make use of the Information Content (IC) of the nodes which is defined as:

$$IC(c) = -\log(P(c)), \tag{4.5}$$

where c is a node and $P(c)$ is the probability of occurrence of c or some of its offspring in the ontology.

Resnik Model

Resnik (1995) first proposed to define an object of an ontology is defined by the members of the class specified.. In a explicit ontology (e.g. GO) the set of members of a class is equivalent to the offspring of the object in question. The information of a given class is defined as the probability P of finding an occurrence of this class or any of its offsprings in the ontology. The entropy of a class a function of the logarithm of such probability. Thus, the **similarity** between two nodes c_1 and c_2 is:

$$SIM_{Resnik}(c_1, c_2) = \max_{c \in S(c_1, c_2)} (-\log(P(c))) = \max_{c \in S(c_1, c_2)} (IC(c)), \tag{4.6}$$

where $S(c_1, c_2)$ represents the set of common antecesors shared by the nodes c_1 and c_2. Thus, Resnik proposed that the similarity between two nodes of an ontology is the specificity (or entropy) of the most specific common antecesor. This measure can take values in the interval $[0, \infty]$ since P takes values between 0 and 1.

Lin Model

An alternative to the Resnik model was proposed by Lin (1998). Similarly, Lin used the IC of the nodes of the ontology for his purpose. In this model the similarity of two nodes is also computed taking into account their common antecesors. However, Lin incorporates the IC of the nodes that are being compared. Thus, Lin defined the **similarity** between two nodes c_1 and c_2 as:

$$SIM_{Lin}(c_1, c_2) = \frac{2 \cdot \max_{c \in S(c_1,c_2)}(-\log(P(c)))}{\log(P(c_1)) + \log(P(c_2))}, \tag{4.7}$$

where the notation of equation 4.6 holds. This measure can take values between 0 and 1. It can be seen as a normalized version of equation 4.6.

Mixed methods

In addition of the two traditional methods (nodes-based and edges-based), there exist other *mixed* approaches that try to integrate the benefits of both of them into a single method. The method proposed by Jiang and Conrath (1997) is the mixed method that has had a great impact in the scientific community (Islam and Inkpen, 2008; Markines et al., 2009).

Jiang and Conrath Model

Jiang and Conrath (1997) proposed a method that is derived from the edges-based models (taking into account node distances), and incorporated the nodes IC as a factor. Particularly, it focused in determining the strength link strength (LS) between a parent node with its child node. The authors proposed that the such strength is a function of the probability of occurrence of the child node c_i given the occurrence of his parent node p:

$$P(c_i|p) = \frac{P(c_i \cup p)}{P(p)} = \frac{P(c_i)}{P(p)}. \tag{4.8}$$

According to theoretical information concepts, Jiang and Conrath (1997) defined the LS as the negative logarithm of the probability shown in equation 4.8:

$$LS(c_i, p) = -\log(P(c_i|p)) = IC(c_i) - IC(p). \tag{4.9}$$

This indicates that the LS can be calculated as the difference of IC of the child and parent nodes. Following these ideas, the authors defined their **distance** semantic measure as:

$$DIST_{J\&C}(c_i, c_j) = IC(c_i) + IC(c_j) - 2 \cdot IC(NCA(c_i, c_j)), \tag{4.10}$$

where the notation of previous equation holds. This measure provides the semantic difference between the pair of nodes (c_i, c_j) and gives values between $[0, \infty]$.

4.2 Gene Ontology Crisp Semantic Measures for Related Proteins Recognition

In this section we show how general ontology semantic measures can be adapted to be applied in GO and we make use of these measures for: *i)* study of correlations between GO and some protein grouping approaches (sequence similarity, expression level, protein-protein interaction, and protein homology), and *ii)* evaluate the performance of several clustering methods in recognition of related proteins (standard c-means, "matrix" c-means, kernel c-means and constant shift embedding c-means).

Our methodology includes the construction of families of related proteins and the comparison of how well the different clustering methods recognize our initial families. We use two GO ontologies (*biological process* and *molecular function*) independently and all three together (*ALL* = *biological process* + *molecular function* + *cellular component*) to compute the three *GO* semantic measures explained above (Resnik, Lin and Jiang & Conrath).

Adapting Classical Measures

As outlined in the previous section, classical semantic similarity measures for ontologies are designed for comparing two individual nodes. Thus, these measures are not directly suitable for measuring semantic similarities of elements defined by a set of nodes. This approach can be used when dealing with domains like WordNet (Fellbaum et al., 1998) where, although a single word can have multiple meanings, it is rare that such ambiguity arises since the appropriate meaning can be inferred from the context where the word is used. However, this is not the case of GO gene products. For example, when a given gene is significantly expressed in a microarray experiment, it is not possible to know in advance which of its different functions or activities will the corresponding gene product perform. To this effect, the most widely used approach is to compute the semantic similarity/distance measure of two gene products as the average inter-set similarity between the GO annotations of the two gene products:

$$SIM(g_i, g_j) = \frac{1}{m \cdot n} \sum_{a_k \in A_i, a_p \in A_j} sim(a_k, a_p) \tag{4.11}$$

$$DIST(g_i, g_j) = \frac{1}{m \cdot n} \sum_{a_k \in A_i, a_p \in A_j} dist(a_k, a_p), \tag{4.12}$$

where g_i and g_j are gene products with the sets of GO annotations A_i and A_j respectively, A_i and A_j have m and n annotations respectively, $sim(a_k, a_p)$ is the semantic similarity of the terms a_k and a_p computed as in equations 4.6 and 4.7 and $dist(a_k, a_p)$ is the semantic distance of the terms a_k and a_p computed as in equation 4.10.

Distances and Similarities

Among the three GO semantic measures (Resnik, Lin and Jiang & Conrath), Resnik and Lin propose similarity measures, while Jiang & Conrath define a distance or dissimilarity measure. We use clustering methods whose inputs are either distances or similarities, so next it is shown how to convert ones into others:

Generally, if we initially have a distance D, we normalize to $[0,1]$ and later we compute the similarity S as $S = 1 - D$. The same is true to convert a similarity into a distance.

However, in case the similarity is an inner product, we can compute the distance between objects i and j as $D_{ij} = S_{ii} + S_{jj} - 2S_{ij}$. We can not get a similarity from a metric distance likewise, since several similarity matrices S can produce the same distance D.

Clustering

Some difficulties arise when we try to use measures defined over GO in learning methods, such as clustering. First of all, Jiang is a dissimilarity measure, but it is not a metric. In this case, the space defined by the measure is not a metric space, i.e. it does not fulfill the triangle inequality $D_{ij} \leq D_{ik} + D_{kj}$ where D_{ij} denotes the distance between points i and j. The same holds true for similarity measures, such as Resnik and Lin. Those measures do not have the properties of an inner product and, therefore, they do not define a norm that fulfills the triangle inequality. Consequently, the learning methods, which are prepared to work with metric distances or inner products, do not work properly with those measures. For example, they do not converge. We take two approaches to solve this issue:

1. If we want a distance D to be metric, we use Constant Shift Embedding (CSE) (Roth et al., 2003a): Given D, CSE sums the minimum possible constant D_0 to the non-diagonal entries in D in order to make D a metric. It can be shown that $D_0 = -2\lambda_n(S^c)$, where $\lambda_n(x)$ is the minimum eigenvalue of x and S^c is the centralized similarity matrix computed as $S^c = -\frac{1}{2}D^c$ and $D^c = QDQ$ with $Q = I - \frac{1}{n}ee^T$, $e = (1,1,...,)^T$.
2. According to that explained in Section 3.7, if we want a similarity S to be an inner product, we have to force it into a kernel. A kernel is a positive semidefinite matrix, i.e. all its eigenvalues should be non-negative. Therefore, we can obtain a matrix S' preserving the positive eigenvalues and corresponding eigenvectors of S. However, the reader should note that this transformation implies losing some information.

We consider four ways to apply c-means to cluster genes from their similarities (distances) in GO: standard c-means (*cmeans*), a naive approach we call matrix c-means (*mcmeans*), kernel c-means (*kcmeans*) and c-means with *CSE* (*csecmeans*).

Briefly, c-means clustering (Hartigan, 1975) is a maximization of expectation algorithm that minimize the cost function shown in Equation 3.8. Next, we explain how we use c-means in the four cases outlined above:

- ***cmeans***: Given a symmetric distance matrix D, we apply c-means as described in equation 3.8. Notice, that D could be a non-metric distance, so we can expect to face convergence problems.
- ***mcmeans***: Given a symmetric distance or similarity matrix Y, we think of every row Y as an object to be clustered. Therefore, we apply c-means over the rows of Y.
- ***kcmeans***: Given a symmetric similarity matrix S, we get rid of negative eigenvalues to produce a kernel S', which is an inner product. We compute the distance matrix $D_{ij} = S_{ii} + S_{jj} - 2 * S_{ij}$ and then apply c-means to cluster. In this case, we avoid convergence problems.
- ***csecmeans***: Given a symmetric distance matrix D, we apply Constant Shift Embedding to get D' and then we apply c-means to cluster. In this case, we also avoid convergence problems.

When the method expects a similarity (distance) and we have a distance (similarity), we use the transformations described in Section 4.2.

Validation

After clustering is performed, we need to assess the quality of the obtained solutions. If we know the "true clustering", as in this case, two scores that can be used to evaluate the computed clustering are Jaccard and Minkowski (Tan et al., 2005).

Let T be the "true clustering" and C the computed clustering. We name n_{11} the number of pairs of data that are in the same cluster in both T and C, n_{01} the number of pairs that are in the same cluster only in T, and n_{10} the number of pairs that are in the same cluster only in C. The scores are defined as:

- **Jaccard Score**: $J(T,C) = \frac{n_{11}}{n_{11}+n_{01}+n_{10}}$. The higher $J(T,C)$, the better C as clustering result.
- **Minkowski Score**: $M(T,C) = \sqrt{\frac{n_{01}+n_{10}}{n_{11}+n_{10}}}$. The lower $M(T,C)$, the better C as clustering result.

These measures do not represent a percentage of hits, but a score of the global quality of the clustering. We use the score of a random clustering as a reference of how good the results are. Also, it can be expected Jaccard and Minskowski are congruent, i.e. the best clustering has higher J and lower M, but this does not always happen.

Datasets

Our experiment setup is intended to give us insights into which GO measure best describes gene proximity and which of the proposed clustering methods is the most appropriate and which GO ontology should be used in each case. In this study we use the annotations corresponding to the November, 2006 GO version.

In order to make the results comparable, the datasets are constructed using related families of proteins with approximately similar size. We also avoid overlapping between families, since we are using crisp methods to recover them.

With these goals, we build families of genes and proteins from the following sources:

- **CluSTr** (Kriventseva et al., 2001a) families (*C*). These families belong to the CluSTr database, which classifies *UniProt KW* (Bairoch et al., 2005) human proteins in groups of related proteins. Proteins in CluSTr are grouped according to pair-wise sequence comparisons with the Smith-Waterman algorithm. We use families of similar size to build each of the five datasets of sizes two, three, four, five, and six clusters. We denote them by *C2*, *C3, C4*, *C5* and *C6*.
- **MIPS** (Mewes et al., 1999) families (*M*). First, we select random nodes from MIPS database. Yeast genes annotated in the same node are considered to be in the same family. Finally, several random families of

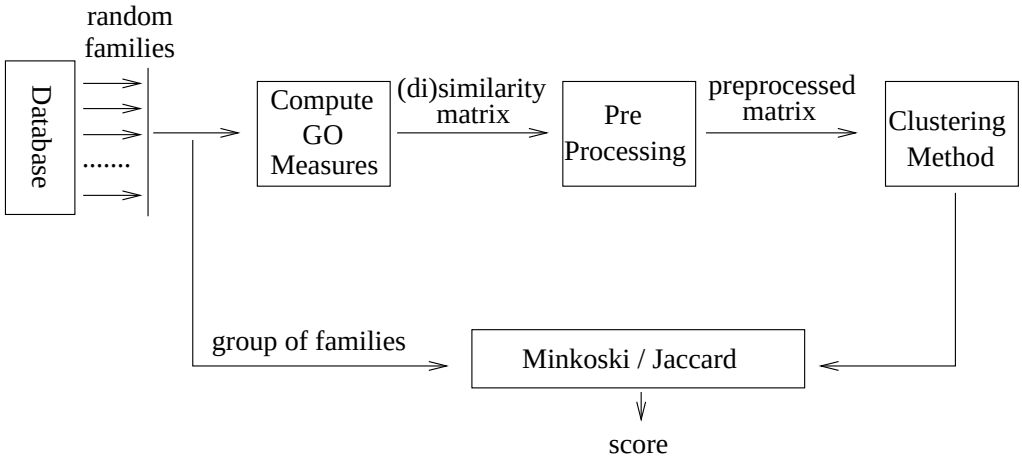

Figure 4.1: ***Workflow.***

approximately the same number of genes are grouped together to form *M3*, *M4*, *M5*, *M6* and *M7*.

- **Expression similarity families** (*E*). These families are obtained from well-known yeast *Saccharomyces cerevisae* clusters reported by Eisen et al. (1998). The clustering with the proposed ten families is considered: *E10*.

- **Pfam** (Bateman et al., 2002) families (*P*). These families are built by sequence alignment and profile-Hidden Markov Models (profile-*HMMs*). Selected a random yeast protein, the rest of the proteins are ranked according to the alignment distance. The closest proteins form a new family. Several families of similar size are used to build the final datasets: *P2*, *P3*, *P9* and *P17*.

- **Gene Sorter** (Hinrichs et al., 2006) yeast protein homology families (*H*). These families are based on the *BLASTP* E-value. Further information can be found in Kent et al. (2005). Same as with *P* families, after selecting a random protein, a family consists of the closest proteins to it. Families are grouped to form *H2*, *H3*, *H4*, *H5*, *H6*, *H7* and *H8*.

Figure 4.2 summarizes all the steps from family construction to clustering validation.

Results

For each experiment, we compute nine matrices by combining each of the three GO measures (Jiang, Lin and Resnik) with the annotations[1] in *BP* ontology, *MF* ontology and all three ontologies together ($ALL = BP + MF + CC$)[2]. The clustering methods *cmeans*, *mcmeans*,*kcmeans* and *csecmeans* are executed over these nine matrices.The obtained clusterings are compared with the true initial families with the Minkowski and Jaccard scores. The random method denotes the scores obtained by a random clustering of the genes and it should be compared with the other scores to evaluate the quality of clusters. In this paper, we show in tables 1 to 5 the results that most intuitively illustrate our findings[3].

[1]We did not take into account the Inferred from Electronic Annotations (IEAs) since they do not provide a reliable source of information

[2]We do not provide the results for *CC* since it presented poor results in terms of family recognition.

[3]We use * aside the value for the score when cmeans fails to converge.

Table 4.1: ***Jaccard scores for* C4.** *Best clusterings are obtained when we compute distances with MF ontology. The same holds for the rest of* C *families, since the genes are grouped according to molecular properties. kcmeans provides best results in this case.*

	ALL, 139 genes			*BP*, 124 genes			*MF*, 105 genes		
Methods	Jiang	Lin	Resnik	Jiang	Lin	Resnik	Jiang	Lin	Resnik
random	0.162	0.164	0.171	0.162	0.153	0.166	0.147	0.164	0.167
cmeans	0.261	0.350	0.396	0.316	0.312	0.297	0.657	0.512	0.466
mcmeans	0.188	0.382	0.387	0.268	0.240	0.305	0.650	0.607	0.649
kcmeans	0.320	0.356	0.346	0.295	0.301	0.335	**0.772**	0.627	**0.772**
csecmeans	0.280	0.335	0.280	0.338	0.294	0.297	0.682	0.430	0.513

Table 4.2: ***Jaccard scores for* M4.** cmeans *gives best results in* ALL *and* BP. *This was expected, since M families comes from MIPS biological function category. Again, the measures do not play an important role in the results.*

	ALL, 292 genes			*BP*, 288 genes			*MF*, 273 genes		
Methods	Jiang	Lin	Resnik	Jiang	Lin	Resnik	Jiang	Lin	Resnik
random	0.154	0.153	0.156	0.155	0.154	0.154	0.152	0.148	0.154
cmeans	0.436	0.464	0.629	0.375	0.560	0.505	0.286	0.287	0.330
mcmeans	0.164	0.618	0.626	0.440	0.615	0.532	0.346	0.318	0.321
kcmeans	**0.668**	0.646	0.675	0.667	0.666	0.624	0.340	0.325	0.282
csecmeans	0.457	0.459	0.558	0.553	0.494	0.433	0.333	0.238	0.309

Table 4.3: *Jaccard scores for* E10. *Very good results were obtained even with a high number of clusters selected from Eisen et al. (1998). As expected,* BP *provides the most stable results. Jiang is the worst measure in this case while Lin and Resnik performs similarly. The best clustering methods are* kcmeans *and* csecmeans.

	ALL, 292 genes			*BP*, 288 genes			*MF*, 273 genes		
Methods	Jiang	Lin	Resnik	Jiang	Lin	Resnik	Jiang	Lin	Resnik
random	0.092	0.087	0.087	0.086	0.092	0.089	0.093	0.088	0.086
cmeans	0.343	0.587	0.607	0.525	0.630	0.574*	0.269	0.585	0.606*
mcmeans	0.369	0.307	0.327	0.369	0.349	0.400	0.401	0.370	0.382
kcmeans	0.388	0.192	0.244	0.461	0.368	0.442	0.411	0.411	0.584
csecmeans	0.435	0.611	**0.727**	0.265	0.643	0.593	0.560	0.577	0.552

Table 4.4: ***Jaccard scores for* P17.** P *families are also distinguished according to molecular properties, so* MF *produces the best clustering again. Resnik and Lin seems to be more stable than Jiang.* kcmeans *beats the other clustering methods.*

	ALL, 292 genes			*BP*, 288 genes			*MF*, 273 genes		
Methods	Jiang	Lin	Resnik	Jiang	Lin	Resnik	Jiang	Lin	Resnik
random	0.070	0.067	0.071	0.073	0.076	0.073	0.077	0.075	0.070
cmeans	0.101	0.082	0.176*	0.068	0.136	0.188*	0.069	0.317	0.245*
mcmeans	0.109	0.151	0.189	0.142	0.165	0.172	0.223	0.230	0.232
kcmeans	0.129	0.105	0.217	0.174	0.188	0.162	0.271	0.310	**0.323**
csecmeans	0.065	0.080	0.161	0.068	0.177	0.166	0.069	0.243	0.253

Table 4.5: Jaccard scores for *H7*. Again, *MF* ontology provides the best results, since proteins were grouped according to homology. *kcmeans* outperforms clearly the other clustering methods.

	ALL, 292 genes			*BP*, 288 genes			*MF*, 273 genes		
Methods	Jiang	Lin	Resnik	Jiang	Lin	Resnik	Jiang	Lin	Resnik
random	0.100	0.103	0.112	0.111	0.116	0.118	0.110	0.122	0.119
cmeans	0.193	0.216	0.362	0.251	0.310	0.294*	0.517	0.413	0.422*
mcmeans	0.170	0.327	0.324	0.210	0.278	0.257	0.343	0.369	0.397
kcmeans	0.209	0.330	0.393	0.262	0.285	0.301	0.454	**0.540**	0.473
csecmeans	0.212	0.201	0.215	0.180	0.294	0.214	0.343	0.484	0.393

The performance of Resnik, Lin and Jiang measures across experiments is very similar. However, Resnik seems to produce the best scores in most of them. This is consistent with results of other experimental works Lord et al. (2003); Sevilla et al. (2005). Resnik also presents less variability but it usually has more convergence problems with the c-means algorithm than the other measures.

Next, we focus our discussion on the clustering methods:

- *cmeans* algorithm is not appropriate to be used directly with GO measures. As commented before, it does not converge in some of the cases. Therefore, we need one of the other methods if we intend to guarantee a stable result.
- *mcmeans* is the first proposed solution. It is clear from the tables that it produces the worst and most variable results.
- *kcmeans* produces the best results for many of the families. It also improves *cmeans* results in most cases, even though it involves loss of information with respect to the original similarity matrix.
- *csecmeans* also produces very good results, similar to those of *kcmeans*. This was expected since this method keeps clustering structure intact (Roth et al., 2003b), favoring family identification.

GO contains a lot of usable knowledge for genomic research. As expected, when we consider the molecular structure of proteins (experiments *C*, *H* and *P*), the *MF* ontology provides the best explanation of the families. Furthermore, when we look at the biological behavior of proteins (experiments *E* and *M*), it is the *BP* ontology which performs the best. For example, families according to gene expression (Table 4.3) are very well recovered, even when the dataset has many genes (over 260) and clusters (10). Also, the performance is not dependent on the organism used for the study, obtaining comparable results for sets of human and yeast proteins.

Measures computed over all of the GO ontologies (*ALL*) provide variable results. Sometimes results are worse (Table 4.4) and sometimes better (Table 4.2). An explanation for this is that the *CC* ontology might hinder recognition in some experiments, since its definition (location of the protein in the cell) is not clearly related to the family construction methods we use.

As a conclusion, GO contains knowledge about molecular alignments of proteins (*C* families), gene expression (*E* families), protein homology according to blast (*H* families), other ontologies (*M* families), and Pfam (*P* families). Computing the different measures defined over GO, we can use such knowledge to analyze sets of annotated proteins.

Cluster Analysis

In this section it is shown that GO separates fairly well different families of proteins. Here, we consider the biological meaning of the results intending to explain of protein misclassification, discussing the analysis of the best cluster for the family *C4*[4], focusing on the protein misplacements and other interesting considerations.

If a protein is not clustered into its appropriate family, there exist four possible reasons:

- The clustering method failed and made the protein belong to a wrong group. Therefore, different clustering methods should be considered.
- The GO measure was not appropriate since the computed distances/similarities did not capture the family structure. The clustering method, although working well, would not be able to place the protein into its family.
- The GO annotations of the protein were very different from the GO annotations of the other proteins in the family. In this case, we should ask whether the protein actually belongs to this family or not.
- The protein was poorly annotated in GO. Therefore, the lack of information would lead to a random behavior of the method.

In Table 4.6, a summary of the best cluster for the family *C4* is shown. We use all three ontologies (*BP*, *MF* and *CC*) in our analysis. Reader should note that, as exposed in section 4.2, *C* families are constructed according to the molecular structure of proteins. Therefore, *MF* usually provides the best explanation.

[4]The best cluster for the family *C4* is using the Resnik measure and the *kcmeans* method

In *C4*, *Cluster 1* contains 23 proteins and 22 of them belong to *Family 1*, which groups 23 proteins. This implies that there is one missing protein (*PTK2B*) and one non-expected protein (*ADRA1A*) in the cluster. The clustering method does not locate *PTK2B* together with its hypothetical family. An analysis of the GO annotations of the family reveals that *PTK2B* is the only protein not annotated to *receptor activity* (*GO:0004872*). However, it is the only one annotated to *nucleotide binding* (*GO:0000166*). Moreover, this protein is the only one not found in the *membrane* node (*CC, GO:0016020*) but in the *intracellular* node (*CC, GO:0054430*). This clearly shows that our method is working properly for this protein. On the other hand, *ADRA1A* is included in *Cluster 1* since it shares the main annotations with the proteins in that cluster.

In *Cluster 2*, there are 25 proteins, all of them belonging to *Family 2*. However, *Family 2* contains 28 proteins, so we have three missing proteins (*ADRA1A*, *HGF* and *LPA*). As explained above, *ADRA1A* belongs to *Cluster 1*. *ADRA1A* is also the only protein not present in the *peptidase activity* node (*GO:0008233*) together with *THRB*. Studying *THRB*, we observe that it can be considered as an outlier since it has its own different set of annotations, e. g. *transcription regulator activity* (*GO:0030528*). However, *THRB* is clustered in its hypothetical family because its annotations are more related to this family than to the rest of the families. *HGF* and *LPA* share the main annotations in *Family 2*, e. g. *catalytic activity* (*GO:0003284*) so, in order to explain why they are misplaced, we need to explore the clusters where they are assigned.

Cluster 3 has 27 proteins, 25 of them in *Family 3*, which also has 27 proteins. The two missing proteins are *CCR1* and *CCBP2*, and the two extra proteins are *HGF* and *MCHR1*. We observe that *HGF* only shares annotations with two proteins in this cluster in the *hidrolase activity* node (*GO:0016787*). Therefore, considering the analysis of *Cluster 2*, the clustering method is not working as expected. In this case, applying fuzzy cmeans gives us a correct classification of this protein. *MCHR1* presents a small number of annotations. Furthermore, it shares an annotation at a very low GO level (high specificity) with two proteins in *Family 2*, e. g. the *neurotransmitter receptor activity* node (*GO:0042923*). Consequently, the GO measure assigns a very small

Table 4.6: ***Best clusters for experiment*** **C4.** *(Jaccard = 0.772).*

fam i	cluster j	% i in j	len(i)	misplaced / cluster j
1	1	22 // 23 (96%)	23	{*PTK2B* / 4}
2	2	25 // 25 (100%)	28	{*ADRA1A* / 1, *LPA* / 4, *HGF* / 3}
3	3	25 // 27 (93%)	27	{*CCBP2* / 4, *CCR1* / 4}
4	4	26 // 30 (87%)	27	{*MCHR1* / 3}

distance between these three proteins, and the clustering method groups them together. For *CCR1* and *CCBP2*, it seems that they should belong to this cluster since they collect most of the general annotations of the family. This fact will be confirmed after the analysis of the last cluster.

Cluster 4 presents 30 proteins, 26 of them belonging to *Family 4*, which has 27 proteins. There is one missing protein (*MCHR1*) and four misplaced proteins (*CCR1*, *CCBP2*, *LPA* and *PTK2B*). *MCHR1* is not annotated to *catalytic activity* node nor in *transferase activity* node (*GO:0003824* and *GO:0016704*), while the other proteins in the family are annotated to one or both of these GO terms. This shows again that the method works according with the information provided. *CCR1* and *CCBP2* do not have important annotations in common with the rest of the proteins in this cluster, however they share annotations with those in *Cluster 3*. In this case, the method does not work properly but, unlike in the case of *HGF*, other clustering methods (e. g. fuzzy kmeans) provide the same results. It is clear then that the GO similarity measures do not work as expected in this case. An study of the annotations of *PTK2B* shows that it shares most of them with the proteins in this family, mainly *catalytic activity* and *transferase activity* (*GO:0003824* and *GO:0016704*). Therefore, the method works properly for this protein. Finally, *LPA* has very few annotations and a correct classification using GO is not feasible.

In Table 4.7 we summarize the information discussed in this section. Observe that for the dataset *C4* the method groups correctly 102 out 105 proteins (97%).

Table 4.7: ***Proteins misplaced for experiment*** **C4.**

Protein	Reason for the misplacement	Method works
ADRA1A	Annotations close to the other family	Yes
CCBP2	GO similarity measure	No
CCR1	GO similarity measure	No
HGF	Cluster method	No
LPA	Poor GO annotations	Yes
MCHR1	Poor GO annotations	Yes
PTK2B	Annotations close to the other family	Yes
THRB[5]	Outlier: very different GO annotations	Yes
All the rest	————————	Yes

4.3 Fuzzy Semantic Similarity Measure for Gene Ontology

In this section we present a new fuzzy similarity measure (FSM) for computing the similarity of two gene products annotated with terms from an ontology. The measures mentioned above do not take into account the reliability of the source of information. In order to do it we propose the aggregation of the information content and the evidence code together. In particular, the evidence codes are translated into weights by means of a Genetic Algorithm (GA).

Adapting Fuzzy Approaches

Fuzzy techniques have been applied in order to solve the problem of computing the semantic similarity between GO gene products. Keller et al. (2004) proposed a Fuzzy Measure-based Similarity (FMS) for computing the similarity of two gene products annotated with terms from an ontology. The advantage of FMS is that it takes into consideration the context of the whole set when computing the similarity.

Definition

The similarity measure will be based on the concept of a general fuzzy measure using the ontology annotations of the given gene products. Let G be a fuzzy measure over a set X (the finite universe of discourse with the subsets $A, B, ...$) satisfying the conditions shown in equation 3.13 and equation 3.14.

Given two gene products, G_1 and G_2, we can consider them as being represented by collections of terms:

$$G_1 = \{T_{11}, \ldots, T_{1n}\}, G_2 = \{T_{21}, \ldots, T_{2n}\}. \tag{4.13}$$

In this context, the terms in a combined set describing two gene products will be considered as "information sources" that support the similarity of two genes. Each annotation T_i will have a fuzzy density value which is interpreted as the importance of the single information source T_i in defining the gene products it belongs to. General fuzzy measures are broad, but it is often the case that the densities can be extracted from the problem domain or supplied by experts. The key to using fuzzy measures involves finding those that can be built out of the densities (Keller et al., 2004), such as Sugeno λ-measures (see Section 3.5).

For constructing the fuzzy densities we use the information-theoretic principles (Resnik, 1995). It has been demonstrated that this type of approach is less sensitive and in some cases not sensitive to the problem of not uniform distribution in the ontology (Budanitsky and Hirst, 2001). Similarly to what was discussed in Section 4.1, for each term, T, we calculate $P(T)$ as the probability of finding T or a child of T in the ontology and then we compute $-\log(P(T))$. This creates a problem when $-\log(P(T)) > 1$. In order to avoid this, we normalize the result dividing by the maximum specificity i.e. $-\log(\frac{1}{\text{Total number of annotations}})$. Finally, for a term T we obtain the importance, $I(T)$, by multiplying its value by a factor depending on the best evidence code annotated in the term:

$$I(T) = \frac{-\log(P(T))}{-\log(P(min))} EC(T). \tag{4.14}$$

Previous studies did not take into account the reliability of the source of information. Using the evidence codes to compute the importance of the terms makes the measure more natural and intuitive since we adjust the importance depending on the credibility of the source of information. In GO there are currently 17 evidences codes[6] (Table 2.3) . In the GO web page it is said that users can and should form their own conclusions as to the reliability of each type of evidence and each individual annotation.

Following this idea we assigned a numeric value in [0, 1] for each evidence code. As we said before, these weights will be multiplied by the information content to get the importance of each term. Hence, we have to solve a search problem in order to find the values for the codes that make the FSM provide the best result. Such problem consists of finding twelve values in [0, 1] that allow the FSM to get the best similarity values. In order to solve this problem we propose the use of a Genetic Algorithm. Genetic algorithms have been widely proved to provide good solutions in this kind of problems with an acceptable time consuming. They have also been proved to outperform other techniques when solving complex problems with many parameters (Pardalos and Resende, 2002). Another advantage of genetic algorithms is that they bring a set of solutions instead of a unique solution[7]. The pseudo-code of the genetic algorithm we used is provided in Figure 4.3.

The main features of the GA are the following:

- Each individual in the population is an array of 10 values. Each value of the array is a weight for the corresponding evidence code.
- The population is initialized randomly.
- Evaluation: for each individual, a similarity matrix is computed and the procedure described in Section 4.4 carried out to get the performance of the FSM with the corresponding weights.
- Mutation operator: changes each weight with probability Pm by a random number in [0, 1].

[6] In what follows, we work with the 11 evidence codes available at the time of moment of developing this work.

[7] See Section 3.7 for more on GAs

```
GENETIC ALGORITHM
1  t ← 0
2  Initialize P(t)
3  Evaluate P(t)
4  while (stop condition not satisfied)
5        do t ← t+1
6           Select P'(t − 1) from P(t − 1)
7           Recombine P'(t − 1) to obtain P(t)
8           Evaluate P(t)
```

Figure 4.2: ***Pseudocode for the genetic algorithm.***

- The selection process is carried out by means of the tournament selection with size 2.
- For recombining the selected population $P'(t-1)$ in each iteration we use the Wright's heuristic crossover (Wright, 1991).
- For each iteration, the two best solutions of the current population are directly copied in the next population.
- We ignored the IEA annotations since they do not represent a reliable source of information.

4.4 Protein Classification using Gene Ontology

One of the central problems in computational biology is the classification of proteins into functional and structural families based on sequence homology. In this section we evaluate the performance of each GO semantic similarity measure based on the effectiveness in recognizing protein families.

Methods

The protein families were downloaded from the CluSTr database of UniProt (Bairoch et al., 2005), which offers a resource for an automatic classification of UniProt Knowledgebase proteins into groups of related proteins. In

the CluSTr database, the clustering is based on analysis of all pair-wise sequence comparisons between proteins using the Smith-Waterman algorithm (Smith and Waterman, 1981) between protein sequences, for various levels of protein similarity (Kriventseva et al., 2001b).

Families within the CluSTr database are univocally identified with an ID code, e.g. HU:3515:141.1. The last number in the ID (141.1) is the z-score for the corresponding family. The higher the z-score is, the more differentiated the family is. We wanted to compare how the different FSMs help to recognize the families according to their GO annotations. Therefore, we selected families of human proteins with medium-high similarity in order to have groups of proteins with some common properties (if not it could lead to random solutions), but not so well differentiated (which could lead to a trivial problem).

We combined the obtained families to form several datasets. The most representative datasets are shown in Table 4.10. The annotations were taken from the GO database, release of October of 2005. In our case, proteins are groped together into families according to their functional similarities. Due to this, we used the molecular function ontology to compute the similarity between the proteins and hence to obtain the similarity matrix. All selected families are comprised of human proteins. Therefore, only the annotations of UniProt of *H. Sapiens* gene products were considered for the experiments.

For the training of the genetic algorithm we used different sets of families in order to obtain as much generalization as possible. In validating their results, previous works implemented a hierarchical clustering algorithm that groups the proteins according to the similarity matrix obtained before (Keller et al., 2004). This approach is too restrictive since it forces the elements to belong or to not belong to a cluster i.e. a family. In reality a protein might be a member of more than one family, and it could belong to one family more than to another if it shares more properties with the proteins from the former than with those from the latter.

Fuzzy clustering could solve the problem of belonging to several families; however, these techniques have the restriction that the sum of the membership degrees is always equal to one. Therefore it creates the problem that a

protein cannot be part of two or more families to a high degree at the same time. Another problem would be the case that a protein does not belong at all to any family with which we are working. In that case, fuzzy clustering approaches also do not work properly since they assign at least one high membership degree in order to fulfil the property that the sum of degrees equals to 1.

We implemented a possibilistic clustering algorithm to deal with these problems. Like fuzzy approaches, possibilistic clustering gives a membership degree to each element for each cluster. However, possibilistic approaches do not have restrictions related to the degrees (see Section 3.3). It solves the problems mentioned above since the possibilistic approaches allow a protein to belong either to several families with a high membership degree or to belong to no family at all. This makes the approach much more realistic and natural since it is more similar to the real situation we are addressing. Therefore, the possibilistic algorithm returns a list of genes with their membership degrees for each cluster. In order to determine the cluster to which a gene belongs, we follow the following rules:

- If a gene presents a membership degree greater than 0.8 for one or more clusters, the gene belongs to all of those clusters.
- If a gene presents a membership degree greater than 0.4 and lower than 0.8 for one or more clusters, the gene belongs to the cluster with the highest membership degree.
- Otherwise, the gene does not belong to any cluster in the experiment.

In addition, if there are too many genes associated to several clusters in a given experiment, the semantic similarity measure might not be suitable for such experiment since it is not able to distinguish the families in a correct manner. We considered that if more than 20% of the genes belonged to more than one cluster, the number of correctly classified genes was 0. All these thresholds were selected after observing the clusters obtained in many runs of the algorithm. We saw that these are good values to get clusters with an acceptable size and biological meaning.

Once we got the clusters, we calculated the proportion of genes of each family that belonged to each cluster to assign a family to every cluster. A

Table 4.8: ***Evidence Weights.***

TAS	IDA	IMP	IGI	IPI	ISS	IEP	NAS	ND	NR	IC
0.930	0.933	0.120	0.030	0.031	0.271	0	0.577	-	-	0.04

family was assigned to the cluster with the highest proportion of genes on it. Finally, we created a measure to compare the quality of the family recognition. This measure counted the number of genes that had been assigned to the cluster to where their family was previously assigned. The more genes are correctly located, the better the measure has performed.

Results

To demonstrate the applicability of our approach we compared the results of our semantic similarity measure to the results of the Resnik, Jian and Conrath, Lin, and Keller SS measures. We took a variety of families, combined them in different manners, and applied the methodology explained before in order to carry out such comparison. We run the GA explained above for obtaining the weights for the evidence codes. Results of the GA are shown in Table 1. The weights follow the loose hierarchy proposed by the GO Consortium (Harris et al., 2004).

We show in Table 4.10 the four most representative experiments. A and B refer to experiments of proteins from two families, C and D refer to experiments of proteins from three families. The IDs of the families correspond with those obtained from CluSTr Database of UniProt, with the corresponding family z-score indicated at the end (i.e. HU:1995:116.8 implies a z-score equals to 116.8).

In Table 4.10 can be seen that the fuzzy semantic similarity measure proposed outperforms the results of the rest of the methods. This was expected and demonstrates that incorporating the information of the technique used for the experiments (i.e. evidence codes) makes the measure more reliable and accurate. It confirms that the GO annotations are a good source of information and should be used to address bioinformatics problems. Our measure

Table 4.9: ***Semantic similarity measures over BP and CC.***

Data set 1 *Data set 2*	HU:1995:116.8 HU:3515:141.1	HU:1897:110.7 HU:398:62.2
Ontology	biological process	cellular component
Family	A_1	A_2
Resnik	30 (53%)	28 (49%)
Lin	27 (47%)	25 (44%)
Jiang	33 (58%)	28 (49%)
Keller	32 (58%)	25 (44%)
FSM w/ EC	33 (58%)	17 (30%)

is the most stable one, and even when the families are not so well differentiated, it still has a high proportion of correctly classified genes (always greater than 70%). For instance, the second family in experiment B presents a medium-low z-score (62.2), 61% of the genes in the experiment belong to this family, and our measure still classifies 70% of the genes correctly.

In order to verify the correct performance of our method, we studied the behavior of the measures when working over the other two subontologies: biological process and cellular component. Results are shown in Table 4.9.

In Table 4.9 we can see that the performance of all the measures decreases substantially compared to the performance for the molecular function subontology. This was expected since, as explained above, families are created according to similarities in the protein sequences, and neither the process where a protein participates, nor the place in the cell where it is located, are directly related to the sequence similarity. The decrease is more significant for the cellular component subontology. This was expected too because besides the problem with the family construction, in this subontology there are less annotations than in the other two, and therefore the SS measures cannot deal with the problems properly.

Table 4.10: ***GO Semantic Measures Performance.***

Data set 1 *Data set 2* *Data set 3*	HU:1995:116.8 HU:3515:141.1 -	HU:1897:110.7 HU:398:62.2 -	HU:1995:116.8 HU:3515:141.1 HU:398:62.2	HU:1995:116.8 HU:1897:110.7 HU:398:62.2
No. Genes	$23+35=58$	$26+41=67$	$22+35+41=98$	$22+26+41=89$
Family	A	B	C	D
Resnik	38 (67%)	45 (67%)	76 (78%)	63 (71%)
Lin	36 (63%)	39 (58%)	68 (69%)	57 (64%)
Jiang	47 (82%)	37 (55%)	77 (79%)	60 (67%)
Keller	38 (67%)	44 (66%)	66 (67%)	61 (69%)
FSM w/ EC	54 (95%)	47 (70%)	83 (85%)	66 (74%)

4.5 Concluding Remarks

In this chapter, we analyzed the classical GO semantic measures (Resnik, Jiang and Lin). Starting with different groups of related proteins according to different criteria (sequence similarity, expression level, protein-protein interaction, and protein homology), we proposed a method to analyze the performance of GO similarity measures in recognizing these relationships.

We used different GO ontologies for computing similarities/distances between proteins, and all three together. We calculated the (di)similarity matrices corresponding to three (di)similarity measures to see how they perform in different experiments. Finally, we applied four clustering methods (cmeans, mcmeans, kcmeans and csecmeans) to these matrices, we validated the results with two scores (Jaccard and Minkowski), and we analyzed the obtained clusters in order to observe the correlation between GO and the different families constructed.

We demonstrated that two of the clustering methods, kcmeans and csecmeans, provide better results than the others and they should be considered when dealing with imprecise data. Furthermore, we showed that selecting an appropriate GO partition is a key issue for each problem, e.g. *molecular function* ontology should be used when dealing with molecular structure of proteins. Thus, we proved that this knowledge can be applied to recognize related proteins, to validate other assessments and to incorporate it to other experiments. In that sense, kernel methods, like kcmeans, are very appropriate for integration of knowledge, so the proposed kernelization of the similarity matrix is specially interesting.

In addition, we presented a new fuzzy similarity measure for computing the similarity of two gene products annotated with terms from an ontology. The alternative measures discussed through this chapter do not take into account the reliability of the source of information. In order to do it, we aggregated the information content and the evidence code together. The evidence codes were translated into weights by means of a genetic algorithm.

Our proposed measure for GO annotations provides results that outperforms previous techniques and are consistent with the biological meaning of

the annotations and with the expected results. It confirms that semantic similarity measures are adequate to face bio-ontology problems. By taking into account the different ways for annotating a gene product (evidence codes) in the GO, we have provided a more realistic approach for calculating the similarity between gene products since we have challenged the techniques used to carry out the experiments.

Semantic measures depend on the annotations in the GO. A source of noise is that some of the proteins we used were not annotated in the GO or that they had poor annotations. These annotations are being increased and improved by the scientific community everyday. As we know more about genes, GO will be more complete and better results are expected.

CHAPTER 5

Comparing TFBS Motifs

One of the main goals in computational biology is to understand how expression of genes is controlled, and to unravel gene regulatory networks. Cells control the abundance and activity of proteins by means of diverse factors in which transcription regulation plays a central role. Transcription factors (TFs) play a key role in gene regulation by binding to target sequences called transcription factor binding sites (TFBSs). In silico prediction of potential binding of a TF to a binding site is a well-studied problem in computational biology. A common question in the context of *de novo* motif discovery is whether a newly discovered, putative motif resembles any previously discovered motif in an existing database. Fuzzy concepts are specially suitable for this problem. In this chapter we introduce the problem of comparing TFBS motifs and we define fuzzy measures motifs that outperforms classical probabilistic measures.

This chapter is organized as follows. Section 5.1 gives an introduction to the problem of comparing TFBS motifs. Section 5.2 discusses the different representations for motifs. Section 5.3 introduces the existing probabilistic measures. In Section 5.4 we show how to adapt different classes of fuzzy measures for our purpose. Section 5.5 present an exploratory study the measures previously commented. In Section 5.6 we conclude the chapter.

5.1 Background

As discussed in Section 2.1, transcription is the process of transcribing a DNA sequence into its corresponding RNA sequence. Multiple events are involved in the initiation of transcription of a gene. One of the most important ones is the binding of several proteins, called transcription factors (TFs), to DNA near the gene, called transcription factor binding sites (TFBSs). TFBSs are usually located close to the transcription start site (TSS) of the gene and upstream from it. Additionally, in some cases TFBSs can be found downstream the TSS or, in rare instances, even within exons (Pan, 2006). These interactions between DNA and proteins play a crucial role in controlling the expression of the genes by activating or inhibiting the transcriptional machinery.

TFs generally have distinct preferences towards specific target sequences. Sometimes a given TF can bind to only one TFBS, but usually the same TF can bind to different DNA sequences. For a known group of binding sites of the same TF, it is possible to construct a model that describe the binding preferences. These groups of TFBSs are known as *regulatory motifs*, and there are a number of studies that discuss the most appropriate representation, which will be introduced in the following section.

The identification of binding sites bound by transcription factors is therefore a key problem in predicting transcription regulation. With the emergence of high-throughput technologies (e.g. ChIP-chip assays, DNA microarrays, etc.) numerous algorithms for finding motifs have appeared (for a review see Das and Dai (2007)). The recognition of *de novo* TFBSs usually includes the issue of comparing putative motifs with one another and with motifs that are already known. In addition, these algorithms usually filter their outputs in order to improve their significance, e.g. merging similar motifs. However, the outcome of these tools, particularly when dealing with large datasets, is usually presented as a large list of motifs that require further post-processing in order to make it meaningful. Methods for comparing motifs are usually applied to give biological significance to the outputs of these programs. This is usually done by comparing the putative motifs provided by these algorithms against known motifs reported in motif databases

such as JASPAR or TRANSFAC (Sandelin et al., 2004a; Matys et al., 2006). Unveiling these relationships might be crucial for the design of appropriate biological experiments.

The existing motif discovery algorithms make use of different strategies to overcome drawbacks of other approaches, usually implying new or different limitations. One common approach involves using several of these algorithms and compounding their outputs (Tompa et al., 2005). In this case, motifs found by different algorithms can either correspond to the same TFBSs or to different ones, making the compounded result very noisy and imprecise. This suggests a need for comparison methods for finding similar motifs to be either removed or merged into a new motif.

5.2 Motif Representations

Many studies discuss the advantages of different regulatory motifs representations (Osada et al., 2004). Regulatory motifs representations aim to describe the binding affinity of the TFs, derived from a multiple alignment of confirmed binding sites for a given transcription factor. Here, we give a brief introduction of the most popular representations for DNA motifs:

- **Consensus Sequences.** It represents a minimum nucleotide sequence of the most common base (although not necessarily identical) in different related binding sites (see Figure 5.1).
- **Extended Consensus Sequences.** Previous notation may be extended if it is necessary to represent more than one base match per position (Figure 5.1). In this case, the International Union of Pure and Applied Chemistry (IUPAC) provides an alphabet that encoding each subset of the four nucleotides (Table 5.1).
- **Position Frequency Matrices (PFMs).** They are used to record the position-dependent frequency of each residue or nucleotide within the motif (Figure 5.2b).
- **Position Weighted Matrices (PWMs).** They are derived from PFMs, and they contain score values that give a weighted match to any given

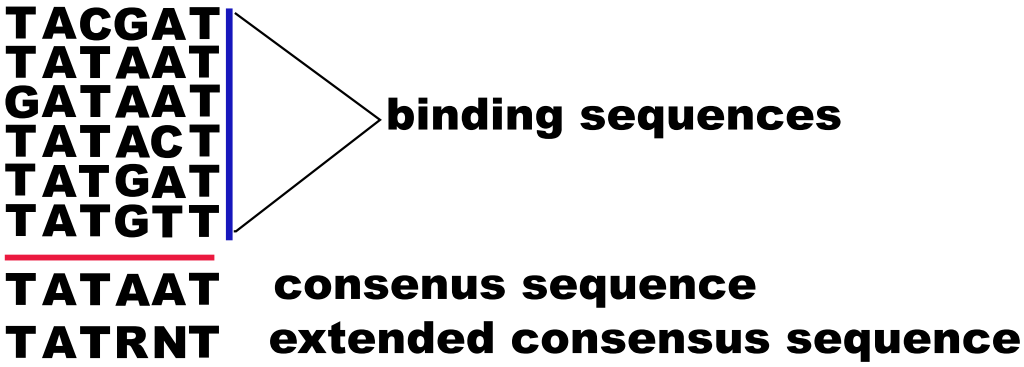

Figure 5.1: *Consensus sequence for TFBSs.*

substring of fixed length, converting normalized frequency values to a log-scale (Figure 5.2c).

- **Sequence Logos.** They provide a very intuitive visual representation of motifs by means of a graphical representation proposed in Crooks et al. (2004) (Figure 5.2d).

Table 5.1: *IUPAC codes for extended consensus sequences.*

Code	Corresponding class
A	[A]
C	[C]
G	[G]
T	[T]
R	[GA]
Y	[TC]
K	[GT]
M	[AC]
S	[GC]
W	[AT]
B	[GTC]
D	[GAT]
H	[ACT]
V	[GCA]
N	[ACGT]

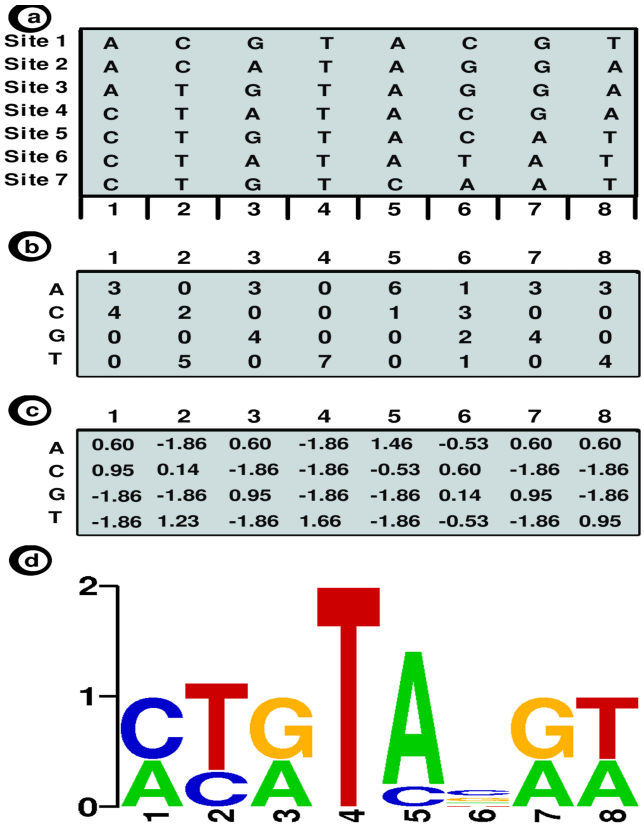

Figure 5.2: ***Motif representations.*** *a) Binding sequences, b) Position Frequency Matrix (PFM), c) Position Weight Matrix (PWM), d) Sequence Logo.*

5.3 Probabilistic Measures

The most common strategy for comparing motifs relies on the assumption that the columns of the matrices are probability distributions. Thus, most measures between motifs are based on statistical techniques that test whether the different columns belong to the same distribution. Pietrokovski (1996) used a straightforward algorithm based on the Pearson correlation coefficient (PCC). Wang and Stormo (2003) proposed the average log-likelihood ratio (ALLR) to compare between motif columns. Schones et al. (2005) made the comparison by means of a Pearson χ^2 test (PCST). They also proposed the Fisher-Irwin exact test (FIET) which provided poorer results. In addition, the Kullback-Leibler divergence (KLD) was used to compare motifs (Roepcke

et al., 2005). Rather than comparing distributions, Choi et al. (2004) used the euclidean distance (ED) between columns, obtaining promising results. Next we briefly introduce such measures.

Pearson correlation coefficient

Pietrokovski (1996) first introduced the Pearson correlation coefficient for comparing motif columns:

$$PCC = \frac{\sum_{b \in B}(b_{C_1} - \overline{C_1})(b_{C_2} - \overline{C_2})}{\sqrt{\sum_{b \in B}(b_{C_1} - \overline{C_1})^2 \sum_{b \in B}(b_{C_2} - \overline{C_2})^2}}. \tag{5.1}$$

The correlations of all the columns are summarized using the mean.

Average log-likelihood ratio

Wang and Stormo (2003) defined the Average log-likelihood ratio (ALLR) statistic to perform motif columns comparisons, which is the sum of two log-likelihood ratios. ALLR is defined as:

$$ALLR = \frac{\sum_{b \in B} N_{b_{C_1}} log\left(\frac{b_{C_2}}{p_b}\right) + \sum_{b \in B} N_{b_{C_2}} log\left(\frac{b_{C_1}}{p_b}\right)}{\sum_{b \in B}(N_{b_{C_1}} + N_{b_{C_2}})}, \tag{5.2}$$

where p_b is the prior for base b. To compare multiples columns, the scores of single columns are summed.

χ^2 test

χ^2 test was proposed by Schones et al. (2005) for comparing motifs. This test is computed under the hypothesis that the columns are observations from the same distribution. The p-value is computed from this χ^2 score with 3 degrees of freedom:

$$\chi^2 = \sum_{j=C_1,C_2} \sum_{b \in B} \frac{(N^o_{jb} - N^e_{jb})^2}{N^e_{jb}}, \tag{5.3}$$

where N^o_{jb} is the observed number of base b at position j, and N^e_{jb} is the expected number of base b at position j (see the work by Schones et al. (2005) for more details). The p-value is considered as an additive score.

Kullback-Leibler divergence

Kullback-Leibler divergence has been used to determine similarities between motifs (Roepcke et al., 2005). Its symmetric form is:

$$KLD = \frac{1}{2}\left(\sum_{b\in B} b_{C_1} log\left(\frac{b_{C_1}}{b_{C_2}}\right) + \sum_{b\in B} b_{C_2} log\left(\frac{b_{C_2}}{b_{C_1}}\right)\right). \tag{5.4}$$

Multiple columns are compared averaging column-to-column divergences.

5.4 Adapting Fuzzy Measures

In recent years, it has been seen that the inherent uncertainty and noise that characterize biological data cannot always be modeled sufficiently well by probabilistic approaches and that, consequently, alternative models for gathering this uncertainty may be required. Furthermore, in the context of motif comparisons, the utilization of PFMs as a representation of the binding preferences of the TFs inherently includes imprecision. In addition to the usual missing values and noisy data associated with biological data, there exist some *hidden* factors apart from the DNA sequence itself that affect the binding preferences of TFs, e.g. cooperative binding and chromatin structure (Lam et al., 2008). Moreover, an arbitrary threshold must usually be chosen in the construction of a PFM itself.

As discussed in Chapter 3, fuzzy set theory, proposed by Zadeh (1965), is especially suitable for dealing with imprecise, noisy and uncertain environments. Fuzzy set theory has been previously used in bioinformatics, however there exist fields where very few or none fuzzy approaches have been applied. During last years, some works have appeared that integrate fuzzy solutions to solve biological problems like microarrays analysis, proteins location, understanding of genomes, etc., showing promising results

(Pan, 2006; Huang and Li, 2004; Lopez et al., 2008). Quantifying similarity within the framework of fuzzy set theory is a crucial concept in approximate reasoning. Similarity measures have been successfully applied in different areas including expert systems, information retrieval or intelligent database systems (Cross and Sudkamp, 2002).

In our case, fuzzy concepts are especially suitable for the tasks of motifs comparison and detection. For example, a given TF might bind to more than one TFBSs, presenting more affinity to bind to some DNA patterns than to others. Therefore, in the matrix representation of the motifs, the binding preferences of each position (column) can be thought as the fuzzy membership degrees to sets of the four DNA nucleotides (A, C, G, T).

In this section we discuss the adaptation for our problem of different classes of classical measures for fuzzy sets including set-theoretic (Jaccard's method (Jaccard, 1908)), proximity-based (Minkowsky's r-metric) (Zwick et al., 1987), and angular coefficient-based (Bhattacharyya distance) (Bhattacharyya, 1946). In addition, we also consider a more recent DNA sequence-oriented dissimilarity measure (Torres and Nieto, 2003). All but the last measure are column-to-column measures and therefore the notation stated above holds. Notation for the remaining measure is given below.

Definitions

Set-theoretic measure: Jaccard coefficient

The Jaccard coefficient is an unparameterized ratio model of similarity (Jaccard, 1908). It is also known as index of communality. The Jaccard coefficient for two PFMs columns is:

$$S_J(C_1, C_2) = \sum_{b \in B} \frac{\mid b_{C_1} - b_{C_2} \mid}{max\left(b_{C_1}, b_{C_2}\right)}. \tag{5.5}$$

For PFMs composed by multiple columns, the average of the obtained measures is considered.

Angular coefficient-based: Bhattacharyya distance

Bhattacharyya distance measures the cosine of the angle between two vectors when the values in each vector are standardized as deviates from the mean of the membership function (Bhattacharyya, 1946). This cosine is taken as the corresponding similarity measure. In the case of two PFMs columns:

$$S_B(C_1, C_2) = \frac{\sum_{b \in B} \left(b_{C_1} \cdot b_{C_2} \right)}{\left(\sum_{b \in B} \left(b_{C_1} \right)^2 \right)^{\frac{1}{2}} \cdot \left(\sum_{b \in B} \left(b_{C_2} \right)^2 \right)^{\frac{1}{2}}}. \tag{5.6}$$

To compare multiple columns, the scores of single columns are averaged.

Proximity-based measure: Minkowsky r-metric

The distance between the partial membership functions of fuzzy sets A and B over a finite universe of discourse $U = \{u_1, u_2, ..., u_n\}$ may be measured using a Minkowsky r-metric (Zwick et al., 1987). A fuzzy set A is represented by a point $[\mu_A(u_1), ..., \mu_A(u_n)]$ in the n-dimensional space. In our case $U = \{A, C, G, T\}$ and the distance is:

$$d_r(C_1, C_2) = \left(\sum_{b \in B} | \, b_{C_1} - b_{C_2} \, |^r \right)^{\frac{1}{r}}, r \geq 1, \tag{5.7}$$

with $r = 1$ d_r becomes the Hamming distance; $r = 2$, the Euclidean distance; and with $r = \infty$, the dominance metric. More on this topic can be found in Cross and Sudkamp (2002). Although further studies are needed for obtaining an optimal value for r, in this work we selected an arbitrary value of $r = 1.3$ which provided the most consistent results throughout the experiments. Again, to compare multiple columns, the scores of single columns are averaged.

DNA sequence-oriented

In Torres and Nieto (2003) polynucleotide chains are transformed to ordered fuzzy sets to define a so-called fuzzy polynucleotide space (FPS). Also, a

Table 5.2: ***PFM toy example.***

#	1	2	3
A	0.7	0.2	0
C	0.1	0	1
G	0	0	0
T	0.2	0.8	0

fuzzy measure for DNA sequences in this space is proposed based on the ideas found in Zadeh (1965). In order to compute this measure the authors proposed the controversial idea[1] of mapping the PFMs into a point in the 12-dimensional unit hypercube $[0,1]^{12}$. For the PFM shown on Table 5.2 this mapping is $M_T = (0.7, 0.1, 0, 0.2, 0.2, 0, 0, 0.8, 0, 0, 1, 0)$. The fuzzy polynucleotide space measure (FPSM) for two PFMs is defined as:

$$FPSM(M_1, M_2) = \frac{\sum_{i=1}^{12} | M_{1_i} - M_{2_i} |}{\sum_{i=1}^{12} max\left(M_{1_i}, M_{2_i}\right)}, \tag{5.8}$$

where M_1 and M_2 are the mapping of the two PFMs to be compared. More details can be found in Torres and Nieto (2003).

5.5 Probability and Fuzzy Measures Analysis

In this section we present an exploratory study of the fuzzy measures for DNA motifs presented in the previous section. We compare their results with those obtained by the probabilistic measures introduced in Section 5.3.

Similar Columns Recognition

We wanted to test the performance of the methods in measuring the differences between sets of random single columns derived from the different distributions. In order to do so, we generated datasets derived from

[1] This idea was later argued by Sadegh-Zadeh in Sadegh-Zadeh (2007) although the authors obtained promising results in Torres and Nieto (2003).

random seed columns where existing methods have shown a good behavior (Pietrokovski, 1996; Schones et al., 2005). We considered 25 randomly generated seed columns. We controlled the information content (IC) to be uniformly distributed in the interval $[0.01, 2]$. For each one of the 25 seed motifs, a true dataset was generated of 50000 single columns. In order to match with the properties of real motifs (Fogel et al., 2005), each column in the true datasets was obtained by sampling its corresponding seed column from a Dirichlet distribution with a random sample size between 25 and 35 (Schones et al., 2005). A false dataset was generated in a similar manner: the process is the same as for the true datasets but we skipped the sampling from seed columns. We tested the effectiveness of the methods in the following experiment: for each of the 25 true datasets, we computed the similarity of each column and its seed column, together with the similarity of each column in the false dataset and the same seed column. We considered a success when a bigger similarity is given to a column in the true dataset. In Figure 5.3 we show the power (selectivity) of the methods when the FDR (False Discovery Rate) is set to 0.01. All the methods present good results when IC gets higher. However, fuzzy approaches perform better when the IC is low. It can be seen that PCC presents the worst behavior while ALLR provides unstable results. Among the fuzzy measures[2], Minkowsky method presents lower selectivity across the experiment while Bhattacharyya performs slightly better for lower ICs values.

Clustering motifs

In order to check the performance of the measures in recognizing related real motifs, we used the freely accessible Jaspar database for our experiments (Sandelin et al., 2004a). Jaspar contains 71 nonzinc-finger motifs divided in 11 classes according to the structural properties of transcription factors (Table 5.3). Familial Binding Profiles (FBPs) are generalized binding profiles that can be used as the representatives of their respective group of

[2]As expected, Jaccard and FPSM measures overlap in Figure 5.3 since their formulas are equivalent when dealing with one column PFMs.

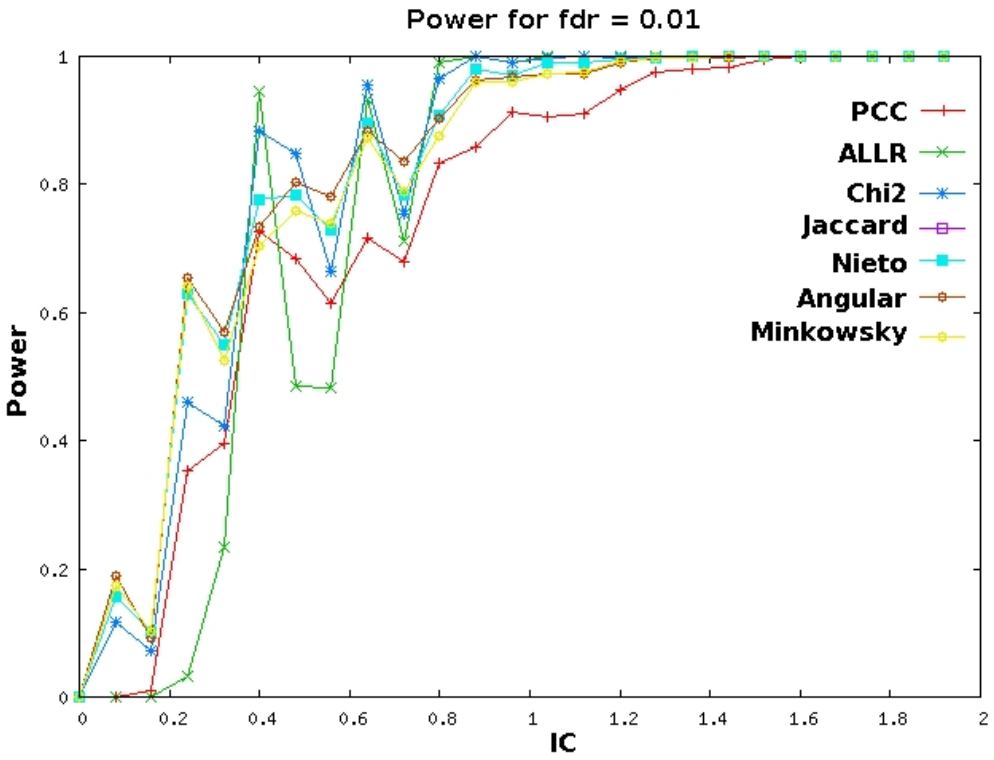

Figure 5.3: ***Recognition of random PFMs columns.*** *Power of the methods to recognize random PFMs columns generated by the same distribution.*

motifs (Sandelin and Wasserman, 2004). In our proposed methodology, we compute FBPs for each Jaspar family from a multiple alignment of the motifs within each family. An example of the logos of the bZIP EBP family and its corresponding FBP can be found in Figure 5.4. Sequence logos are computed using the freely accessible weblogo application (Crooks et al., 2004). For each of the measures we tested the similarity obtained from each FBP to its corresponding family member. In the case that the measure is a distance we converted its value into a similarity by normalizing the distance D to $[0,1]$ and later computing the similarity S as $S = 1 - D$. Table 5.4 shows the average similarity obtained from each measure for each Jaspar family. It can be easily seen that Angular, χ^2, ALLR and PCC measures provide the worst results. It is noteworthy that Angular and ALLR measures provide unstable results, yielding high similarities for some families and very poor similarities

Table 5.3: ***Jaspar family distribution.***

Family	No of motifs	Family	No of motifs
ETS	7	TRP	5
FORKHEAD	8	HMG	6
BHLH	10	HOMEO	8
bZIP EBP	4	NUCLEAR	8
MADS	5	bZIP CREB	4
REL	6		

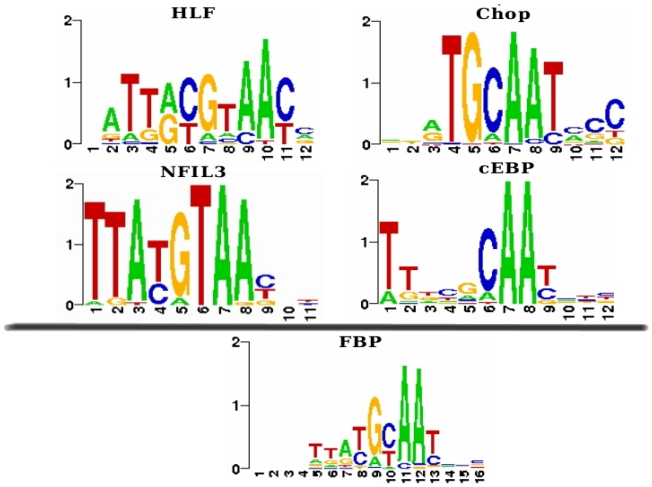

Figure 5.4: ***Logos for the bZIP EBP family and its corresponding FBP.***

for others. On the other hand, Jaccard and FPSM measures provide similar robust results in all the families. However, the Minkowsky metric performs in a better way. Using a paired t-test, the hypothesis that the distances obtained with the Minkowsky metric are not greater than those of the Jaccard or FPSM measures can be rejected with a p-value $p \leq 0.01$.

Case study. Skeletal muscle-specific

Wasserman and Fickett (1998) curated a set of PFMs for skeletal muscle-specific TFBSs. Their research was focused on locating binding sites for tran-

Table 5.4: ***Computed scores for Jaspar families and their FBPs.***

Family	Jac.	FPSM	Mink.	Ang.	χ^2	ALLR	PCC
ETS	0.63	0.62	0.71	0.75	0.19	0.71	0.24
Forkhead	0.46	0.45	0.55	0.02	0.01	0.05	0.1
bHLH	0.57	0.54	0.64	0.24	0.02	0.43	0.05
EBP	0.64	0.63	0.72	0.25	0.11	0.62	0.07
MADS	0.62	0.61	0.69	0.04	0.01	0.70	0.16
REL	0.70	0.68	0.76	0.27	0.19	0.77	0.05
TRP	0.50	0.49	0.58	0.26	0.03	0.17	0.01
HMG	0.55	0.54	0.65	0.90	0.05	0.50	0.02
HOMEO	0.47	0.47	0.59	0.73	0.02	0.37	0.07
Nuclear	0.45	0.44	0.53	0.91	0.01	0.07	0.28
CREB	0.70	0.69	0.77	0.09	0.26	0.92	0.14
Mean	0.57	0.55	**0.65**	0.41	0.08	0.48	0.10

scription factors associated with skeletal muscle-specific expression. They developed PFMs for the binding sites of Mef-2, Myf, Sp-1, SRF, and Tef transcription factors. They ensured maximum specificity for the PFMs, selecting only those sites for which there was clear and direct evidence both for function and for the identity of the factor bound were selected. In addition, for each transcription factor, they compiled its corresponding PFM using data obtained from *in vitro* binding studies and regulatory sequences from genes not specifically expressed in muscle cells. Likewise, this set is the independent counterpart of the muscle-specific one. Although gene regulation is a very complex process, the study of the transcription factor binding preferences in the two sets can shed light on deciphering the associated regulatory machinery. Figure 5.5 shows the logos for the two sets of PFMs.

It is well-known that *in vitro* binding of transcription factors to a given DNA sequence is not always confirmed by *in vivo* experiments and vice versa. Given the two curated sets of PFMs, we describe these differences in quantifiable terms, comparing the analogous PFMs from the muscle-specific and in-

Table 5.5: ***Measures for muscle-specific VS independent PFMs.***

Motif	Mink.	Ang.	Jac.	FPSM
Tef	0.24	0.94	0.26	0.32
Mef-2	0.29	0.87	0.31	0.36
Myf	0.44	0.78	0.45	0.50
Sp1	0.45	0.73	0.45	0.54
SRF	0.78	0.53	0.70	0.76

dependent sets using the fuzzy measures introduced above. Table 5.5 shows the scores obtained from the four fuzzy measures proposed. Once again, the results suggest that the Minkowsky measure is the robustest and stablest one. It can be seen that for all the measures the order of proximity of the different PFMs is the same[3]: Tef is the closest one, followed by Mef-2, Myf and Sp1, finally SRF is considered as the most dissimilar. Reader should note that, although there is not a true similarity order for these motifs that can be used as a reference, the Minkowsky measure provides lower/higher distances for the considered more/less similar motifs than the rest of the measures. It is out of the scope of this work to investigate the biological implications of the obtained similarities. However, the adequacy of this order can be confirmed by observing the logos in Figure 5.5. Roughly speaking, it can be seen that the lower the differences between high conserved positions (crucial in computing the similarity between motifs), the lower the obtained fuzzy distances are.

For the rest of the measures this order is not always obtained. For example, the closest motifs for the ALLR measure are the Mef-2 PFMs. Also, the distances tend to be bigger than for the fuzzy approaches, confirming the insights given in previous sections.

[3]Reader should note that the Angular measure is a similarity and therefore, the bigger the similarity, the closer the two PFMs

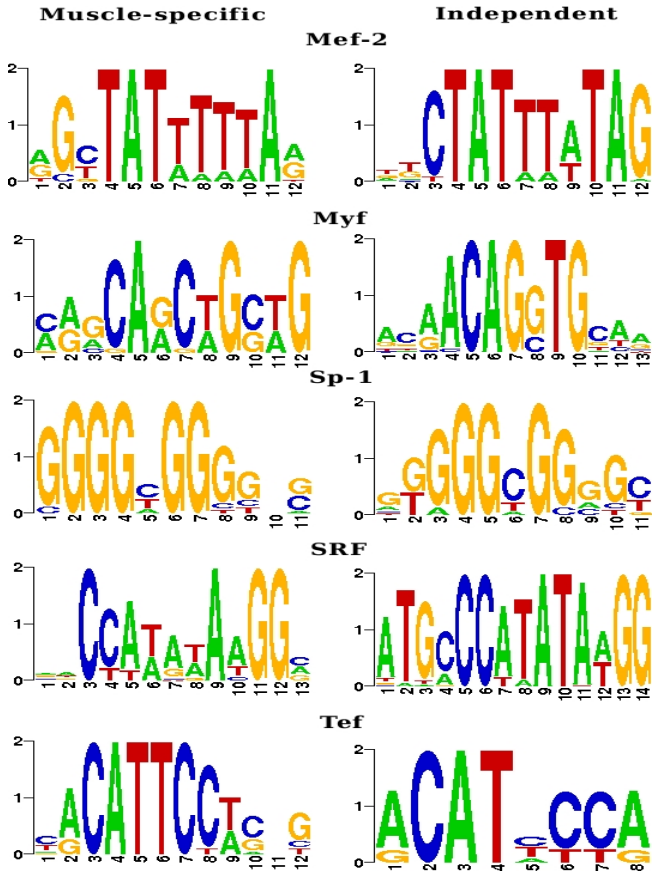

Figure 5.5: ***Logos for the muscle-specific and independent motifs studied in Wasserman and Fickett (1998).***

5.6 Concluding Remarks

TFBSs are known as regulatory motifs and can be represented as position frequency matrices (PFMs). The de novo identification of transcription factor binding sites (TFBSs) is a crucial problem in computational biology and includes the issue of comparing putative TFBSs to one another and to already known TFBSs. To date there is no fuzzy approach for this problem. In this chapter we propose the use of fuzzy measures to deal with motif comparison tasks. We investigate the behavior of different classes of classical measures for fuzzy sets including set-theoretic (Jaccard's method), proximity-based (Minkowsky's r-metric), angular coefficient based (Bhattacharyya's distance)

and a measure defined for the fuzzy polynucleotide space. We show that fuzzy measures provide excellent results when dealing with sets of randomly generated motifs, outperforming other existing measures when facing datasets of real motifs. This chapter shows the adequacy of fuzzy technology within motif comparison issues.

CHAPTER 6

Advances in Motif Measures

Although the methods proposed in Chapter 5 have been shown to work well, there is still room for improvement. Recently, new approaches have appeared aiming to improve the performance of existing measures. In this chapter we propose a new similarity measure for DNA motifs called FISim (Fuzzy Integral Similarity). FISim is based on the fuzzy integral of the distance of the nucleotides with respect to the information content of the positions. Unlike existing methods, FISim is designed to consider the higher contribution of better conserved positions to the binding affinity. Furthermore, we propose a new cluster methodology based on kernel theory together with FISim to obtain groups of related motifs potentially bound by the same TFs. We compare the performance of our proposed approaches with the measures introduced in Chapter 5 and with two newly developed methods that have been shown to be very promising.

Section 6.1 motivates the chapter. Section 6.2 gives a brief introduction of the two newly developed methods. Section 6.3 proposes our novel similarity measure FISim. Section 6.4 describes our proposed cluster methodology. Section 6.5 presents the experiments while Section 6.6 analyzes the results. The conclusion for the chapter is given in Section 6.7.

6.1 Motivation

There are several properties that are desirable for a motif similarity measure:

• Greater importance should be given to the similarity of high information content positions of the motifs than to the similarity of low information content positions.

• Methods should be designed to deal with the inherent uncertainty associated with motif comparison tasks.

• The use of parameters should be minimized.

Existing methods fail to follow one or more of these considerations. In general their approaches are not designed to deal with imprecise scenarios. In addition, these methods are not designed to consider the higher contribution of better conserved positions to the binding affinity. Some methods intrinsically tend to give greater importance to better conserved positions (e.g. ED). However, this can be improved. There is therefore a need for similarity measures for motifs that deal with these kinds of problems. In this section we present FISim (Fuzzy Integral Similarity), a novel similarity measure for comparing two motifs with one another based on the fuzzy integral with respect to a fuzzy measure.

One of the most popular tools for information aggregation is the weighted average method. It is simple, intuitive and easy to implement. This method assumes that the different information sources are non-interactive/independent and, hence, their weighted effects are viewed as additive. Due to some inherent interaction/inter-dependencies among diverse information sources, the weighted average method does not work well in many real problems. In our case, the affinity of a TF to a specific TFBS is typically correlated with how well the site matches the consensus sequence of the corresponding motif. However not all mismatches at a given position have the same effect and some interactions between positions have been observed (D'haeseleer, 2006). Here, we propose the use of the fuzzy integral to formally incorporate the different degrees of importance of the positions according to their information content level. Fuzzy integrals are a type of non-linear function dependent on fuzzy measures, and have been shown to be very useful for multiple information source fusion (Sugeno, 1977; Popescu et al., 2006). The combination of multiple information sources is very valuable with regard to overcoming the inherent ambiguities present in single information

sources. Fuzzy integrals are capable of representing the interaction among the information sources (e.g. motif columns) and of combining them to make the result more significant than just the sum of the individual comparisons, enabling the individual importance of each source to be considered in the final result (e.g. information content level). More information about fuzzy integrals can be found in Section 3.6.

6.2 New approaches

In the last years, some methods have appeared that aim to overcome the difficulties outlined above. Before we present our proposed approach, in this section we give an introduction to two newly-developed methods that have quickly become very popular.

Gupta et al. (2007) proposed an algorithm (Tomtom) that allows any column-to-column measure. They obtained best results when using euclidean distance. More recently, Pape et al. (2008) introduced the concept of a *natural* measure between motifs. They proposed that two motifs should be considered to be similar if they yield a high number of overlapping hits on a random sequence. They considered the number of hits as a random variable and described a method based on covariance to measure the correlation between the random variables of two PFMs.

Later on this chapter, we will compare the results obtained by these methods on different experiments with those obtained by our proposed measure.

Tomtom

Gupta et al. (2007) developed an algorithm (Tomtom) that admits any column-to-column measure to compute the p-values of the match scores for the columns of the query motif aligned with a given target motif. Best results are obtained when using euclidean distance (Choi et al., 2004). ED is defined as:

$$ED = -\sqrt{\sum_{b \in B} (b_{C_1} - b_{C_2})^2}. \tag{6.1}$$

In the Tomtom algorithm, a null distribution is approximated in order to obtain a p-value for the sum of the distances for all positions in the motif. The probability of observing a minimum p-value of p^* among a collection of N independent p-values is $1-(1-p)^N$. This value is the motif p-value.

Natural measure

Pape et al. (2008) defined their measure under what they called the *natural* assumption that two motifs should be considered as similar if they yield a high number of overlapping hits on a random sequence, and the number of hits is correlated between both motifs using the asymptotic covariance. Let A and B the motifs to be compared. They compute the score distributions s_A and s_B for the fixed thresholds t_A and t_B. Let $Q^k_{n_A+k}(s_A, s_B)$ be the probability to observe score s_A starting at position j and score s_B starting at position $j+k$ (see Pape et al. (2008) for more details). The overlap probability is:

$$\gamma_{A,B}(k) = \sum_{s_A \geq t_A} \sum_{s_B \geq t_B} Q^k_{n_A+k}(s_A, s_B). \tag{6.2}$$

6.3 Fuzzy Integral Similarity for Motifs (FISim)

Definition

Using PFMs for the representation of the motifs, we propose a novel column-to-column motif similarity measure called FISim (Fuzzy Integral Similarity). FISim is based on the Sugeno fuzzy integral of the distances of the nucleotide frequencies with respect to the level of conservation of the positions. In our case, the binding preferences of each position (column) are taken as the fuzzy membership degrees to sets of the four DNA nucleotides (A, C, G, T). The reader should note that uniform background distribution is assumed. When dealing with a biased background, PFMs should be modified as stated in D'haeseleer (2006).

Let $C_1 = (A_{C_1}, C_{C_1}, G_{C_1}, T_{C_1})$ and $C_2 = (A_{C_2}, C_{C_2}, G_{C_2}, T_{C_2})$ be the two columns to be compared. Let $X = \{(A_{C_1}, A_{C_2}), (C_{C_1}, C_{C_2}), (G_{C_1}, G_{C_2}), (T_{C_1}, T_{C_2})\}$ be the

set of information sources. To simplify the notation we label the pairs with a single letter so that $X = \{A, C, G, T\}$.

As it was stated in Section 3.6, Sugeno fuzzy integrals need of a function to be integrated (the so-called h function). h can be defined as:

$$h(i) = 1 - |i_{C_1} - i_{C_2}|, \tag{6.3}$$

where $i = \{A, C, G, T\}$, i.e. the similarity of the nucleotide i in the two columns C_1 and C_2.

In addition, a fuzzy measure is needed to determine the relative importance of the subset of elements being considered. Taking advantage of the properties of the fuzzy measures explained in Section 3.5, we can define a λ-fuzzy measure μ, constructed from the fuzzy densities of the individual elements μ^i. In our case, μ^i is defined as:

$$\mu^i = \max(i_{C_1}, i_{C_2}), \tag{6.4}$$

where $i \in \{A, C, G, T\}$, i.e. the maximum level of conservation of the two nucleotides, which favors the importance of better conserved positions.

At this point, we can just apply Equation 3.19 to obtain λ, and Equation 3.18 to finally obtain the fuzzy measure μ. It can be easily proven that μ fulfils properties 1 and 2 of the fuzzy measures. Once we have h and μ, it is a straightforward task to obtain the fuzzy integral applying equation 3.20.

Similarity between two PFMs comprising multiple columns needs to be constructed from the aggregation of the column-wise similarities. We proceed by averaging the similarities of the columns considering the best of all possible alignments between the PFMs as well as their reversed complementary sequences. This technique has been shown to work well in previous approaches (Schones et al., 2005; Roepcke et al., 2005). The algorithm pseudocode can be found in Figure 6.1. The source code can obtained from http://genome.ugr.es/fisim. Next, we provide an example of the computation.

FUZZY INTEGRAL SIM(C_1, C_2)

```
1  bases ← {A, C, G, T}
2  for i ∈ bases
3       do h(i) ← 1 − |i_{C_1} − i_{C_2}|
4          μ^i ← max(i_{C_1}, i_{C_2})
5  bases ← Sort bases with respect to h
6  μ ← {μ^{h(bases(1))}}
7  curSet ← {bases(1)}
8  λ ← SOLVE equation (3.19)
9  for i ← 2 to 4
10      do curSet ← curSet + bases(i)
11         μ ← μ + μ(curSet), applying equation (3.18) for λ
12 sim ← min(h(bases(1)), μ(1))
13 for i ← 2 to 4
14      do curSim ← min(h(bases(i)), μ(i))
15         if curSim > sim
16           then sim ← curSim
17 return sim
```

Figure 6.1: ***FISim pseudocode.***

FISim example

Let $C_1 = (0, 0.9, 0.1, 0), C_2 = (0.1, 0.05, 0.05, 0.8)$ the columns from the PFMs. $FISim(C_1, C_2)$ is obtained as follows: First, we need to compute h. Following the formula explained above $h(i) = 1 - |i_{C_1} - i_{C_i}|$. Therefore:

$$\begin{aligned} h(A) &= 1 - |0 - 0.1| &= 0.9 \\ h(C) &= 1 - |0.9 - 0.05| &= 0.15 \\ h(G) &= 1 - |0.1 - 0.05| &= 0.95 \\ h(T) &= 1 - |0 - 0.8| &= 0.2 \end{aligned}$$

Next, h is arranged in a decreasing order: $\{G, A, T, C\}$. From here, the sets $A_i = \{x_1, \ldots, x_i\}$ can be obtained:

$$
\begin{aligned}
A_1 &= \{G\} \\
A_2 &= \{G,A\} \\
A_3 &= \{G,A,T\} \\
A_4 &= \{G,A,T,C\}
\end{aligned}
$$

For the second part of the fuzzy integrals, a fuzzy measure μ, is needed. Since we have defined a λ-fuzzy measure, we can obtain μ from the individual importances $\mu(\{x_i\}) = \mu^i$. As we explained above $\mu^i = \max(i_{C_1}, i_{C_2})$. Hence:

$$
\begin{aligned}
\mu^A = &\quad \max(0,0.1) &&= 0.1 \\
\mu^C = &\quad \max(0.9,0.05) &&= 0.9 \\
\mu^G = &\quad \max(0.1,0.05) &&= 0.1 \\
\mu^T = &\quad \max(0,0.8) &&= 0.8
\end{aligned}
$$

Next, we need to obtain the value for the parameter λ. This can be done by solving Equation 3.19, for example by applying Newton's method. In our case $\lambda = -0.979$.

Now, it is easy to compute $\mu(A_i)$ by applying Equation 3.18: $\mu(A_1) = \mu(\{G\}) = \mu^G = 0.1$, $\mu(A_2) = \mu(\{G,A\}) = \mu(\{G\}) + \mu(\{A\}) + \lambda\mu(\{G\})\mu(\{A\}) = 0.1 + 0.1 - 0.979 \cdot 0.1 \cdot 0.1 = 0.190$. Similarly, we obtain $\mu(A_3) = 0.841$, and $\mu(A_4) = 1$.

Now, we are ready to compute the value of the fuzzy integral by solving Equation 3.20. In our case it reduces to:

$$FISim(C_1, C_2) = \max(0.1, 0.190, 0.2, 0.15) = 0.2.$$

Table 6.1 shows a summary of the computation. The reader should note that FISim will assign a high similarity between two columns when their similar values also correspond to well-conserved nucleotides. If a well-conserved

Table 6.1: ***FISim example.*** *Summary of the computation of the fuzzy integral for the given example ($\lambda = -0.979$). In bold are the minimum between $h(i)$ and $\mu(A_i)$. The fuzzy integral value is the maximum value of such minimums, i.e.* 0.2.

i	$h(i)$	μ^i	A_i	$\mu(A_i)$
G	0.95	0.1	$\{G\}$	**0.1**
A	0.9	0.1	$\{G,A\}$	**0.190**
T	**0.2**	0.8	$\{G,A,T\}$	0.841
C	**0.15**	0.9	$\{G,A,T,C\}$	1

position in one column (say 0.9) clearly differs from its corresponding position in the other column (say 0.2), the high value for the importance between these positions (0.9) is ignored. On the contrary, the similarity (0.3) will be the value chosen to proceed with the fuzzy integral computation explained in the previous section.

The reader might ask what are the advantages of FISim over the weighted sum: $\sum_{i=1}^{n} h(i)\mu_i$. Apart from benefits such as the combination of multiple information sources discussed in previous sections, FISim captures much more effectively the concept of similarity in this context, as can be seen in the example. Computing the weighted sum results:

$$WA(C_1, C_2) = 0.9 \cdot 0.1 + 0.15 \cdot 0.9 + 0.95 \cdot 0.1 + 0.2 \cdot 0.8 = 0.48.$$

This score gives the wrong impression that C_1 and C_2 present medium similarity. On the other hand, the result provided by FISim (0.2) is much more realistic, as the similarity between C_1 and C_2 is expected to be low.

6.4 Novel Clustering Methodology for Motifs

One of the main applications of motif measures is that they can be incorporated into clustering procedures for grouping related motifs. There exist two previously proposed approaches: application of hierarchical clustering methods (Mahony et al., 2007); or adaptation of the PAM (Partition Around

Medoids) algorithm (Pape et al., 2008). Hierarchical methods present problems when dealing with noisy data. They also suffer from a lack of robustness and solutions may be dependent on the data order. Moreover, PAM implementations have the drawbacks that they can converge to local optima and cannot identify clusters that are non-linearly separated in the input space. We propose a novel clustering methodology called kcmeans (kernel c-means) based on the well-known c-means algorithm, kernel methods, and our FISim measure.

The c-means algorithm uses the distances between the objects to group them into clusters (see Section 3.3). As FISim is a similarity measure, we first need to convert the similarities into distances. If the similarity (S) is an inner product, we can compute the distance (D) between objects i and j as $D_{ij} = S_{ii} + S_{jj} - 2 * S_{ij}$.

Furthermore, if we want a similarity S to be an inner product, we have to force it into a kernel. According to the kernel theory (Section 3.7), we can obtain a kernel matrix S' preserving the positive eigenvalues and corresponding eigenvectors of S. The reader should note that this transformation implies losing some information, however it is expected to be the least significant. Similarly to that explained in Section 4.2, the clustering methodology we propose works as follows: we obtain a symmetric matrix of motifs similarities S using FISim, we eliminate negative eigenvalues to produce a kernel S', which is an inner product. Finally, we compute the distance matrix $D_{ij} = S_{ii} + S_{jj} - 2 * S_{ij}$ and then apply c-means to cluster.

6.5 Experiments

Distinguishing randomized motifs

Random motifs

We tested the performance of FISim in measuring the differences between sets of random motifs. We considered 20 randomly generated *seed* motifs of a fixed length of 6 nucleotides. Following the JASPAR motif properties, the information content was uniformly ranged from 1.5 to 10.5 (see Figure 6.2

for some statistics of JASPAR motifs). For each one of the 20 *seed* motifs, a true dataset was generated containing 10000 motifs. In order to match with the properties of real motifs (Fogel et al., 2005), each motif in the true datasets was obtained as follows:

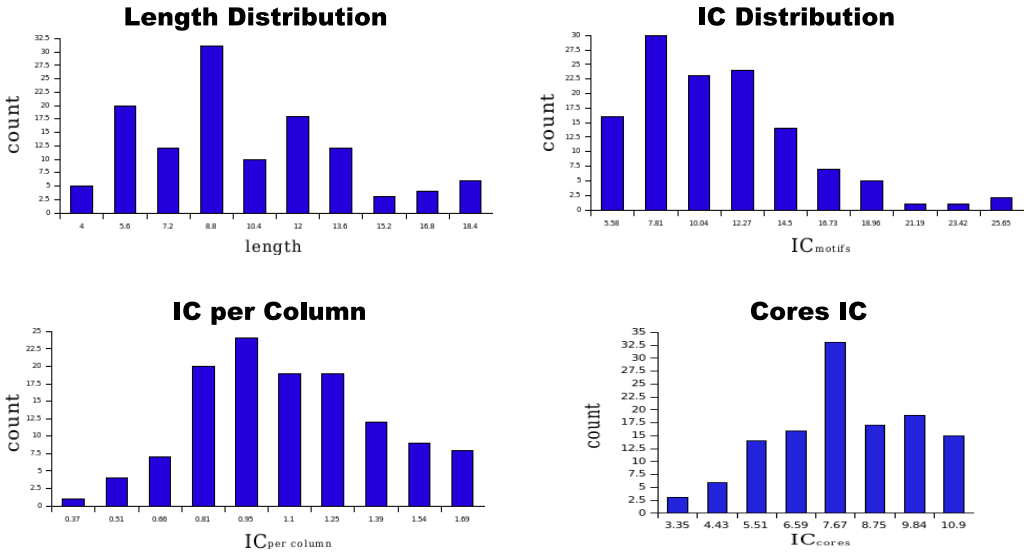

Figure 6.2: ***JASPAR statistics.***

A *random* motif of a random length between 6 and 14 was generated. The information content of this *random* motif is controlled to be low in order to create a non-conserved flanking region for the motif. The corresponding *seed* motif was sampled from a Dirichlet distribution with a random sample size between 25 and 35 (Schones et al., 2005), which generated a *sample* motif of length 6. Finally, starting in a random position, the columns in the *random* motif are replaced by the *sample* motif.

Similarly, a false dataset was generated. The process is the same as for the true datasets but we omitted the insertion of samples from *seed* motifs and the information content is not controlled. Figure 6.3 shows the power (selectivity) of the methods in recognizing motifs generated from the *seed* motifs when the FDR is 0.01. FISim shows a very good performance in a random dataset.

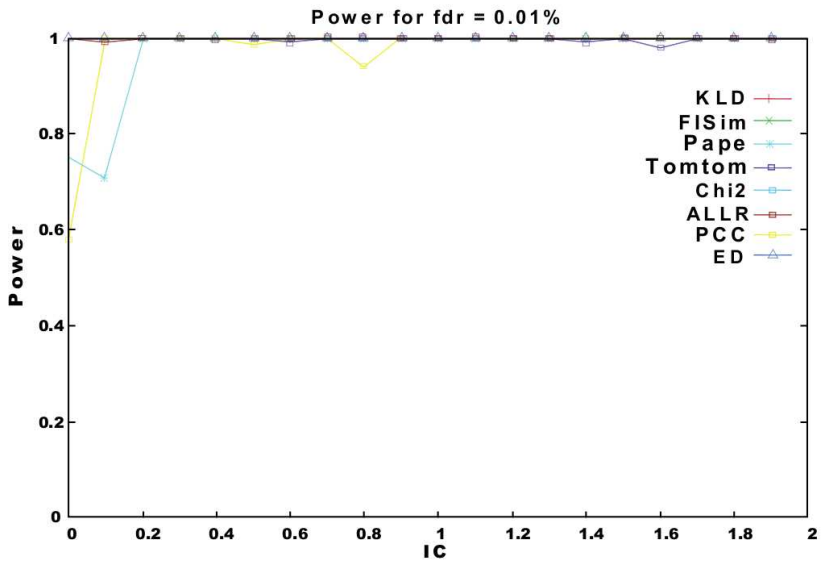

Figure 6.3: ***Random Motifs.*** *Power of the methods to recognize random PFMs generated by the same distribution.*

Distinguishing conserved and non-conserved motifs

Case study

We wanted to demonstrate the ability of the measures in discriminating the importance of non-conserved positions and well-conserved positions. In Figure 6.4 we show three motifs. We used the middle one as a *reference*. It has well-conserved positions in the odd locations (permutations of the column vector $[10, 2, 2, 2]$), and non-conserved positions in the even locations (from column vector $[4, 4, 4, 4]$). This *reference* motif was compared with the other two motifs to check how each measure performs:

• Motif *A* is composed of non-conserved columns. It therefore matches perfectly with the even positions of the *reference* motif. However, the similarity between odd positions (well-conserved) is expected to be low.

• Motif *B* is made up of two kind of columns: *a)* well-conserved positions in the odd locations that match perfectly with the corresponding positions of the *reference* motif, and *b)* medium-conserved positions (derived from permutations of the vector $[7, 7, 1, 1]$) in the odd locations that differ from the odd positions of the *reference* motif.

Note that both motifs *A* and *B* perfectly match half of the positions of the *reference* motif, while they differ in the other half of the positions. These differences are controlled for balance, in the sense that the *raw distance* of the different positions is the same, e.g. *raw distance* between $[10, 2, 2, 2]$ and $[4, 4, 4, 4]$ (*reference* motif and motif *A* differences) equals to the *raw distance* between $[4, 4, 4, 4]$ and $[7, 7, 1, 1]$ (*reference* motif and motif *B* differences). We call *raw distance* to the sum of the absolute value of the four differences between the counts of the nucleotides of the two columns.

We then considered two cases for each of the measures: *case 1*: distance between motif A and the *reference* motif, and *case 2*: distance between motif B and the *reference* motif. As has been explained above, it would be desirable that the distance for *case 2* be lower than the distance *case 1*, as, unlike motif *A*, motif *B* and the *reference* motif share the similarities in the most conserved positions of the motifs. In Figure 6.4 we show the ratio of the distances for *case 1* against *case 2*. Results for the measures proposed by Gupta et al. (2007) and Pape et al. (2008) are not shown since they require a background dataset to function correctly. Three of the measures (χ^2, KLD and ALLR) failed to capture the expected differences, and provided a lower distance for *case 1*. On the other hand, our measure obtained a more realistic distance between the motifs, providing a much lower distance for *case 2*.

Related motifs

We extended the last experiment to check the performance of the methods in datasets of related motifs. We generated a *reference* motif of length 8 comprising four well-conserved positions and four non-conserved positions used as a *reference* (see previous section for more details). We then obtained a pair of *seed* motifs comprising one *close* motif and one *distant* motif with respect to the *reference* one. Each of these motifs present three positions dissimilar to the *reference* motif. The *close* motif present the dissimilarities in the non-conserved positions, while the *distant* motif present the dissimilarities in the conserved positions (Figure 6.5). We generated a true dataset for the *close* motif and a true dataset for the *distant* motif following the procedure

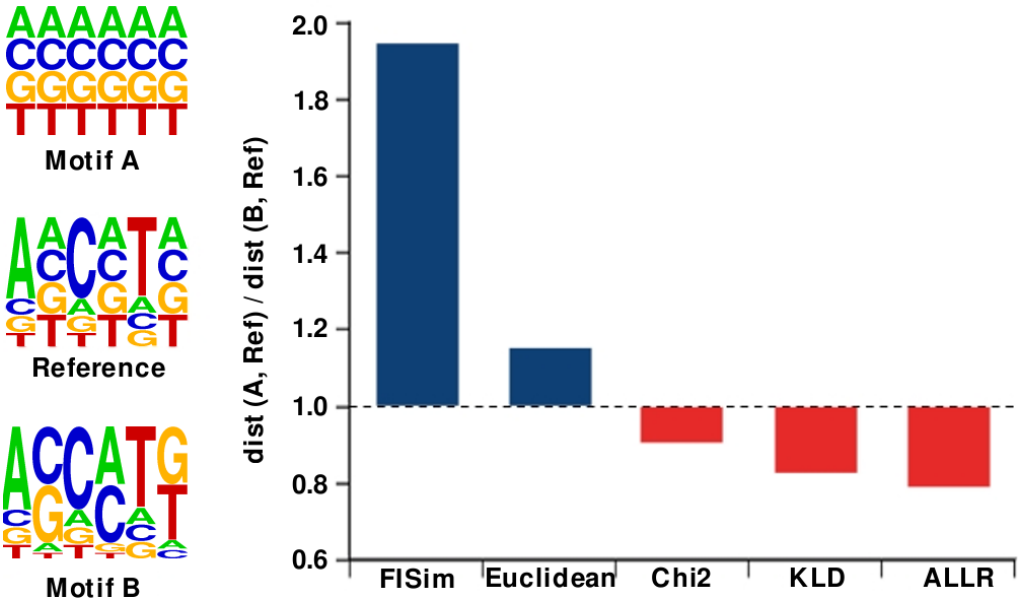

Figure 6.4: ***Case study***. *Ratio of distances. In order to facilitate the visual comparison of the non-conserved positions, fraction-based logos are used. We do not show results for the measures proposed by Pape et al. (2008) nor Gupta et al. (2007) since they need a background dataset to work properly.*

of above experiments. For each motif in the datasets we computed its distance to the *reference motif*. We determined a correct classification when a smaller distance is assigned to the *close* motif, and determined an incorrect classification otherwise. We arranged the motifs according to their distances, and from this arranged set of motifs we computed an ROC (Receiver Operating Characteristic) curve (Hanley and McNeil, 1982). ROC curves plot the percentage of correct classifications as a function of incorrect classifications. In Figure 6.6 we show the ROC curves obtained from the different approaches. It can be seen that our FISim method proposed outperforms the other methods. Similar results are obtained when varying the number of dissimilar positions of the seed motifs. The area under the curves (AUC) scores and the logos for the motifs can be found in Table 6.2 and Figure 6.7 respectively.

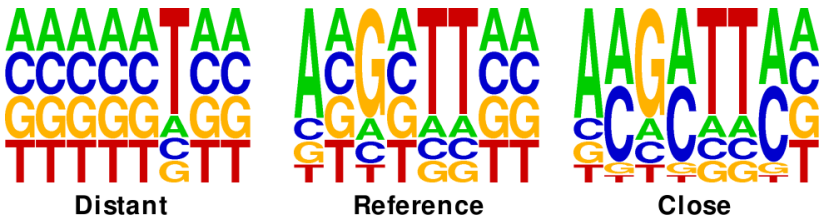

Figure 6.5: ***Related motifs.*** *Three dissimilar positions are observed between the reference motif and both close and distant motif. Again, fraction-based logos are used to ease the visual comparison of the non-conserved positions.*

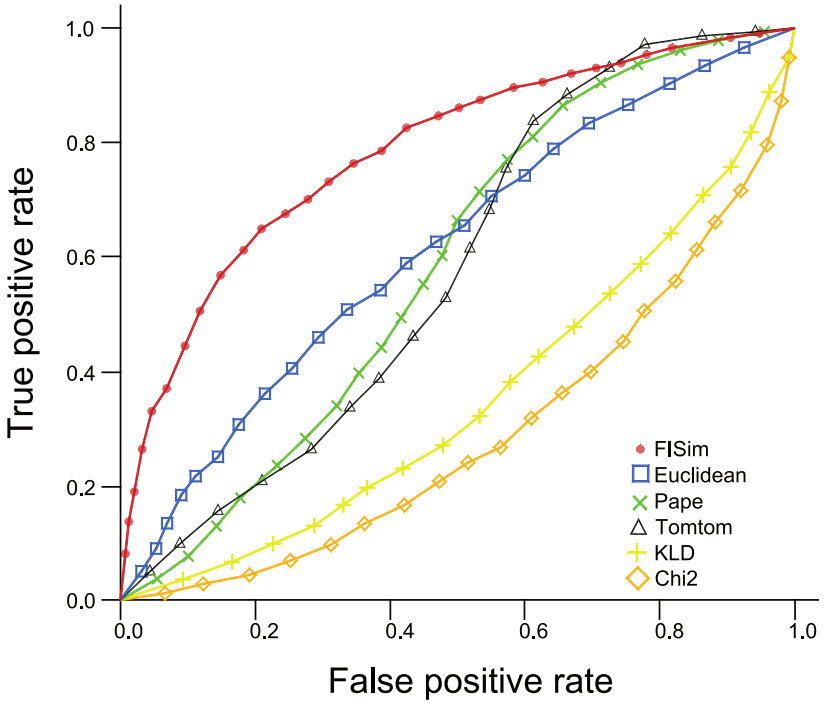

Figure 6.6: ***ROC curves.*** *ROC curves for the case of three different columns. FISim provides a more consistent classification than the rest of the methods.*

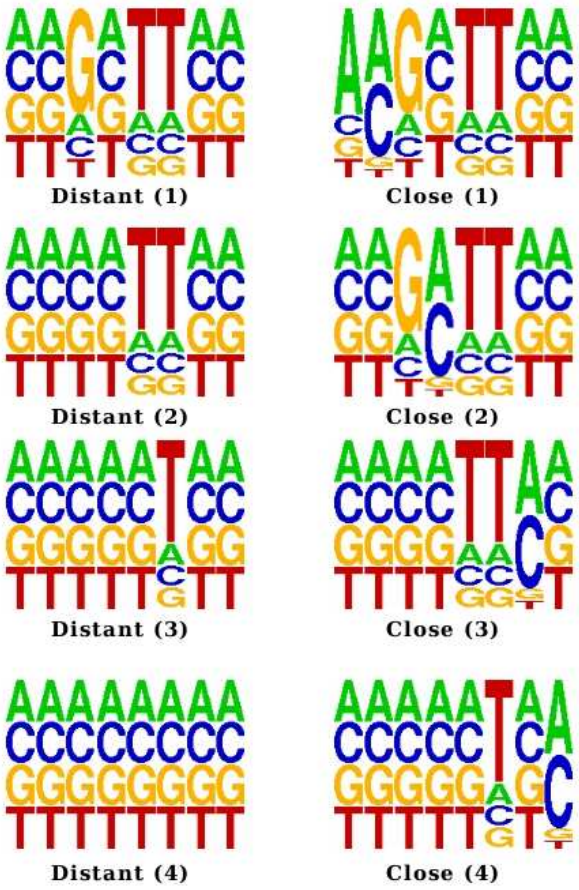

Figure 6.7: ***Pairs of seeds of related motifs.*** *The number of the columns not shared with the reference motifs is indicated between brackets.*

Table 6.2: ***Area Under the Curve scores for the related motifs experiment.*** *Note that the more differences between the motifs the more difficult the discrimination task is.*

Differences	FISim	KLD	Chi2	ALLR	Tomtom	Euclidean	Pape
1	0.785	0.328	0.242	0.232	0.894	0.657	0.701
2	0.785	0.350	0.282	0.315	0.795	0.633	0.430
3	0.783	0.353	0.291	0.276	0.697	0.615	0.590
4	0.682	0.202	0.157	0.213	0.573	0.404	0.632

Clustering real data

In order to check the performance of kcmeans in separating related motifs, we used the freely accessible JASPAR (Sandelin et al., 2004a) database for our experiments. JASPAR contains 71 nonzinc-finger motifs divided into 11 classes according to the structural properties of the transcription factors. The distribution of the families of the JASPAR motifs can be found in Table 6.3. For each motif we computed the core region, following the suggestions of Schones et al. (2005). In order to obtain a symmetric matrix, comparisons between two motifs were made by averaging the similarity between the core region of the first motif and the second motif, and the similarity between the first motif and the core region of the second motif. Once we obtained the similarity matrix, we applied the *kcmeans* clustering method as described in the Methods section. For each cluster, the FBP is automatically obtained from a multiple alignment of its corresponding motifs.

Table 6.3: ***JASPAR family distribution.*** *Summary of the JASPAR classification. There exist 71 motifs divided into 11 families.*

Family	Number of motifs	Family	Number of motifs
ETS	7	TRP	5
FORKHEAD	8	HMG	6
bHLH	10	HOMEO	8
bZIP EBP	4	NUCLEAR	8
MADS	5	bZIP CREB	4
REL	6		

To obtain the optimal number of clusters (k) we used the Silhouette coefficient (Kaufman and Rousseew, 1990). The optimal clustering of the 11 motifs classes was found for $k = 15$. The 15 clusters can be found in Figure 6.8. Next, we discuss the clustering.

All five TRPs motifs are contained in two homogeneous clusters with two and three motifs. Five of ten BHLHs motifs form one homogeneous cluster. Four of the remaining five BHLHs are grouped together with one ETS motif. All six remaining ETSs motifs form one cluster together with three of

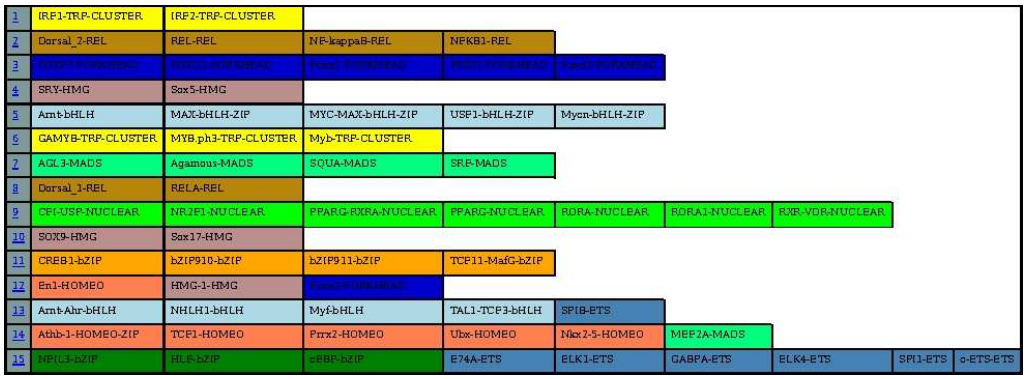

Figure 6.8: ***Clusters obtained by kcmeans for the Jaspar motifs.*** *The motifs that share the same background color belong to the same Jaspar family.*

four bZIP EBPs motifs. Four of five MADSs motifs are grouped in one homogeneous cluster. The remaining MADS motif is contained in one cluster together with five of six HOMEOs motifs. There exists one heterogeneous cluster formed by the remaining HOMEO motif, one HMG motif and one FORKHEAD motif. Four of the remaining five HMGs motifs are grouped in two homogeneous clusters of two motifs each. Five of the remaining six FORKHEAD motifs form one homogeneous cluster. Seven of eight NUCLEAR motifs form one homogeneous cluster. Finally, all five bZIP CREBs motifs form one homogeneous cluster, and all six RELs motifs form two homogeneous clusters with four and two motifs. Figure 6.8 shows the clustering result. Logos for the different clusters can be found in the next section. As a summary, ten out of 15 of the obtained clusters are homogeneous, while eight motifs are not clustered and are considered as outliers.

To ensure the quality of the clustering, we compared our results with those provided by Pape et al. (2008).

Two identical clusters are obtained: NUCLEAR and bZIP CREB. The same MADS and HOMEO groups are provided but we yielded a MADS motif, *MEF2A*, within the HOMEO group. MADSs motifs present the consensus CCA*A, while HOMEO motifs present the consensus ATTA. *MEF2A* motif contains the consensus ATT showing that the FISim measure certainly gives greater importance to better conserved positions (for sequence logos see Appendix A). We presented the REL family in two clusters, while in Pape et al.

(2008), this appears together in the same cluster. We obtained the same two TRPs clusters, but added one extra TRP motif (*MYB.ph3*) to one cluster which Pape et al. (2008) considered as an outlier. The *MYB.ph3* motif shares the consensus AAC*G with the motifs in its cluster. The same bZIP cEBP group is provided, although we added six out of the seven ETSs motifs. Here, the common high degree of conservation of the consensus TTCC forces them to belong to the same cluster. We yielded the same two bHLH clusters, but added one bHLH motif (*Arnt-Ahr*), considered as an outlier in Pape et al. (2008), as well as the remaining ETS motif to one of the clusters. Pape et al. (2008) presented the FORKHEAD and HMG groups in one single cluster in comparison with three homogeneous clusters obtained. Finally, the heterogeneous cluster that we produced comprises one extra FORKHEAD motif *Foxd3* that does not contain the consensus GTTTA present in the FORKHEAD group.

In short, we obtained 15 clusters (eleven homogeneous) and found eight outliers (i.e. motifs not clustered), compared to 14 clusters (ten homogeneous) and twelve outliers in Pape et al. (2008). Hence, we found more motifs in the final clustering, reducing the number of non-classified motifs, and maintaining a homogeneous structure. Figure 6.5 shows the sequence logos of one REL group as well as its corresponding FBP.

Motif identification in co-regulated genes

As discussed in previous sections, one of the most common applications of a motif similarity measure is its use for comparing putative motifs of co-regulated genes obtained from motif discovery algorithms to those reported in motif databases such as JASPAR or TRANSFAC. In this section we present the results of applying FISim to this workflow with the data studied in Sørlie et al. (2001).

The aim of this study was to classify breast carcinomas based on their gene expression profiling derived from 85 microarray experiments and to correlate tumor characteristics to clinical outcome. The authors classified the tumor samples into two main branches, each of these separated into three subgroups. For this experiment, we selected the “Luminal Subtype A” subgroup, which contains 15 clones (13 genes) clearly involved in pathological

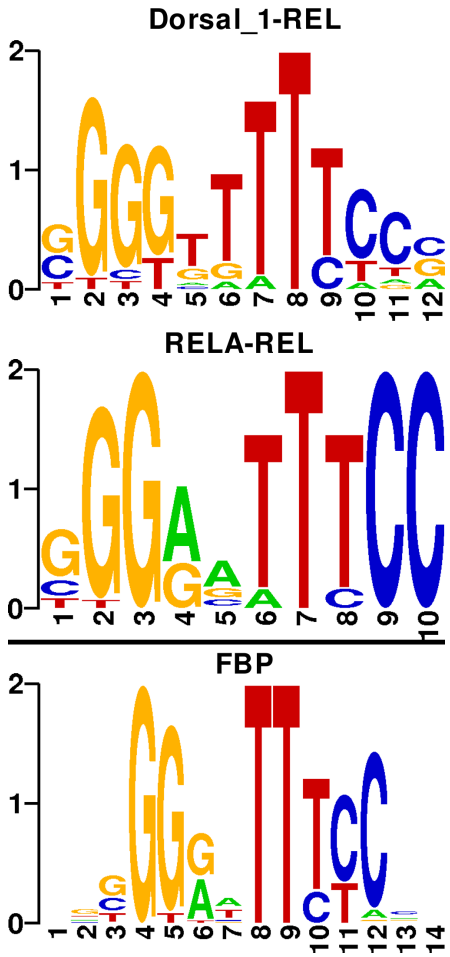

Figure 6.9: ***REL group retrieved by kcmeans.*** *The FBP is computed from the multiple alignment of the TFs Dorsal_1 and RELA.*

processes of breast cancer. This cluster includes genes implicated in transcription, development and differentiation such as *ESR1*, *GATA3*, *LIV1*, and *XBP1* (see Table 6.4 for a whole list of genes).

We applied the motif discovery tool WebMOTIFS (Romer et al., 2007) to further investigate regulation of the predicted cluster of genes. We used WebMOTIFS to find putative motifs in the promoter regions of these 15 clones, setting the options to default, i.e. selecting AlignACE, MDscan, MEME and Weeder methods (Hughes et al., 2000; Liu et al., 2002; Bailey and Elkan, 1994; Pavesi et al., 2006) and no Bayesian information. For each method,

Table 6.4: ***Co-regulated genes.*** *Genes that belong to the cluster termed "Luminal epithelial gene cluster containing ER" in Sørlie et al. (2001).*

Gene	GenBank Acc.	Description
GPR160	H50224	G protein-coupled receptor 160
ACADSB	H95792	Acyl-Coenzyme A dehydrogenase
ESR1	AA291749	Estrogen receptor 1
TFF3	N74131	Trefoil factor 3 (intestinal)
GATA3	R31441	GATA binding protein 3
XBP1	W90128	X-box binding protein 1
FOXA1	T74639	Forkhead box A1
AFF3	H99588	AF4/FMR2 family, member 3
LIV1	H29315	Estrogen-regulated protein LIV-1
NPNT	AA029948	Nephronectin
TUBA1C	N54508	Tubulin, alpha 1c
NAT1	R91802	N-acetyltransferase 1
MYO6	AA625890	Myosin VI
MYO6	AA030004	Myosin VI

Table 6.5: ***Best JASPAR matches for the MDSCAN algorithm.***

Motif	Best JASPAR match	FISim value
MDSCAN-1	CREB1	0.745
MDSCAN-2	ABI4	0.700
MDSCAN-3	ESR1	0.731
MDSCAN-4	TP53	0.732
MDSCAN-5	Pax4	0.704
MDSCAN-6	NFKB1	0.743
MDSCAN-7	TFAP2A	0.701

we selected the most significant motifs and compared these to the publicly available JASPAR motifs using FISim. Some of the most similar motifs found in JASPAR include ESR1, CREB1, TAL1-TCF3, TP53, NFKB1 and PAX5. For a complete list of motifs, as well as their similarities with JASPAR motifs, see Tables 6.5-6.7. Logos for those motifs are available in Appendix B.

As expected, the link between these motifs is the estrogen receptor alpha (*ESR1*) gene. Estrogens play an important role in both female and male

reproductive function, as well as in female cancers, and they have multiple effects on the nervous, skeletal, and cardiovascular systems. *ESR1* is over-expressed in the “Luminal Subtype A" subgroup together with, among others, the *GATA-3*, *LIV-1* and *XBP1* genes. Previous studies described how these genes are coordinately expressed with *ESR1* in breast cancers (Wilson and Giguere, 2008; Gomez et al., 2007). A wide variety of non-DNA binding molecules, called coactivators, have been identified that are able to enhance ligand-induced activity of steroid receptors, including *ESR1*, through direct or indirect binding to these receptors (Dutertre and Smith, 2003). Among them, *CREB*-binding protein is critical for ligand-induced, nuclear receptor-mediated transcription activation (Torchia et al., 1997). In addition, there is evidence that estrogen and progesterone together with *TGF-β* signaling are necessary for maintenance of p53 activity in the mammary epithelium (Becker et al., 2005), and for an *ESR*-mediated inhibition of the *NFKB* signaling pathway. *NFKB* target genes are significantly elevated in *ESR*-negative versus *ESR*-positive breast tumors, which indicates a potential crosstalk between *NFKB* and *ESR* (Van Laere et al., 2006).

6.6 Analysis of the Results

We have introduced a new measure of similarity for regulatory motifs called FISim. The uncertainty associated with motif comparison tasks makes fuzzy concepts particularly suitable for handling this kind of data. FISim is based on the fuzzy integral and takes advantage of the fuzzy concepts to overcome

Table 6.6: ***Best JASPAR matches for the MEME algorithm.***

Motif	Best JASPAR match	FISim value
MEME-1	MNB1A	0.790
MEME-2	SPIB	0.793
MEME-3	SPI1	0.808
MEME-4	ZNF42	0.775
MEME-5	GATA2	0.745

Table 6.7: ***Best JASPAR matches for the Weeder algorithm.***

Motif	Best JASPAR match	FISim value
Weeder-1	Pax5	0.727
Weeder-2	TAL1-TCF3	0.744
Weeder-3	ESR1	0.682
Weeder-4	TCF11	0.789
Weeder-5	CREB1	0.701

some of the known difficulties that arise in measuring motifs tasks. There are three main differences from other approaches: *i)* it considers not only the distance between the PFMs columns, but also the relative importance of each occurrence within each column, *ii)* it enables the inherent uncertainty of the PFMs to be handled, and *iii)* it does not make use of any user-provided parameter.

A simple experiment shows how other measures fail in capturing realistic differences, while FISim provides good results (Figure 6.4). These results are confirmed on extending the experiment to long datasets (Figure 6.6). Furthermore, it is noteworthy how the naive euclidean distance (Choi et al., 2004) inherently appears to assign greater importance to better conserved positions (see Figure 6.4). This might explain why Gupta et al. (2007) and Mahony et al. (2007) found the best performance of their methods when using the euclidean distance to compare the motifs.

As explained above, FISim is based on the fuzzy integral theory. Fuzzy integrals have been proven to be very suitable for information fusion. The combination of the evidence supplied by the information sources (nucleotide frequencies) and the importance of each subset of information sources (nucleotide conservation level) is very interesting in motif recognition tasks. When dealing with long random datasets, we show that FISim provides excellent results in terms of motif recognition, similar to those obtained applying existing methods. This was expected, since the probability of overlapping within random motifs is low, which facilitates the discrimination of the origins of the motifs. Some methods perform poorly when the information

contents are low (e.g. ALLR and PCC), however, FISim also provides good results under these circumstances.

This task gets more complicated when motifs are interrelated. In this case, it is noteworthy that the Tomtom algorithm provides very good results for higher information content values. However, FISim provides better results, especially when the information content of the motifs is lower, i.e. when it is more difficult to recognize the motifs. This makes FISim particularly interesting when dealing with real problems. For example, as motif discovery algorithms become more and more powerful, motifs with lower information content will be produced as putative motifs and these will need to be tested.

Another advantage of our method is that it does not require any additional parameter. This makes FISim a more robust and fully automated method, thus avoiding the need to select parameters via expert knowledge or trial-and-error approaches.

We used FISim to investigate the motifs found by popular motif discovery algorithms in a well-known set of co-regulated genes corresponding to the subgroup "Luminal Subtype A" of breast carcinomas. Comparison of the obtained motifs with those reported in JASPAR suggested that the *ESR1* gene plays a crucial role in this kind pathology. Furthermore, *ESR1* interacts with other motifs also present among the most significant motifs obtained. These findings confirm previous studies and show the reliability of FISim in real-life problems.

Our proposed cluster methodology (kcmeans) makes use of FISim and the kernel theory to avoid problems found when applying other classical methods (i.e. definition of a medoid, data order dependence, etc.). The study of the performance of kcmeans in real data shows promising results in terms of accuracy and cluster compactness. Comparison of our results with those from similar experiments shows a better global behavior and a more accurate grouping of the motifs.

6.7 Concluding Remarks

In this chapter we introduce FISim, a new similarity measure for motifs and a novel clustering methodology, based on the fuzzy integral and on kernel technology respectively. Our main objectives were to favor the influence of the better conserved positions of the motifs and to exploit the tolerance for imprecision and uncertainty of fuzzy technology. FISim corrects a design flaw of the most popular methods, whose measures favor similarity of low information content positions. Our measure takes into account the relative importance of each nucleotide within a given position. We use our measure to successfully identify motifs that describe binding sites for the same TF and to solve real-life problems. We show that FISim outperforms other approaches in motif recognition tasks, and prove how it can be successfully applied to day-to-day research problems. In this chapter the reliability of fuzzy technology for motif comparison tasks is proven.

CHAPTER 7

Scoring DNA Sequences against TFBS Motifs

Pattern discovery in DNA sequences is one of the most important problems in bioinformatics with applications in finding regulatory elements and transcription factor binding sites. An important task in this problem is to search (or predict) known binding sites in a new DNA sequence. The most common approach for this matter, is to score all the subsequences of the given DNA sequence by means of an scoring function. Most of the available tools for transcription factor binding site prediction are based on methods which assume no sequence dependence between the binding site base positions. New approaches aim to consider position dependencies. In this chapter, our primary objective is to propose a novel scoring method based on the intuitionistic fuzzy set theory that outperforms existing methods.

Section 7.1 introduces the problem. Section 7.2 reviews the existing approaches. Our new scoring method is proposed in Section 7.3. In Section 7.4 we compare the methods with different experiments. Section 7.5 provides an analysis of the results of the experiments performed in the previous section. In Section 7.6 we use our proposed scoring method for the study of single nucleotide polymorphisms. Finally, Section 7.7 concludes the chapter.

7.1 Introduction

Cells control the abundance of proteins by means of diverse mechanisms. One such mechanism is the regulation of transcription, which is a continuous process whereby many factors combine to ensure appropriate rates of protein synthesis. Understanding such complex processes is one of the main objectives in computational biology. In its early stages, transcription is controlled by the binding of proteins called transcription factors (TFs) to specific regions of a given chromosome called transcription factor binding sites (TFBSs). These interactions between proteins and DNA usually take place upstream from the gene, close to the transcription start site (TSS), in the so-called promoter region of the gene.

One of the biggest issues in identifying TFBSs is that a single binding protein can bind to different DNA sequences. Related DNA sequences to which the same TF can bind are grouped together into a TFBS motif. The identification of TFBSs within a given set of DNA sequences is an active area of research. In this context there exist two main approaches: i) the *de novo* discovery of motifs, and ii) the detection of TFBSs using motifs that are already known.

De novo methods aim to find significant sub-sequence patterns within a set of TFBS sequences. Some of the most popular approaches are MEME (Bailey and Elkan, 1994), Gibbs sampling (Lawrence et al., 1993) and AlignACE (Hughes et al., 2000). For a review of these methods see Das and Dai (2007).

Detection methods, on the other hand, focus on inferring new TFBSs from known binding motifs. Early detection methods assumed independence between positions within a putative TFBS sequence, e.g. in Patser (Hertz et al., 1990) and ConSite (Sandelin et al., 2004b). However, it is now well established that this assumption is wrong (Benos et al., 2002; Bulyk et al., 2002; Eisen, 2005), and two recent methods have been developed that take into account interdependencies between TFBS positions. Tomovic and Oakeley (2007) proposed a method that incorporates a measure of positional interdependence into the overall score. More recently, Zare-Mirakabad et al. (2009)

developed a method based on joint information content and mutual information. In this method, positional dependencies are taken into account by considering all pairwise combinations of positions (See Section 7.2 for more information).

The fact that TFBS sequences are usually very short means that the same or very similar sequences tend to occur by chance at a relatively high frequency. Consequently one of the main goals in the prediction of TFBSs is to reduce the false positive rate without compromising sensitivity. Methods that take into account positional dependencies tend to be significantly more effective at meeting this challenge. However, there remains room for improvement. As we will show in Section 7.4, existing methods have some drawbacks, such as overlearning of the training data, arbitrary threshold selection for testing dependencies, etc. The purpose of the work presented here is to provide a new method for measuring sequence-motif affinity that improves on existing approaches in precisely these areas.

Zadeh (1965) proposed fuzzy set theory to mathematically model the imprecision inherent in certain concepts. Briefly, fuzzy set theory allows an object to partially belong to a set with a membership degree between 0 and 1. Classical set theory is a special case of its fuzzy counterpart in which membership and certainty degrees are restricted to either 0 or 1. Attanasov (1986) proposed intuitionistic fuzzy set (IFS) theory as an extension of the fuzzy set theory. IFSs generalize the notion of a fuzzy set representing uncertainty with respect to both the degree of membership (μ) and non-membership (ν) of a set by allowing that the sum $\mu + \nu \leq 1$.

Owing to the fact that IFSs are capable of modelling the uncertainty present in real-life situations, they have been widely applied during the past decades to a variety of problems (see Section 7.3). In recent years, it has been seen that the inherent uncertainty and noise that characterize biological data cannot always be modeled sufficiently well using probabilistic approaches and that, as a consequence, alternative approaches to modeling this uncertainty may be required (Garcia et al., 2009; Lopez et al., 2008; Liang et al., 2008). As it was previously discussed, in addition to the usual problems of missing values and noisy data associated with biological data,

there exist some additional hidden factors that affect binding affinities in the context of sequence-motif scoring, e.g. cooperative binding and chromatin structure (Lam et al., 2008). Furthermore, the described motifs are subject to change as new experiments confirm new binding sites. In this work we make use of IFS theory to formally model the uncertainty associated with the problem of scoring DNA sequences against TFBS motifs.

7.2 Alternative approaches

In recent years, several scoring methods for the prediction of TFBSs have been proposed. In this section we give a brief overview of those methods that take account of positional dependencies, as they have been shown to outperform methods that assume independence. Let us first introduce the notation. Let $B = \{A, C, G, T\}$ be the set of the four DNA nucleotides. Let $D = d_1, \ldots, d_n$ be an ordered DNA sequence on B of length n. Let us suppose that we have a motif $M = S_1, \ldots, S_t$, where $S_i \in D$ consists of t aligned binding sites of length n . The problem is then reduced to assigning a score to the pair formed by a given putative TFBS, $S \in D$, and a given motif, M.

In what follows we will follow the notation of Wasserman and Sandelin (2004) where $F(b,i)$, for $b \in B$ and $1 \leq i \leq n$ shows the occurrences of nucleotide b in position i, and $P(b,i) = \frac{F(b,i)}{t} + a(b)$, for $b \in B$ and $1 \leq i \leq n$ is the corrected probability of base b at position i, where $a(b)$ is a smoothing parameter ($a(b) = 0.001$)[1].

Statistical dependencies.

Tomovic and Oakeley (2007) extended the previous method that assumed positional independence. The authors also followed the notation of Wasserman and Sandelin (2004) and defined $W_{b,i}$ as a position weighted matrix (PWM) of base b in position i computed as:

$$W_{b,i} = log_2 \frac{P(b,i)}{P(b)}. \tag{7.1}$$

[1] $a(b) = 0.01$ is usually reported but our experiments show that smaller values provide more accurate results.

In the case where independence is assumed, the score for a given DNA sequence S can be computed by summing all the values of $W_{b,i}$ for every base in S:

$$SC_{indep}(S) = \sum_{i=1}^{n} W_{S_i,i}. \tag{7.2}$$

The first step for extending this score involves testing the dependencies between each pair of positions i and j. The authors introduced three different methods:

- χ^2 test:

$$\chi^2 = \sum_{b_i,b_j} \frac{(P(b_i,b_j,i,j) - P(b_i,i)P(b_j,k))^2}{P(b_i,i)P(b_j,k)}, \tag{7.3}$$

  where $P(b_i,b_j,i,j) = \frac{F(b_i,b_j,i,j)}{t}$. Here $F(b_i,b_j,i,j)$ is the frequency of base pairs $b_i b_j$ at positions i and j.

- G statistics:

$$G = 2\sum_{b_i,b_j} P(b_i,b_j,i,j) \ln\left(\frac{(P(b_i,b_j,i,j)}{P(b_i,i)P(b_j,k)}\right). \tag{7.4}$$

- Bayesian hypothesis testing of the hypotheses i) H_0: both distributions $P(b_i,b_j,i,j)$ and $P(b_i,i)P(b_j,k)$ are the same; and ii) H_1: otherwise. The authors show that the Bayes Factor $BF(H_0,H_1)$ can be approximated as:

$$\log_2(BF(H_0,H_1)) \approx -tM_{ij}, \tag{7.5}$$

  where M_{ij} is the mutual information between positions i and j given by:

$$M_{ij} = \sum_{b_i,b_j} P(b_i,b_j,i,j) \log_2\left(\frac{(P(b_i,b_j,i,j)}{P(b_i,i)P(b_j,k)}\right). \tag{7.6}$$

The authors used these three methods to calculate the dependencies between pairs of positions in the motifs available in the public database JASPAR

(Sandelin et al., 2004a). The reader should note that the accurate computation of positional dependencies is still an open problem since different results are obtained depending on the method and parameters used in their computation (see Supplementary Material 2-4 in Tomovic and Oakeley (2007)). Further details about obtaining the position dependencies and multiple test corrections can be found in Tomovic and Oakeley (2007).

In order to compute the new score, the corrected probability for the bases $b_1 b_2 \dots b_m$ in the dependent positions $i_1 i_2 \dots i_m$ is defined by:

$$P(b_1,\dots,b_m,i_1,\dots,i_m) = \frac{F(b_1,\dots,b_m,i_1,\dots,i_m)}{t} + \\ + a(b_1,\dots,b_m), \tag{7.7}$$

where $a(b_1,\dots,b_m) = a(b_1)\dots a(b_m)$ is a smoothing parameter.

It is straightforward then to obtain values that correspond to the PWM values:

$$W_{b_1,\dots,b_m,i_1,\dots,i_m} = \log_2\left(\frac{P(b_1,\dots,b_m,i_1,\dots,i_m)}{P(b_1)\dots P(b_m)}\right). \tag{7.8}$$

Finally, their proposed scoring function, which incorporates positional dependencies, can be computed as:

$$SC_{dep}(S) = \sum_{i=1}^{k_1} W_{S_i,i} + \sum_{i=1}^{k_2} W_{S_{j_i},S_{j_{i+1}},j_i,j_{i+1}} + \dots + \\ + \sum_{i=1}^{k_m} W_{S_{j_i},\dots,S_{j_{i+m-1}},j_i,\dots,j_{i+m-1}}, \tag{7.9}$$

where, k_1 is the number of independent positions, k_2 is the number of dependent positions of order 2 (nucleotides at positions j_i and j_{i+1}) and k_m the number of dependent positions of order m (in other words, nucleotides at positions $j_i, j_{i+1}, \dots, j_{i+m-1}$).

For both the SC_{indep} and SC_{dep} it is advisable to perform the following normalization:

$$N_{SC} = \frac{SC - \min(SC)}{\max(SC) - \min(SC)}. \tag{7.10}$$

Matrix based.

Zare-Mirakabad et al. (2009) proposed a new scoring function based on the dependency between all pairwise combinations of binding site positions. Their method is based on the mutual information matrix (see equation (7.6)) and on the joint information content (JIC), defined as:

$$JIC = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{b_1 \in B} \sum_{b_2 \in B} P(b_1, b_2, i, j) \log \left(\frac{P(b_1, b_2, i, j)}{p(b_1) p(b_2)} \right). \tag{7.11}$$

In order to compute their score, the authors defined a PWM, W^{PW}, containing 16 rows and $(n\Delta(n-1)/2)$ columns for all the pairwise combinations of the positions:

$$W^{PW}_{b_1,b_2,i,j} = \log \left(\frac{P(b_1, b_2, i, j)}{p(b_1) p(b_2)} \right) + \log \left(\frac{P(b_1, b_2, i, j)}{p(b_1, i) p(b_2, j)} \right), \tag{7.12}$$

where $b_1, b_2 \in B$ and $1 \leq i, j \leq n$ and $i \neq j$.

Finally, for a given DNA sequence $S \in D$ of length n the score SC_{mat} is computed as:

$$SC_{mat} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} W^{PW}_{S_i,S_j,i,j}. \tag{7.13}$$

In order to obtain a normalized value for the score, equation (7.10) should be applied.

7.3 New Intuitionistic Approach (SC_{intuit})

Intuitionistic fuzzy sets

Intuitionistic fuzzy set (IFS) theory was proposed by Attanasov (1986) and has been applied in such diverse fields as decision making (Szmidt and Kacprzyk, 1996), logic programming (Atanassov and Gargov, 1990) medical diagnosis (De et al., 2001b; Khatibi and Montazer, 2009), pattern recognition (Hung and Yang, 2004). IFS theory is an extension of the fuzzy sets theory previously proposed by Zadeh (1965) that allows the degrees of membership

and non-membership to be independently uncertain, which makes the representation more flexible at capturing the current state of our understanding given inconclusive data (Atanassov, 1994; De et al., 2001a). Next, we introduce some basic IFS concepts.

Let X be the universe of discourse. An intuitionistic fuzzy set A in X is an object having the form:

$$A = \{(x, \mu_A(x), \nu_A(x)) : x \in X\}, \tag{7.14}$$

where $\mu_A, \nu_A : X \to [0,1]$ denote membership function and non-membership function of A, satisfying $0 \leq \mu_A + \nu_A \leq 1$ for every $x \in X$. Therefore, the degree of uncertainty of x to A is $\pi_A(x) = 1 - \mu_A - \nu_A$. For more on this topic please refer to Attanasov (1986); Atanassov (1994); De et al. (2001a).

Intuitionistic representation of motifs

For our approach, a given motif M is represented as the set of IFSs of all the pairwise combinations of its positions: $I^M = \{I^M_{i,j}\}$, where $1 \leq i, j \leq n$ and $i \neq j$. Each of the i, j combinations for the motif positions is then an IFS of 16 elements defined as:

$$I^M_{i,j} = \{b, \mu_{I^M_{i,j}}(b), \nu_{I^M_{i,j}}(b) : b \in B \times B\}, \tag{7.15}$$

where $B \times B$ is the universe of discourse, i.e. the set of all 16 possible combinations of bases for two given positions i and j (AA, AC, ..., TT).

Membership degree computation

$\mu_{I^M_{i,j}}$ represents the degree of membership of the pairs for the basis $b_1, b_2 \in B$ in a given pair of positions i, j in a motif M. It can be automatically computed as:

$$\mu_{I^M_{i,j}}(b_1, b_2) = P(b_1, b_2, i, j) + (1 - P(b_1, b_2, i, j))\frac{p(b_1, i) + p(b_2, j)}{2}, \tag{7.16}$$

where the above notation holds. Obviously, $0 \leq \mu_{I^M_{i,j}}(b_1, b_2) \leq 1$ and the degree increases as do the corrected probabilities of bases b_1 and b_2 in positions i and j, as well as the individual corrected probabilities $p(b_1, i)$ and $p(b_2, j)$.

Non-membership degree computation.

$\nu_{I^M_{i,j}}$ represents the non-membership degree of the pairs for the basis $b_1, b_2 \in B$ in a given pair of positions i and j in a motif M. It can be automatically computed as:

$$\nu_{I^M_{i,j}}(b_1, b_2) = \left(\frac{IC_i^{b_1} + IC_j^{b_2}}{2} \right) (1 - \mu_{I^M_{i,j}}(b_1, b_2)), \tag{7.17}$$

where $IC_p^b = \frac{p(b,p)\log_2(p(b,p))}{2}$ is the normalized information content of base b in position p and $\nu_{I^M_{i,j}}(b_1, b_2)$ is in the range $0 \leq \nu_{I^M_{i,j}}(b_1, b_2) \leq 1$. As $\nu_{I^M_{i,j}}(b_1, b_2)$ increases, the information content of the two basis increases and the corresponding membership degree decreases. It is easy to prove that $\mu_{I^M_{i,j}}(b_1, b_2) + \nu_{I^M_{i,j}}(b_1, b_2) \leq 1$.

Scoring

In order to define our proposed score, we first introduce the simplest case of scoring a length-2 DNA subsequence $D = b_1, b_2$ in the positions i and j of a motif M:

$$SC^{i,j}_{intuit}(b_1, b_2) = \mu_{I^M_{i,j}}(b_1, b_2)(\max(\nu_{I^M_{i,j}}) - \nu_{I^M_{i,j}}(b_1, b_2)), \tag{7.18}$$

where $\max(\nu_{I^M_{i,j}})$ is the maximum degree of non-membership in M found in the pair of positions i and j considering all the possible combination of basis $b_1, b_2 \in B^2$, and $\mu_{I^M_{i,j}}(b_1, b_2)$ and $\nu_{I^M_{i,j}}(b_1, b_2)$ are the membership degree and non-membership degree of the pairs for the basis $b_1, b_2 \in B$ in the pair of positions i, j of M, computed as stated in sections 7.3 and 7.3 respectively.

As with the previously defined scores, a normalization step needs to be performed in order to obtain comparable results

$$NSC^{i,j}_{intuit}(b_1, b_2) = \frac{SC^{i,j}_{intuit}(b_1, b_2) - \min(SC^{i,j}_{intuit}(b_1, b_2))}{\max(SC^{i,j}_{intuit}(b_1, b_2)) - \min(SC^{i,j}_{intuit}(b_1, b_2))}. \tag{7.19}$$

Finally, for a given DNA sequence $S \in D$ of length n the score SC_{intuit} is computed as:

$$SC_{intuit} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} NSC^{i,j}_{intuit}(S_i, S_j). \tag{7.20}$$

7.4 Comparative Study of the Performance of SC_{intuit}

Case study

In the majority of cases, the sequences known to belong to a given TFBS motif have very similar nucleotide compositions and highly conserved positions. However, in the databases of known motifs there are a number of examples where individual sequences differ from the majority in highly-conserved positions. Such a binding sequence can be considered an outlier with respect to the motif. When scoring new nucleotide sequences against a given TFBS motif, we should generally tolerate small, additional variations in the sequence with respect to non-outliers, but be far less tolerant of mutations to outlier sequences. Here we evaluate the extent to which each scoring method is able to discriminate between sequences belonging to these two categories.

Take, as a preliminary example, the binding sequences for motif MZF1 in the JASPAR database, as shown in Figure 7.1(A). It can be observed how the highlighted outlier sequence GGAGGA does not contain the highly-conserved base G at the third position, while the highlighted sequence TGGGGA is clearly a non-outlier (see motif logo in Figure 7.1(B)). We selected the highlighted sequences and additionally created two new sequences by mutating its sixth position giving GGAGGG and TGGGGT. We scored each pair of sequences against the motif by means of the different methods. Figure 7.1(C) shows the results.

In the case of the non-outlier binding sequence, all three methods performed similarly for the two sequences. They gave very high scores for the original sequence, indicating a high binding probability, using, for example, a standard threshold as discussed in section 7.4 below. In addition, high scores are obtained for the mutated sequence, as expected, since the sequence still shares most of the highly conserved positions with the motif. This situation changes in the case of the outlier binding sequence. As can be seen, all three methods gave a high score for the original sequence, indicating a high binding probability, whereas, the results for the mutated sequence are signif-

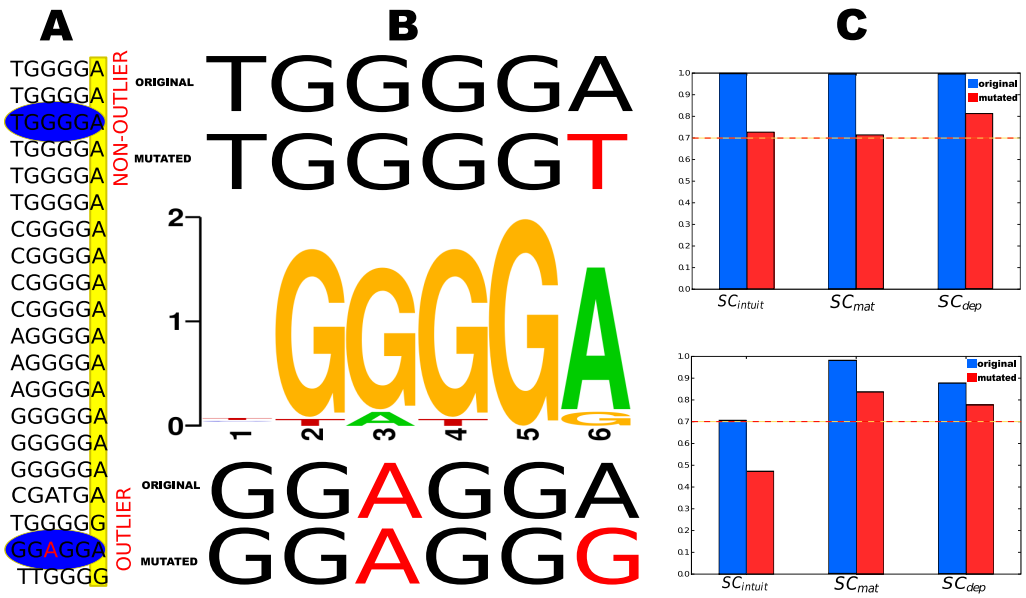

Figure 7.1: ***Motif MZF1.*** *A) Shows the binding sequences found in JASPAR. B) Shows the logo representation of the motif (center), with the original and mutated sequences for the non-outlier binding sequence (above), and the outlier binding sequence (below). C) Shows the results for each of the three methods when scoring the original and mutated sequence of the non-outlier binding sequence (top) and the outlier binding sequence (bottom).*

icantly different. In reality one would expect a low binding probability owing to the fact that the original sequence is an outlier and the mutated sequence has an additional mismatch at one of the highly conserved positions of the motif. As desired, a low score is obtained by our proposed method SC_{intuit}. However, high scores are obtained by the SC_{mat} and SC_{dep} methods, giving the incorrect impression that binding might occur.

These insights are confirmed in the following sections where the experiments are extended to use large datasets, and the results are measured in terms of discovery rates without regard to the chosen threshold.

Prediction of TFBSs

Synthetic sequences

In order to compare the performance of the different methods in predicting TFBSs, we used the publicly available JASPAR motifs database (Sandelin et al., 2004a) for our experiments. We selected all motifs for which binding sequences are available (not only matrix profiles), resulting in a dataset of 124 motifs. For each of these motifs, a binding site was randomly selected and inserted in a random sequence from a third-order Markov model background distribution obtained from the program RSA (Van Helden, 2003). For each position of each sequence we computed the score for its corresponding motif. We consider a correct classification to have occurred when a higher score is assigned to a position where a binding site was inserted, and an incorrect classification otherwise. In order to control the number of *true negatives* for the different lengths of the motifs, we fixed the length of the random sequences to 50 bases more than the length of the inserted binding site. Thus, we obtained a dataset comprising 124 *true positives* and 6200 *true negatives*.

Usually, methods have a high sensitivity (i.e. can detect true positives), so that the key difference between them is the number of false positives. Following the recommendations of Tomovic and Oakeley (2007), we selected a threshold of 0.7 indicating a match for a binding site. The thresholded results for the different methods indicate that our proposed scoring function performed best, giving the smallest number of false positives per TF whilst simultaneously giving a high number of true positives.

In order not to rely on the selection of an arbitrary threshold for evaluating the results, we arranged the motifs according to their distances, and from this arranged set of motifs we computed a ROC (Receiver Operating Characteristic) curve (Hanley and McNeil, 1982). ROC curves plot the percentage of correct classifications as a function of incorrect classifications. In Figure 7.2 we show the ROC curves obtained using the different approaches. Area under the curve (AUC) values can be observed in Table 7.4. It can be seen that our proposed method SC_{intuit} outperforms the other methods.

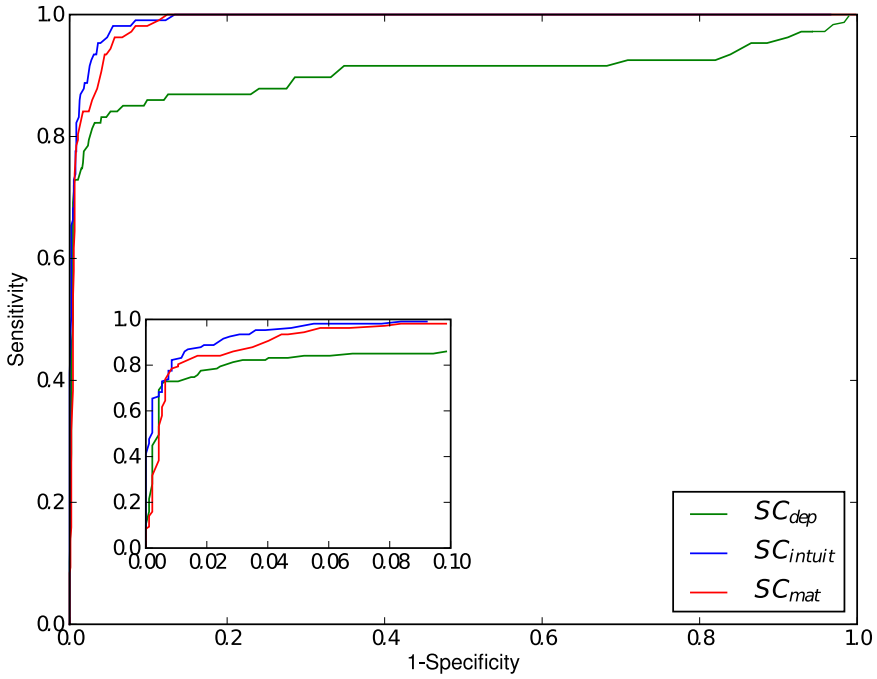

Figure 7.2: ***ROC curves of the three scoring methods for the synthetic sequences experiment.*** *SC_{intuit} provides a more consistent classification than the rest of the methods.*

Precision-recall (PR) curves are commonly used in information retrieval for evaluating classification performance and give a more informative picture of a method's performance when dealing with highly skewed datasets as is the case here (Fawcett, 2006). Figure 7.3 shows the PR graphs together with a representation of the F-measure results[2] in function of the selected threshold for the three methods. SC_{intuit} produces a better PR graph than the remaining methods (see Table 7.4 for AUC values). This is confirmed by the F-measure performance at the selected threshold. Positive values are obtained when the F-scores for SC_{mat} and SC_{dep} are subtracted from that for SC_{intuit} (1.45 and 0.89 respectively).

[2]F-measure is the weighted harmonic mean of precision and recall.

Table 7.1: *AUC values for the synthetic and mutated sequence experiments.*

	Synthetic		Mutated	
	AUC ROC	AUC PR	AUC ROC	AUC PR
SC_{dep}	0.904	0.822	0.878	0.730
SC_{mat}	0.987	0.891	0.984	0.819
SC_{intuit}	0.992	0.942	0.991	0.924

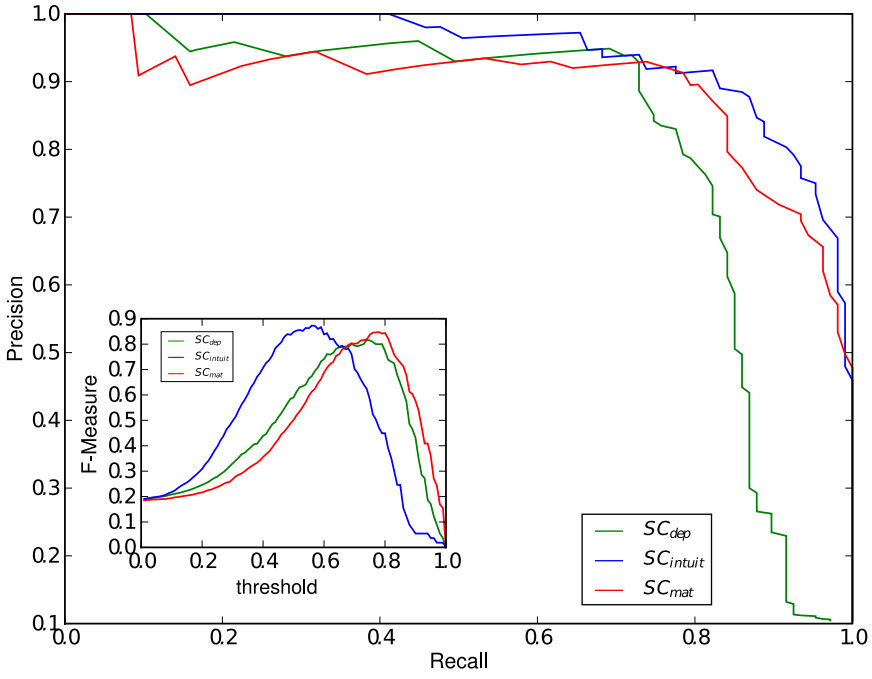

Figure 7.3: ***Precision-recall and F-measure graphs of the three scoring methods for the synthetic sequences experiment.***

Mutated sequences

To further evaluate our proposed method, we obtained a set of putative binding sites that are very similar to those that are already known. This is a common scenario in motif discovery, where the set of known sequences belonging

to a given binding motif is incomplete. In order to simulate this situation, we proceeded in a similar way to our previous experiment; all the steps were the same except that we gave a single base mutation at a random position within the selected binding site for each motif. ROC curves and AUC values were computed to compare the performance of the different methods (Figure 7.4 and Table 7.4). In addition, Figure 7.5 shows the precision-recall and F-measure graphs for the three methods. SC_{intuit} clearly obtains better results than the remaining methods. The ROC and precision-recall graphs shows how SC_{intuit} gives consistently superior values, with a higher AUC value (Table 7.4), which are confirmed by the representation of the F-measure results. Here again, positive values are obtained when the F-scores for SC_{mat} and SC_{dep} are subtracted from that for SC_{intuit} (2.16 and 4.37 respectively). It can be observed that the improvement of the performance of our method compared to SC_{mat} and SC_{dep} grew with respect to the synthetic sequences experiment discussed in the previous section. Arguably this makes it more reliable to be used in real problems.

Real Data

We analyzed the performance of the proposed methods when dealing with real experimental data. In order to do so, we made use of the published ChIP-seq data on binding of TFs in embryonic stem cells from mouse by Chen et al. (2008), as provided in the supplementary material of Sharov and Ko (2009). We considered the three TFs (SMAD1, Myc, and STAT3) that have binding sequences available in the TRANSFAC database (Matys et al., 2006). Thus, we obtained three sets of 200 bp sequence segments centered at TF binding locations, and we randomly selected 50 sequence segments from each set for our study. We scanned each set of sequences using the 124 TFs from JASPAR for which binding sequences are available. The results demonstrate the superior performance of our new scoring method, as it gives the smallest number of false positives per nucleotide and per TF, and has an excellent true-positive rate (Figure 7.6). It can be seen how our method maintains consistently low false-positive rate with all three set of sequences, whilst the performance of the other methods varies with respect to the different dataset.

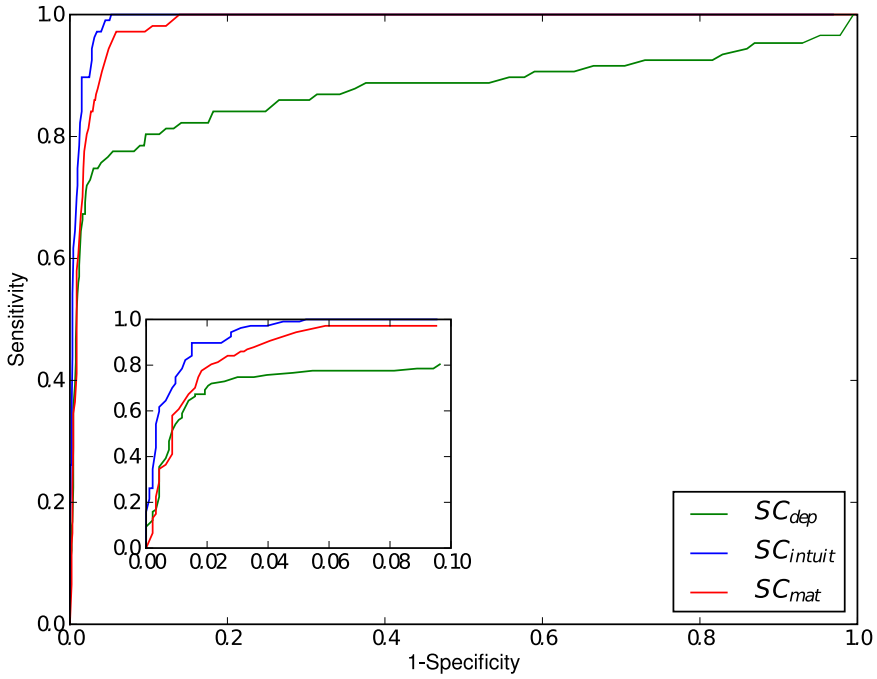

Figure 7.4: ***ROC curves of the three scoring methods for the mutated sequences experiment.***

7.5 Analysis of the Results

We have introduced a new IFS-based approach for scoring DNA sequences against DNA motifs called SC_{intuit}. In section 7.2 we presented three scoring schemes. These approaches have several drawbacks. SC_{indep} is based on an incorrect assumption that the nucleotides of a given TFBS are independent. In that context, SC_{dep} extended the score in order to account for positional dependencies. Zare-Mirakabad et al. (2009) pointed out the problems associated with unnormalized scores at each position. In addition, the results vary depending on the choice of the method and parameters for testing the dependencies (see section 7.2). The main drawback with SC_{mat} is outlined in section 7.4: it has a tendency to overlearn the training data and consequently its performance decreases when applied to real problems. There is

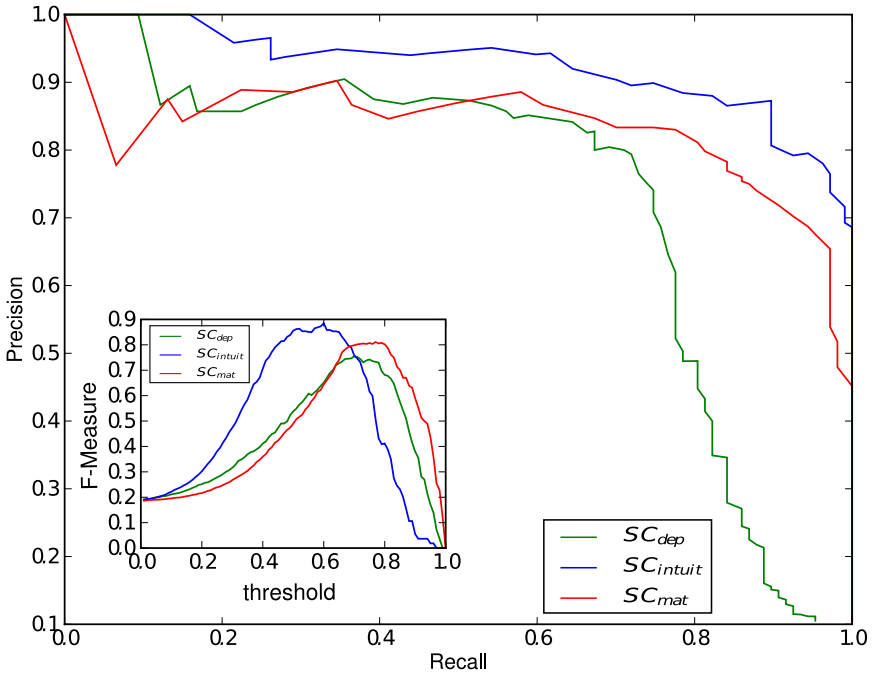

Figure 7.5: ***Precision-recall and F-measure graphs of the three scoring methods for the mutated sequences experiment.***

therefore a need for a scoring method that accounts for positional dependencies without compromising either the consistency or the accuracy of the results.

As explained above, SC_{intuit} is based on the IFS theory, which has been successfully applied to problems that suffers from noisy and imprecise data. IFS theory represents uncertainty with respect to both the degree of membership and non-membership. The uncertainty associated with the tasks of scoring DNA sequences against motifs makes intuitionistic concepts particularly suitable for handling this kind of data. Taking advantage of such properties, we define the membership and non-membership degrees of a given pair bases at a given position not only as a function of their combined probability of occurrence, but also taking into account the importance of each individual base at its corresponding position.

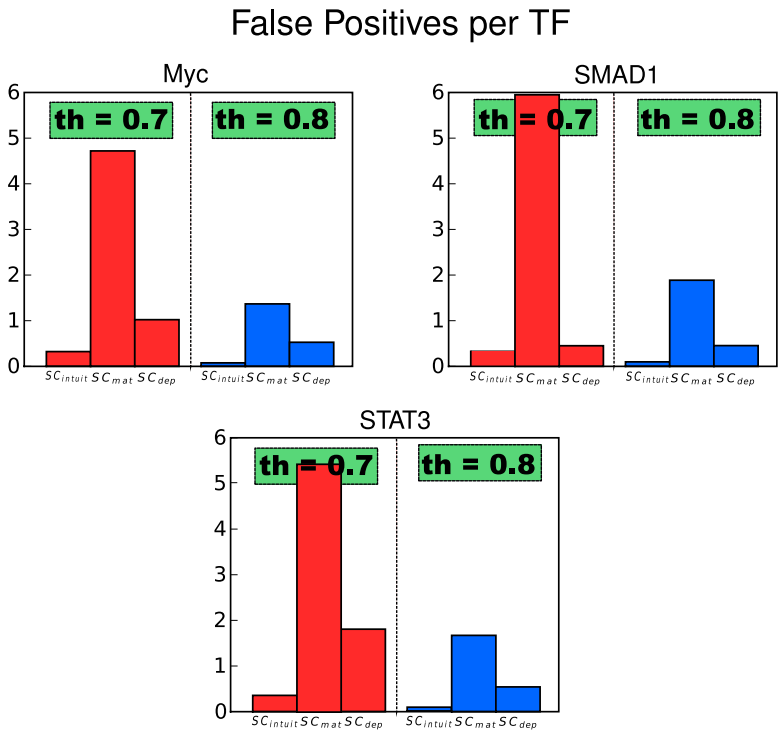

Figure 7.6: ***Average false-positive ratio per TF for different thresholds for the proposed scoring methods.***

One of the biggest issues for this kind of scoring methods is giving high scores for the known binding sequences of the motifs without overfitting. Our proposed approach adequately solves the problem of computing the score of a given sequence against a given motif by considering the binding sequences that comprise the motif not only individually but also as part of such set of sequences. A simple experiment shows how other methods fail in capturing realistic differences, while SC_{intuit} provides good results (Figure 7.1). Our method assigned high scores for known binding sites, disfavoring mutations in the conserved positions of the binding site.

These insights are confirmed from experiments for predicting TFBSs in large datasets (Section 7.4). We compared the performance of the proposed scoring methods on recognizing motifs in sets of random sequences from a

third-order Markov model background distribution in two circumstances: *i)* when inserting known binding sequences, and *ii)* when inserting mutated binding sequences. In both situations we found that our proposed method gave the smallest number of false positives per TF whilst simultaneously giving a high number of true positives (Figures 7.2-7.5). More importantly, our method outperforms the other approaches when dealing with real experimental data derived from Chip-seq assays. In this case, again, the number of false positive is significantly reduced (Figure 7.6).

In general, the obtained results on the different experiments demonstrated that the proposed intuitionistic approach provide a better and more accurate model for the detection of motifs and for the relationships between positions of the TFBSs.

7.6 Study of Single Nucleotide Polymorphisms in TNFR1 Gene for the Response against *Aspergillus Fumigatus*

Motivation

Hematological patients are typically treated by chemotherapy and/or radiation. These treatments usually produce immunosuppression and severe neutropenia. This clinical situation can be exploited by opportunistic pathogens such as *Aspergillus fumigatus* to cause a deadly infection called Invasive Pulmonary Aspergillosis (IPA) (Denning, 1998; Offner et al., 1998). *Aspergillus fumigatus* is then an important fungal pathogen that can cause IPA in immunocompromised patients with high morbidity and mortality rates. The importance of finding ways to combat this pathogen is evidenced by the fact that IPA occurs in roughly 10% to 40% of hematological patients, with overall mortality rates ranging from 50% to 90% (Chamilos et al., 2006; Diop et al., 2005).

Tumor necrosis factor (TNF) is primarily secreted by macrophages and activates T lymphocytes in response to fungal infections through TNF receptors. One of the most important TNF receptors is TNFR1, which triggers a

pro-inflammatory response and, therefore, plays a crucial role in immune regulation and host immune responses. Experimental studies with TNFR1 knockout mice indicate that TNFR1 is indispensable in host resistance against several infections (Hehlgans and Pfeffer, 2005). Our hypothesis is that single nucleotide polymorphisms (SNPs) in TNFR1 gene may influence the innate immune response against *Aspergillus fumigatus*.

Identification of patients who are more susceptible for infection could facilitate the development of effective prevention strategies. Genetic factors explain, at least in part, why some people resist infection more successfully than others. Gene disruptions can cause fatal vulnerability to specific pathogens (Kwiatkowski, 2000). Indeed, SNPs in the promoter and coding regions of cytokine genes have been described associated with a difference in the cytokine production (Kim et al., 2008b) or function (Knight, 2005) and, therefore, they might influence susceptibility to infections.

The gene encoding TNFR1 contain numerous polymorphisms (Bochud et al., 2008; Baker et al., 1991). By means of different experiments, we concluded that TNFR1$_{\text{-609(G/T)}}$ polymorphism is critical in the development of the response against *Aspergillus* because it might be regulating the cell-mediated Th1 immune response[3]. In this section, we use our proposed scoring method SC_{intuit} to know whether the TNFR1$_{\text{-609(G/T)}}$ promoter polymorphism is involved in the disruption of the recognition of a potential binding site for a critical transcription factor that could influence TNFR1 transcription level.

TNFR1$_{\text{-609(G/T)}}$ Polymorphism Binding Affinity

As commented above, there are several databases where TFBSs are available for the scientific community. For this experiment we used motifs found in TRANSFAC database (Matys et al., 2006), which has been widely used in research works involving regulatory elements (Wingender, 2008). In order to find interesting dependences between the TNFR1$_{\text{-609(G/T)}}$ SNP and

[3]Details on these experiments are out of the scope of this dissertation and can be consulted in Sainz et al. (2009).

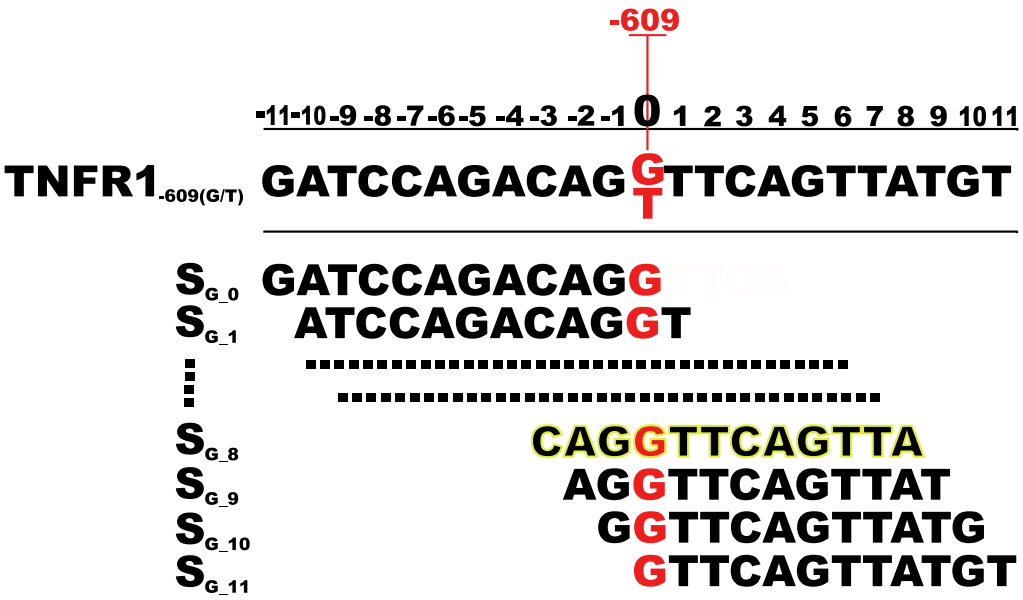

Figure 7.7: ***Set of putative binding sequences for the G allele of the TNFR1$_{-609(G/T)}$.*** *Best results are found for the highlighted S_{G_8} sequence. A corresponding set was also obtained for the T allele.*

TFs binding affinity we scored the human TRANSFAC TFBSs against the TNFR1$_{-609(G/T)}$ polymorphism by means of the SC_{intuit} method.

TFs bind to short parts of the TNFR1 promoter region and, therefore, for each trial, we need to define a fragment of the promoter sequence containing the TNFR1$_{-609(G/T)}$ SNP that might be considered as the putative TFBS. To this, we need to determine the length of the sub-sequences and the relative offset to the position of the TNFR1$_{-609(G/T)}$ SNP. For each of the 446 human TF in TRANSFAC, we generated a set of putative binding sequences by using a window size of a fixed length equals to the number of position of the corresponding TF. Moving the window across the sequence in 5'-3' direction gave us the sub-sequences for the TNFR1$_{-609(G/T)}$ SNP that we considered to be putative TFBSs (see Figure 7.7 for an example). Next, we scored each pair of sub-sequences (one sub-sequence for G allele, and for T allele) against the given TF applying the SC_{intuit} method.

We were interested in those sub-sequences that fulfil two properties: *i)* they have a high score in one allele (G or T) so they can be considered

as candidates to be binding sites, and *ii)* the score is substantially lowered when considering the remaining allele so the SNP might affect to the binding affinity. For this matter, we kept those TFBSs that presented a score over 0.7[4] After that, we compared the selected TFBSs with their corresponding alleles (sequences with a G(T) instead of a T(G) at position -609). Results are shown in Table 7.2. In the next section we discuss these finding from a biological point of view.

Table 7.2: SC_{intuit} ***scores for the two alleles.***

TF	Starting position	Direction	TNFR1$_{-609(T)}$	TNFR1$_{-609(G)}$
AREB6	603	-	0.59	0,70
E2A	606	-	0.64	0.79
HNF4	605	+	0.52	0.78
ICSBP	606	+	**0.81**	**0.69**
MYB	601	-	0.76	0.77
Pax-2	604	-	0.76	0.58
SMAD	603	+	0.73	0.73

Functional Effect of ICSBP/IRF-8 in the TNFR1-609(C/T) SNP

In the previous section, we obtained predictive results using SC_{intuit} method and TRANSFAC database (Table 7.2). From them, we selected four candidates according to the two properties outlined in the previous section, i.e E2A, HNF4, ICSBP, and Pax-2. We did not find described relations between IPA response for any of E2A, HNF4, and Pax-2 TFs. Logos for these TFs are provided in Figure 7.8.

On the other hand, we found ICSBP (also known as IRF-8) to be directly related with the purpose of our study. ICSBP/IRF-8 shows a preference for binding the T allele (see Table 7.2). As a member of IRF family of transcription factors it is an important modulator of IFNγ signalling cascade and was

[4]This arbitrary threshold is used as a conservative level. Later on this section we show how the most interesting insights are found for the ICSBP TF which presents the highest scoring among all the human TRANSFAC motifs.

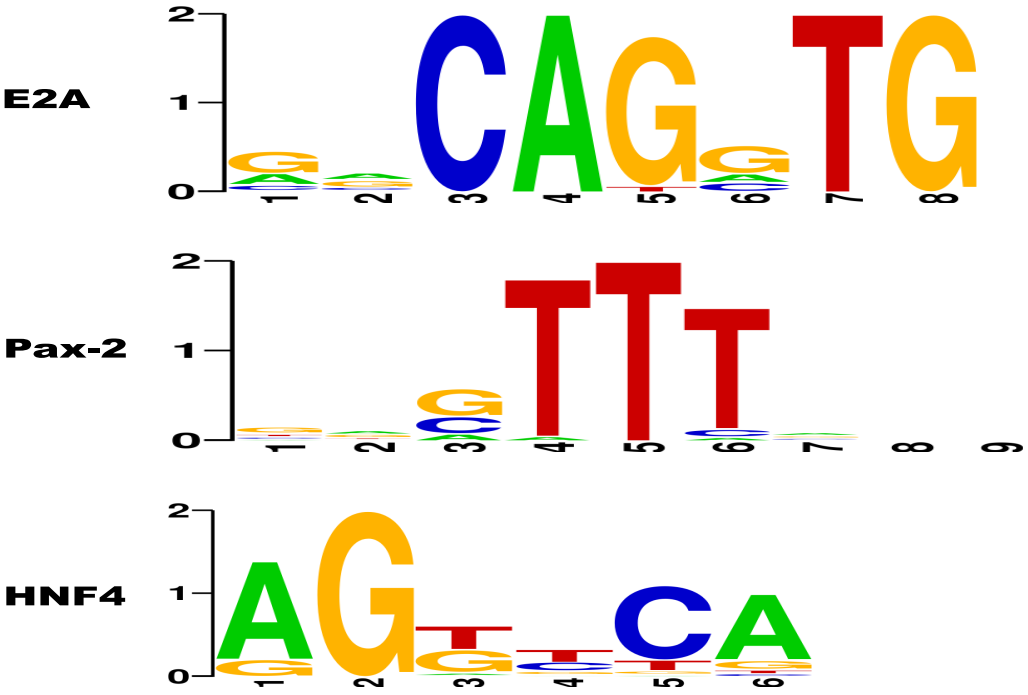

Figure 7.8: ***Discarded TFs.***

identified in association on the promoter region of numerous macrophage essential genes such as IL12, IL1β, IL18, iNOS or ISG15 (Dror et al., 2007).

In addition, several genes regulated by ICSBP/IRF-8, such as MAP4K4, IL-17R, and SOCS7, are involved in different stages of the nuclear factor κB (NFκB) signaling pathway (Dror et al., 2007). Therefore, we can hypothesize that ICSBP/IRF-8 transcription factor might be also regulating the NFκB signaling pathway through the control of the first gene of this signalling cascade, the TNFR1 gene. In support of this hypothesis, (Zhao et al., 2006) established that ICSBP/IRF-8 and TNFR1 are closely related genes. They found ICSBP/IRF-8 to be associated with an enhanced ubiquination of TNFR associated factor 6 (TRAF6), a protein that mediate the signal transduction from members of the TNF receptor superfamily, and the activation of AP-1 and NFκB transcription factors.

On the other hand, several studies demonstrated that ICSBP/IRF-8 promotes the differentiation and activation of dendritic cells and macrophages cells (Tamura and Ozato, 2002; Tamura et al., 2000) and that, at the same

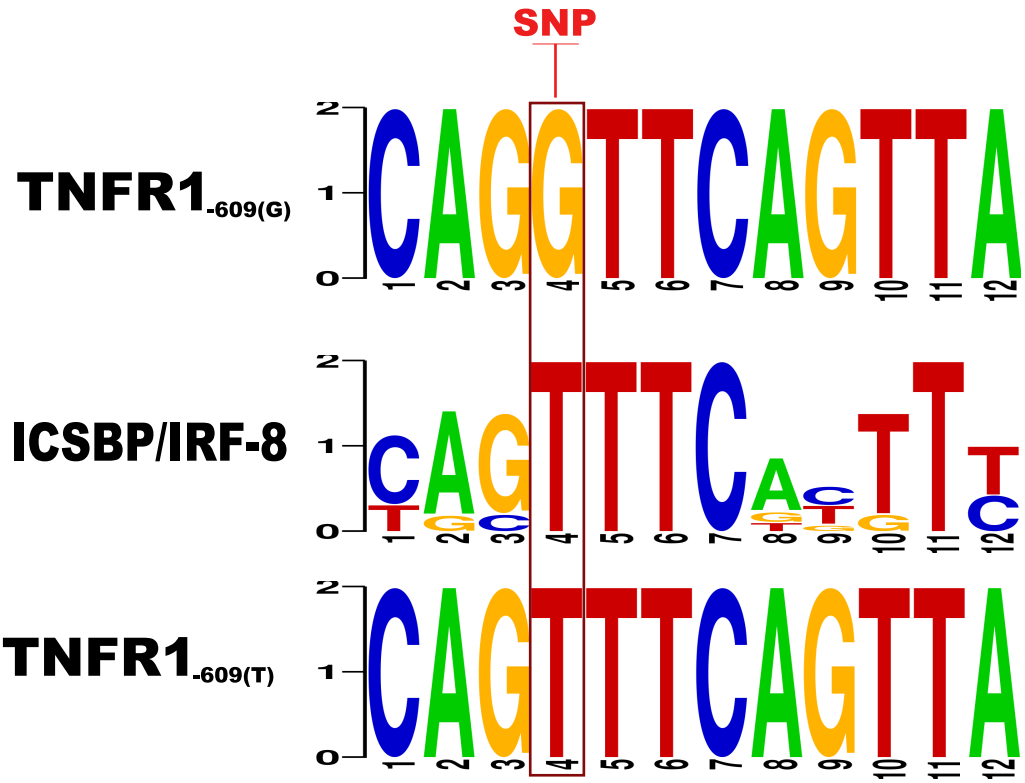

Figure 7.9: ***ICSBP against TNFR1$_{-609(G/T)}$ polymorphism.***

time, TNFR1 mRNA level is increased during this biological process (Schling et al., 2006).

Taken into account these observations, we hypothesize that the presence of TNFR1$_{-609(G/T)}$ promoter polymorphisms can modify the binding affinity to ICSBP/IRF-8 (see Figure 7.9) and, therefore, it could be used to predict susceptibility to infection and to facilitate risk stratification of hematological patients. However, the question of whether the TNFR1 polymorphisms have biological relevance regulating mRNA TNFR1 levels through ICSBP/IRF-8 transcription factor remains unanswered. Functional analysis should be performed to demonstrate the role of TNFR1$_{-609(G/T)}$ polymorphism mediating the binding of ICSBP/IRF-8 to TNFR1 promoter.

7.7 Concluding Remarks

In the present study, we have introduced SC_{intuit}, a new scoring method for measuring sequence-motif affinity, based on IFS theory. Our main objective was to improve the prediction quality for TFs of the existing approaches, reducing the false positive rate without compromising sensitivity. We show that SC_{intuit} outperforms other approaches in motif recognition tasks, and prove how it can be successfully applied to real research problems like ChIP-chip experiments or SNP analysis. Results for this last task suggest that the presence of G allele at position -609 might disrupt the binding affinity to interferon consensus binding protein (ICSBP/IRF-8). In the absence of T allele, the binding of ICSBP/IRF-8 to this region is predicted to be disrupted, potentially disturbing the transcriptional activation (Figure 7.9). We hypothesize that this putative transcription factor binding site might be involved in the initiation of TNFR1 transcription process, which should be confirmed with *in vitro* and/or *in vivo* experiments.

We have used our approach as a scanning method for the prediction of TFBSs, but it also can be incorporated with methods for *de novo* discovery of motifs. As intuitionistic theory is specially suitable for problems that deal with imprecise concepts, we are currently working on a fuzzy approach that applies the proposed scoring in an *ab initio* method to find motifs in large sets of related DNA sequences.

Part IV

Conclusions

CHAPTER 8

Conclusions and Future Work

8.1 Conclusions

This section summarizes the contributions of this thesis to the field of bioinformatics, analyzing the results in accordance with the initial objectives. This dissertation means an important step in the application of fuzzy technology, capable of representing and handling imprecise, vague and uncertain scenarios, to bioinformatics research problems. Throughout the document, we have presented examples in different real and synthetic experiments, such as protein classification, motif identification in co-regulated genes, or single nucleotide polymorphisms.

Objective 1. The first part of Chapter 4 fulfills the first of our objectives, which was to perform a critical review of the state of the art in GO semantic measures. We have applied these GO crisp semantic measures combined with different clustering methods in order to achieve protein family recognition. We have demonstrated that two of the clustering methods (kcmeans and csecmeans), provide better results than the others and they should be considered when dealing with imprecise data. Furthermore, we have showed that selecting an appropriate GO partition is a key issue for each problem. Although the results are promising and show that we can use the knowledge contained in GO to analyze sets of annotated proteins, we have seen that

they present some limitations. Some of them have been overcome in the next chapter.

Objective 2. The definition of a new fuzzy semantic similarity measure for GO in the second part of Chapter 4 fulfills the second of our objectives. We have extended the approach proposed by Keller et al. (2004) so that the reliability of the source of information is taken into account by means of the evidence codes of the annotations. We have provided a comparison of our new method with previously defined methods in terms of protein classifications. We have proved that our proposed measure provides results that outperforms previous techniques and are consistent with the biological meaning of the annotations and with the expected results.

Objective 3. Chapter 5 fulfills the third objective. We have exposed the state of the art in motif comparison measures, providing a review of the most popular probabilistic measures for motifs. We have also adapted different classes of classical measures for fuzzy sets and a measure defined for the fuzzy polynucleotide space, so that they can be applied to motif comparison tasks. We have proved the adequacy of fuzzy technology within motif comparison issues. The results showed that fuzzy measures provide excellent results when dealing with sets of randomly generated motifs and outperforms other existing measures when facing datasets of real motifs.

Objective 4. Chapter 6 proposes a new similarity measure for DNA motifs called FISim (Fuzzy Integral Similarity), which fulfills our fourth objective. FISim takes into account the relative importance of each nucleotide within a given position of the motif. For the validation of our new approach, we have introduced the recent advances in motif measures, including two recent methods that have shown better results than probabilistic approaches (Gupta et al., 2007; Pape et al., 2008). We have demonstrated that FISim outperforms other approaches in motif recognition tasks, and it can be successfully applied to real-life research problems. In addition, we have also proposed a novel clustering methodology for motifs based on our FISim measure and

kernel theory, showing promising results in terms of accuracy and cluster compactness. All of this proves the reliability of fuzzy technology for motif comparison tasks.

Objective 5. The definition of SC_{intuit}, a novel scoring method for measuring sequence-motif affinity, in Chapter 7 fulfills our fifth objective. SC_{intuit} is based on IFS theory, which makes our method more flexible at capturing the real meaning of the affinity sequence-motif given the uncertain available data. We have introduced the latest advances on this topic, showing that considering dependencies among motif positions is a key point for this matter. We have proved that SC_{intuit} outperforms other approaches in motif recognition tasks, and we have shown how it can be successfully applied to real research problems like ChIP-chip experiments or SNP analysis.

8.2 Future Work

The application of fuzzy techniques for bioinformatics problems is a relatively new field of research. Therefore, there is still much room for further work that improves and extends current approaches. Particularly, from the work presented in this thesis, some ideas that should be considered next are:

- Taking advantage of the DAG structure of GO, it would be very interesting to have a GO semantic measure that is directly a kernel so it can be incorporated in any kernel method.
- To extend classical fuzzy measures for motifs by taking into account the different importances of the motif positions regarding the observed conservation, e.g. incorporating weights to the positions according to their corresponding information content.
- To incorporate FISim and SC_{intuit} into a publicly available web platform that allows the scientific community to easily use them for their experiments.
- One of the main problems of the scoring methods for measuring sequence-motif affinity is that outlier sequences of the a given motif tend

to have relatively low scores when scoring them against their corresponding motif. It would be necessary to have work on a scoring method that increases those scores without overestimating of lower conserved positions of TFBSs.

- As fuzzy technology is especially suitable for problems that involving imprecise concepts, it is a natural next step to work on a fuzzy algorithm that applies both FISim and SC_{intuit} for finding *de novo* motifs in large sets of related DNA sequences.

8.3 Publications

Derived from this thesis

International Journals

- **F. Garcia**, A. Blanco, A. Shepherd, "An intuitionistic approach to scoring DNA sequences against transcription factor binding site motifs", BMC Bioinformatics (**Submitted**).
- **F. Garcia**, M. Masip, A. Blanco, J. Garcia-Castro, "Genome-Wide Differential Gene Expression Profiling of Human Peripheral Blood-Derived Mesenchymal Stem Cells", BMC Genomics (**Submitted**).
- **F. Garcia**, F. J. Lopez, C. Cano, A. Blanco, "FISim: A new similarity measure between TFBSs based on the fuzzy integral", BMC Bioinformatics, **10**:224, 2009.
- J. Sainz, I. Salas, E. Lopez, C. Olmedo, A. Comino, **F. Garcia**, A. Blanco, S. Oyonarte, P. Bueno, M. Jurado, "TNFR1 mRNA Expression Level and TNFR1 Gene Polymorphisms Are Predictive Markers for Susceptibility to Develop Invasive Pulmonary Aspergillosis", International Journal of Immunopathology and Pharmacology, **22**(3), 557-565, 2009.
- C. Olmedo, J. Martin-Cano, **F. Garcia-Alcalde**, A. M. Comino, K. Alexandrova, K. Muffak, L. Arseniev, D. Garrote, A Blanco, P. Bueno, J. A. Ferron, "Microarray study of gene expression profile in steatotic hepatocytes", Transplant International, **22**(Supp 2), 250, 2009.

- L. Hassan-Montero, P. Bueno, C. Olmedo, A.M. Comino, K. Muffak-Granero, **F. Garcia-Alcalde**, M. Serradilla, J.M. Villar, D. Garrote, A. Blanco, and J.A. Ferron, "Gene Expression Profiling in Liver Transplant Recipients With Alcoholic Cirrhosis", Transplantation Proceedings **40**, 2955-2958, 2008.
- K. Muffak-Granero, P. Bueno, C. Olmedo, A.M. Comino, L. Hassan, **F. Garcia-Alcalde**, M. Serradilla, A. García-Navarro, A. Mansilla, J.M. Villar, D. Garrote, A. Blanco, J.A. Ferron, "Study of Gene Expression Profile in Liver Transplant Recipients With Hepatitis C Virus", Transplantation Proceedings **40**, 2971-2974, 2008.

Conference Publications

- **F. Garcia**, F. J. Lopez, C. Cano, A. Blanco, "Study of fuzzy resemblance measures for DNA motifs", 2009 IEEE International Conference on Fuzzy Systems, Jeju, South Korea, 2009.
- **F. Garcia**, L. Adarve, F. J. Lopez, A. Blanco, "Assessment of gene ontology based recognition of related proteins", IADIS International Conference on Applied Computing, Algarve, Portugal, 2008.
- C. Benner, **F. Garcia**, S. Subramaniam, C. Glass, "HOMER: An algorithm for the de novo discovery of cis-regulatory elements from high throughput data", Algorithmic Biology 2006, San Diego, US, 2006.
- **F. Garcia**, F. J. Lopez, A. Blanco, C. Cano, "An ontology-driven similarity providing reliable protein family recognition", IADIS International Conference on Applied Computing, San Sebastian, Spain, 2006.
- M.D. Collado, J. Sainz, A. Blanco, **F. Garcia**, J. Lopez, R. Caliz, "Gene expression study by microarrays of peripheral blood white cells from patiens with developed rheumatoid arthritis", American College Rheumatology Conference, Washington, US, 2006.
- M.D. Collado, J. Sainz, **F. Garcia**, A. Blanco, J. Lopez, R. Caliz, "Peripheral blood white cells gene expression profile in early rheumatoid

arthritis patients by Code Link human whole genome bioarrays", American College Rheumatology Conference, Washington, US, 2006.

- J. Sainz, R. Caliz, L. Hassan, A. Blanco, **F. Garcia**, J. Lopez, "Differences in gene expression profile between early and developed rheumatoid arthritis", Annual European Congress of Rheumatology EULAR, Amsterdam, Netherlands, 2006.

- J. Sainz, M. D. Collado, A. Blanco, **F. Garcia**, J. Lopez, R. Caliz, "Peripheral blood white cells gene expression profile in early rheumatoid arthritis patients by Code Link human whole genome bioarrays", Annual European Congress of Rheumatology EULAR, Amsterdam, Netherlands, 2006.

- M.D. Collado, J. Sainz, A. Blanco, L. Hassan, **F. Garcia**, J. Lopez, "Gene expression characterization of peripheral blood white cells from patients with developed rheumatoid arthritis by microarrays", Annual European Congress of Rheumatology EULAR, Amsterdam, Netherlands, 2006.

- J. Sainz, A. Barroso, A. Blanco, **F. Garcia**, C. Cano, A. Concha, "Gene Expression Profiling in Mouse Embryonic Stem Cells", Simposio Internacional sobre Nuevos Avances en Medicina Reproductiva, Valencia, Spain, 2005.

- J. Sainz, A. Barroso, A. Blanco, **F. Garcia**, C. Cano, A. Concha, "Microarray Analysis of Mouse Embrionic Stem Cells", 32 Symposium Internacional Fertilidad, Barcelona, Spain, 2005.

- J. Sainz, M.D. Collado, L. Hassan, A. Blanco, **F. Garcia**, J. Lopez, "Perfil de expresion genica en la artritis reumatoide temprana", XVI Congreso de la Sociedad Andaluza de Reumatologia, Cadiz, Spain, 2005.

- J. Sainz, M.D. Collado, L. Hassan, A. Blanco, **F. Garcia**, J. Lopez, "Perfil de expresion genica en la artritis reumatoide tardia", XVI Congreso de la Sociedad Andaluza de Reumatologia, Cadiz, Spain, 2005.

Related with this thesis (collaborator)

International Journals and Book Chapters

- F. J.Lopez, **F. Garcia**, A. Blanco, C. Cano "Extracting biological knowledge by association rule mining", Data mining in biomedicine using ontologies. Artech, Boston, US, 2009. (**Book Chapter**)
- 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.
- F. J. Lopez, A. Blanco, **F. Garcia**, C. Cano, A. Marin, "Fuzzy association rules for biological data analysis: a case study on yeast", BMC Bioinformatics, **9**(1), 1-18, 2008.
- C. Cano, F. J. Lopez, **F. Garcia**, A. Blanco, "Evolutionary algorithms for finding interpretable patterns in gene expression data", IADIS international journal on computer science and information systems, pp. 88-99, 2006.

Conference Publications

- F. J. Lopez, C. Cano, **F. Garcia**, A. Blanco, "A fuzzy approach for studying combinatorial regulatory actions of transcription factors in yeast", IDEAL09, Burgos, Spain, 2009.
- F. J. Lopez, **F. Garcia**, A. Blanco, C. Cano, A. Marin, "A fuzzy approach for the study of functional and structural features of the yeast genome", 13th Evolutionary Biology Meeting at Marseilles, Marseilles, France, 2009.
- F. J. Lopez, **F. Garcia**, A. Blanco, S. Blanco, "Aplicacion de las reglas de asociacion difusas en proteomica", XIV Congreso sobre Tecnologias y Logica Fuzzy ESTYLF 2008, Oviedo, Spain, 2008.
- C. Cano, F.L. Adarve, **F. Garcia**, F. J. Lopez, A. Blanco, "Non-supervised identification of gene regulatory modules by possibilistic biclustering

of microarray data" 11TH International Conference on Cognitive and Neural Systems, Boston, US, 2007.

- F. J. Lopez, **F. Garcia**, A. Blanco, A. Marin, "Extracting biological knowledge by fuzzy association rule mining", 2007 IEEE International Conference on Fuzzy Systems, London, UK, 2007.
- C. Cano, A. Blanco, **F. Garcia**, F. J. Lopez, "Evolutionary Algorithms for Finding Interpretable Patterns in Gene Expression Data", IADIS International Conference on Applied Computing, San Sebastian, 2006.

Bibliography

SF Altschul, TL Madden, AA Schaffer, J. Zhang, Z. Zhang, W. Miller, and DJ Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389, 1997.

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.

K. Atanassov and G. Gargov. Intuitionistic fuzzy logic. *Compt. Rend. Acad. Bulg. Sci*, 43:9–12, 1990.

K.T. Atanassov. New operations defined over the intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, 61(2):142, 1994.

KT Attanasov. Intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, 20:87–96, 1986.

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, pages 28–36. Citeseer, 1994.

A. Bairoch, R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, et al. The universal protein resource (UniProt). *Nucleic Acids Research*, 33(Database Issue):D154, 2005.

E. Baker, LZ Chen, CA Smith, DF Callen, R. Goodwin, and GR Sutherland. Chromosomal location of the human tumor necrosis factor receptor genes. *Cytogenetics and cell genetics*, 57(2-3):117, 1991.

S. Bandyopadhyay. An efficient technique for superfamily classification of amino acid sequences: feature extraction, fuzzy clustering and prototype selection. *Fuzzy Sets and Systems*, 152(1):5–16, 2005.

A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S.R. Eddy, S. Griffiths-Jones, K.L. Howe, M. Marshall, and E.L.L. Sonnhammer. The Pfam protein families database. *Nucleic acids research*, 30(1):276, 2002.

J.C. Bean. Genetic Algorithms and Random Keys for Sequencing and Optimization. *INFORMS Journal on Computing*, 6(2):154, 1994.

K A Becker, S Lu, E S Dickinson, K A Dunphy, L Mathews, S S Schneider, and D J Jerry. Estrogen and progesterone regulate radiation-induced p53 activity in mammary epithelium through tgf-beta-dependent pathways. *Oncogene*, 24(42):6345–6353, 2005.

P.V. Benos, A.S. Lapedes, and G.D. Stormo. Probabilistic code for DNA recognition by proteins of the EGR family. *Journal of molecular biology*, 323(4): 701–727, 2002.

DA Benson, MS Boguski, DJ Lipman, J. Ostell, BF Ouellette, BA Rapp, and DL Wheeler. GenBank. *Nucleic acids research*, 27(1):12, 1999.

JM Berg, JL Tymoczko, L. Stryer, and L. Stryer. *Biochemistry, Ed 5th*. WH Freeman, New York, 2002.

HM Berman, T. Battistuz, TN Bhat, WF Bluhm, PE Bourne, K. Burkhardt, Z. Feng, GL Gilliland, L. Iype, S. Jain, et al. The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, 58(6):899–907, 2002.

A. Bernstein, E. Kaufmann, C. Burki, and M. Klein. How similar is it? towards personalized similarity measures in ontologies. In *7th International Conference Wirtschaftsinformatik (WI-2005), Bamberg, Germany*, pages 1347–1366. Springer, 2005.

J.C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers Norwell, MA, USA, 1981.

A. Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, pages 401–406, 1946.

P.Y. Bochud, J.W. Chien, K.A. Marr, W.M. Leisenring, A. Upton, M. Janer, S.D. Rodrigues, S. Li, J.A. Hansen, L.P. Zhao, et al. Toll-like receptor 4 polymorphisms and aspergillosis in stem-cell transplantation. *New England Journal of Medicine*, 359(17):1766, 2008.

A. Budanitsky and G. Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources*, volume 2. Citeseer, 2001.

M.L. Bulyk, P.L.F. Johnson, and G.M. Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic acids research*, 30(5):1255, 2002.

S. Carbon, A. Ireland, C.J. Mungall, S.Q. Shu, B. Marshall, S. Lewis, et al. AmiGO: online access to ontology and annotation data. *Bioinformatics*, 25 (2):288, 2009.

G. Chamilos, M. Luna, RE Lewis, GP Bodey, R. Chemaly, JJ Tarrand, A. Safdar, I.I. Raad, and DP Kontoyiannis. Invasive fungal infections in patients with hematologic malignancies in a tertiary care cancer center: an autopsy study over a 15-year period (1989-2003). *Haematologica*, 91(7): 986, 2006.

X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V.B. Vega, E. Wong, Y.L. Orlov, W. Zhang, J. Jiang, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6): 1106–1117, 2008.

I G Choi, J Kwon, and S H Kim. Local feature frequency profile: A method to measure structural similarity in proteins. *PNAS*, 101:3797–2892, 2004.

O. Cordon, F. Gomide, F. Herrera, F. Hoffmann, and L. Magdalena. Ten years of genetic fuzzy systems: current framework and new trends. *Fuzzy sets and systems*, 141(1):5–31, 2004.

G E Crooks, G Hon, J Chandonia, and S E Brenner. Weblogo: A sequence logo generator. *Genome Res*, 14:1188–1190, 2004.

V. Cross and T.A. Sudkamp. *Similarity and compatibility in fuzzy set theory: assessment and applications*. Springer, 2002.

M. Das and H.K. Dai. A survey of DNA motif finding algorithms. *BMC bioinformatics*, 8(Suppl 7):S21, 2007.

MO Dayhoff. Computer analysis of protein evolution. *Scientific American*, 221(1):86, 1969.

S.K. De, R. Biswas, and A.R. Roy. An application of intuitionistic fuzzy sets in medical diagnosis. *Fuzzy Sets and Systems*, 117(2):209–213, 2001a.

S.K. De, R. Biswas, and A.R. Roy. An application of intuitionistic fuzzy sets in medical diagnosis. *Fuzzy Sets and Systems*, 117(2):209–213, 2001b.

J. Demeter, C. Beauheim, J. Gollub, T. Hernandez-Boussard, H. Jin, D. Maier, J.C. Matese, M. Nitzberg, F. Wymore, Z.K. Zachariah, et al. The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic acids research*, 2006.

D.W. Denning. Invasive aspergillosis. *Clinical infectious diseases*, pages 781–803, 1998.

P D'haeseleer. What are dna sequence motifs? *Nature Biotechnology*, 24(4): 423–425, 2006.

G. Diop, J.L. Spadoni, H. Do, T. Hirtzig, C. Coulonges, T. Labib, W. Issing, J. Rappaport, A. Therwath, M. Lathrop, et al. Genomic approach of AIDS pathogenesis: exhaustive genotyping of the TNFR1 gene in a French AIDS cohort. *Biomedicine & Pharmacotherapy*, 59(8):474–480, 2005.

N. Dror, M. Alter-Koltunoff, A. Azriel, N. Amariglio, J. Jacob-Hirsch, S. Zeligson, A. Morgenstern, T. Tamura, H. Hauser, G. Rechavi, et al. Identification of IRF-8 and IRF-1 target genes in activated macrophages. *Molecular immunology*, 44(4):338–346, 2007.

D. Dubois, F. Esteva, L. Godo, and H. Prade. *Fuzzy-set based logics: an history-oriented presentation of their main developments*. The Many Valued and Nonmonotonic Turn in Logic, North Holland, 2007.

M Dutertre and C L Smith. Ligand-independent interactions of p160/steroid receptor coactivators and creb-binding protein (cbp) with estrogen receptor-α: Regulation by phosphorylation sites in the a/b region depends on other receptor domains. *Molecular Endocrinology*, 17(7):1296–1314, 2003.

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, et al. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Research*, 30(1):69, 2002.

M.B. Eisen. All motifs are not created equal: structural properties of transcription factor-DNA interactions and the inference of sequences specificity. *Genome Biology*, 6(5):P7, 2005.

M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863, 1998.

T. Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27 (8):861–874, 2006.

C. Fellbaum et al. *WordNet: An electronic lexical database*. MIT press Cambridge, MA, 1998.

G B Fogel, D G Weekes, G Varga, E R Dow, A M Craven, H B Harlow, E W Su, J E Onyia, and C Su. A statistical analysis of the transfac data. *Biosystem*, 81(2):137–154, 2005.

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, 2009.

GEO. Gene expression omnibus. National Center for Biotechnology Information, 2009. URL http://www.ncbi.nlm.nih.gov/geo/. http://www.ncbi.nlm.nih.gov/geo/.

B P Gomez, R B Riggins, A N Shajahan, U Klimach, A Wang, A C Crawford, Y Zhu, A Zwart, M Wang, and R Clarke. Human x-box binding protein-1

confers both estrogen independence and antiestrogen resistance in breast cancer cell lines. *FASEB J.*, 21(14):4013–4027, 2007.

S Gupta, J A Stamatoyannopoulos, T L Bailey, and W S Noble. Quantifying similarity between motifs. *Genome Biol*, 8:R24, 2007.

P. Hájek. Making fuzzy description logic more general. *Fuzzy Sets and Systems*, 154(1):1–15, 2005.

J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, 1982.

MA Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, et al. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(Database issue):D258, 2004.

J.A. Hartigan. *Clustering algorithms*. Wiley New York, 1975.

T. Hehlgans and K. Pfeffer. The intriguing biology of the tumour necrosis factor/tumour necrosis factor receptor superfamily: players, rules and the games. *Immunology*, 115(1):1, 2005.

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, 1990.

AS Hinrichs, D. Karolchik, R. Baertsch, GP Barber, G. Bejerano, H. Clawson, M. Diekhans, TS Furey, RA Harte, F. Hsu, et al. The UCSC genome browser database: update 2006. *Nucleic acids research*, 34(Database Issue):D590, 2006.

J.H. Holland. Genetic algorithms. *Scientific American*, 267(1):66–72, 1992.

Y Huang and Y Li. Prediction of protein subcellular locations using fuzzy k-nn method. *Bioinformatics*, 20:21–28, 2004.

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.

W.L. Hung and M.S. Yang. Similarity measures of intuitionistic fuzzy sets based on Hausdorff distance. *Pattern Recognition Letters*, 25(14):1603–1611, 2004.

A. Islam and D. Inkpen. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2(2):10, 2008.

P. Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Societe de Vaud des Sciences Naturelles*, 44:223, 1908.

K. Janowicz, M. Raubal, A. Schwering, and W. Kuhn. Semantic similarity measurement and geospatial applications. *Transactions in GIS*, 12(6):651–659, 2008.

J. Jiang and D. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of International Conference on Computational Linguistics (ROCLING X)*, 1997.

L Kaufman and P Rousseew. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Son, New York, 1990.

J Keller, P Gader, and A Hocauglu. Fuzzy integrals in image processing and recognition. In *Fuzzy Measures and Integrals: Theory and Applications*, pages 435–466. Springer, Berlin, 2000.

JM Keller, M. Popescu, and J. Mitchell. Taxonomy-based soft similarity measures in bioinformatics. In *2004 IEEE International Conference on Fuzzy Systems, 2004. Proceedings*, volume 1, 2004.

W.J. Kent, F. Hsu, D. Karolchik, R.M. Kuhn, H. Clawson, H. Trumbower, and D. Haussler. Exploring relationships and mining data with the UCSC Gene Sorter. *Genome research*, 15(5):737, 2005.

V. Khatibi and G.A. Montazer. Intuitionistic fuzzy set vs. fuzzy set application in medical pattern recognition. *Artificial Intelligence In Medicine*, 47(1): 43–52, 2009.

K. Kim, M. Kim, and Y. Woo. A DNA sequence alignment algorithm using quality information and a fuzzy inference method. *Progress in Natural Science*, 18(5):595–602, 2008a.

S. Kim, S.M. Moon, Y.S. Kim, J.J. Kim, H.J. Ryu, Y.J. Kim, J.W. Choi, H.S. Park, D.G. Kim, H.D. Shin, et al. TNFR1 promoter- 329G/T polymorphism results in allele-specific repression of TNFR1 expression. *Biochemical and biophysical research communications*, 368(2):395–401, 2008b.

O.D. King, R.E. Foulger, S.S. Dwight, J.V. White, and F.P. Roth. Predicting gene function from patterns of annotation. *Genome research*, 13(5):896, 2003.

T. Kislinger, K. Rahman, D. Radulovic, B. Cox, J. Rossant, and A. Emili. PRISM, a generic large scale proteomic investigation strategy for mammals. *Molecular & Cellular Proteomics*, 2(2):96–106, 2003.

G.J. Klir and B. Yuan. *Fuzzy sets and fuzzy logic: theory and applications*. Prentice Hall Upper Saddle River, NJ, 1995.

J.C. Knight. Regulatory polymorphisms underlying complex disease traits. *Journal of molecular medicine*, 83(2):97–109, 2005.

R. Krishnapuram and J.M. Keller. A possibilistic approach to clustering. *IEEE transactions on fuzzy systems*, 1(2):98–110, 1993.

R. Krishnapuram and JM Keller. The possibilistic c-means algorithm: insights and recommendations. *IEEE Transactions on Fuzzy Systems*, 4(3):385–393, 1996.

E.V. Kriventseva, W. Fleischmann, E.M. Zdobnov, and R. Apweiler. CluSTr: a database of clusters of SWISS-PROT+ TrEMBL proteins. *Nucleic acids research*, 29(1):33, 2001a.

E.V. Kriventseva, W. Fleischmann, E.M. Zdobnov, and R. Apweiler. Clustr: a database of clusters of swiss-prot+ trembl proteins. *Nucleic acids research*, 29(1):33, 2001b.

D. Kwiatkowski. Science, medicine, and the future: susceptibility to infection. *British Medical Journal*, 321(7268):1061, 2000.

A. Lægreid, T.R. Hvidsten, H. Midelfart, J. Komorowski, and A.K. Sandvik. Predicting gene ontology biological process from temporal gene expression patterns. *Genome Research*, 13(5):965, 2003.

F.H. Lam, D.J. Steger, and E.K. O'Shea. Chromatin decouples promoter threshold from dynamic range. *Nature*, 453(7192):246–250, 2008.

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.

C.S. Lee, Y.F. Kao, Y.H. Kuo, and M.H. Wang. Automated ontology construction for unstructured text documents. *Data & Knowledge Engineering*, 60 (3):547–566, 2007.

J.H. Lee, M.H. Kim, and Y.J. Lee. Information retrieval based on conceptual distance in IS-A hierarchies. *Journal of Documentation*, 49(2):188–207, 1993.

B. Lewin. *Genes VIII*. Pearson Prentice Hall, 2004.

L. Liang, V. Mandal, Y. Lu, and D. Kumar. MCM-test: a fuzzy-set-theory-based approach to differential analysis of gene pathways. *BMC bioinformatics*, 9 (Suppl 6):S16, 2008.

D. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Citeseer, 1998.

X S Liu, D L Brutlag, and J S Liu. An algorithm for finding protein-dna binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology*, 20(9):835–839, 2002.

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 (1):107, 2008.

PW Lord, RD Stevens, A. Brass, and CA Goble. Semantic similarity measures as tools for exploring the gene ontology. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 8, pages 601–612, 2003.

S Mahony, P E Auron, P V Benos, and G Stormo. Dna familial binding profiles made easy: Comparison of various motif alignment and clustering strategies. *PLoS Comput Biol*, 3:578–591, 2007.

C.D. Manning and H. Schutze. *Foundations of statistical natural language processing*. MIT Press, 2002.

B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th international conference on World wide web*, pages 641–650. ACM New York, NY, USA, 2009.

V Matys, O V Kel-Margoulis, E Fricke, I Liebich, S Land, A Barre-Dirrie, I Reuter, D Chekmenev, M Krull, K Hornischer, N Voss, P Stegmaier, B Lewicki-Potapov, H Saxel, A E Kel, and E Wingender. Transfac®and its module transcompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34(Database issue):D108–D110, 2006.

V.A. McKusick and S.E. Antonarakis. *Mendelian inheritance in man: a catalog of human genes and genetic disorders*. Johns Hopkins University Press, 1998.

HW Mewes, K. Heumann, A. Kaps, K. Mayer, F. Pfeiffer, S. Stocker, and D. Frishman. MIPS: a database for genomes and protein sequences. *Nucleic acids research*, 27(1):44, 1999.

A. Mukhopadhyay and U. Maulik. Towards improving fuzzy clustering using support vector machine: Application to gene expression data. *Pattern Recognition*, 42(11):2744–2763, 2009.

T. Murofushi and M. Sugeno. An interpretation of fuzzy measures and the Choquet integral as an integral with respect to a fuzzy measure. *Fuzzy Sets and Systems*, 29(2):201–227, 1989.

NIH. Pubmed database. National Library of Medicine, 2009. URL http://www.nlm.nih.gov/. http://www.nlm.nih.gov/.

F. Offner, C. Cordonnier, P. Ljungman, HG Prentice, D. Engelhard, D.D. Bacquer, F.M., , and B.D. Pauw. Impact of previous aspergillosis on the outcome of bone marrow transplantation. *Clinical infectious diseases*, 26(5): 1098–1103, 1998.

R Osada, E Zaslavsky, and M Singh. Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics*, 20:3516–3525, 2004.

L. Painton and J. Campbell. Genetic algorithms in optimization of system reliability. *IEEE Transactions on Reliability*, 44(2):172–178, 1995.

Y Pan. Advances in the discovery of cis-regulatory elements. *Current Bioinformatics*, 1:321–336, 2006.

U J Pape, S Rahmann, and M Vingron. Natural similarity measures between position frequency matrices with an application to clustering. *Bioinformatics*, 24:350–357, 2008.

P.M. Pardalos and M.G.C. Resende. *Handbook of applied optimization*. Oxford University Press New York;, 2002.

G Pavesi, P Mereghetti, F Zambelli, M Stefani, G Mauri, and G Pesole. Mod tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. *Nucleic Acids Research*, 34(Web server issue):W566–W570, 2006.

S Pietrokovski. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res*, 24:3836–3845, 1996.

M. Popescu, J.M. Keller, and J.A. Mitchell. Fuzzy measures on the gene ontology for gene product similarity. *IEEE/ACM Transactions on computational biology and bioinformatics*, pages 263–274, 2006.

P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 448–453. Citeseer, 1995.

H. Ressom, R. Reynolds, and R.S. Varghese. Increasing the efficiency of fuzzy logic-based gene expression data analysis. *Physiological Genomics*, 13(2): 107, 2003.

R. Richardson and A.F. Smeaton. Using WordNet in a knowledge-based approach to information retrieval. In *Proceedings of the BCS-IRSG Colloquium, Crewe*. Citeseer, 1995.

S Roepcke, S Grossmann, S Rahmann, and M Vingron. T-reg comparator: an analysis tool for the comparison of position weight matrices. *Nucleic Acids Res*, 33:438–441, 2005.

K A Romer, G R Kayombya, and E Fraenkel. Webmotifs: automated discovery, filtering and scoring of dna sequence motifs using multiple programs and bayesian approaches. *Nucleic Acids Research*, 35(Web server issue):W217–W220, 2007.

V. Roth, J. Laub, J. Buhmann, and K.R. Muller. Going metric: Denoising pairwise data. *Advances in Neural Information Processing Systems*, pages 841–848, 2003a.

V. Roth, J. Laub, M. Kawanabe, and J.M. Buhmann. Optimal cluster preserving embedding of nonmetric proximity data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1540–1551, 2003b.

K. Sadegh-Zadeh. The fuzzy polynucleotide space revisited. *Artificial Intelligence in Medicine*, 41(1):69–80, 2007.

J. Sainz, I. Salas, E. Lopez, C. Olmedo, A. Comino, F. Garcia, A. Blanco, S. Oyonarte, P. Bueno, and M. Jurado. TNFR1 mRNA Expression Level and TNFR1 Gene Polymorphisms Are Predictive Markers for Susceptibility to Develop Invasive Pulmonary Aspergillosis. *International Journal of Immunopathology and Pharmacology*, 22(3):557–565, 2009.

A Sandelin and W W Wasserman. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol*, 338:207–215, 2004.

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, 2004a.

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, 2004b.

P. Schling, C. Rudolph, S. Heimerl, S. Fruth, and G. Schmitz. Expression of tumor necrosis factor alpha and its receptors during cellular differentiation. *Cytokine*, 33(5):239–245, 2006.

B. Schölkopf, K. Tsuda, and J.P. Vert. *Kernel methods in computational biology*. MIT press Cambridge, MA, 2004.

D.E. Schones, P. Sumazin, and M.Q. Zhang. Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, 21(3):307, 2005.

J.L. Sevilla, V. Segura, A. Podhorski, E. Guruceaga, J.M. Mato, L.A. Martinez-Cruz, F.J. Corrales, and A. Rubio. Correlation between gene expression and GO semantic similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2(4):338, 2005.

A.A. Sharov and M.S.H. Ko. Exhaustive Search for Over-represented DNA Sequence Motifs with CisFinder. *DNA Research*, 2009.

TF Smith and MS Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

T Sørlie, C M Peroua, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, M B Eisen, M van de Rijn, S S Jeffrey, T Thorsen, H Quist, J C Matesec, P O Brown, D Botstein, P E Lønning, , and A Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *PNAS*, 98(19):10869–10874, 2001.

M Sugeno. Fuzzy measures and fuzzy integrals: A survey. In *Fuzzy Automata and Decision Processes*, pages 89–102. North Holland, Amsterdam, 1977.

E. Szmidt and J. Kacprzyk. Intuitionistic fuzzy sets in group decision making. *Notes on IFS*, 2(1):11–14, 1996.

T. Tamura and K. Ozato. Review: ICSBP/IRF-8: its regulatory roles in the development of myeloid cells. *Journal of Interferon & Cytokine Research*, 22(1):145–152, 2002.

T. Tamura, T. Nagamura-Inoue, Z. Shmeltzer, T. Kuwata, and K. Ozato. ICSBP directs bipotential myeloid progenitor cells to differentiate into mature macrophages. *Immunity*, 13(2):155–165, 2000.

P.N. Tan, M. Steinbach, and V. Kumar. *Introduction to data mining*. Pearson Addison Wesley Boston, 2005.

L. Tari, C. Baral, and S. Kim. Fuzzy c-means clustering with prior biological knowledge. *Journal of Biomedical Informatics*, 42(1):74–81, 2009.

A. Tomovic and E.J. Oakeley. Position dependencies in transcription factor binding sites. *Bioinformatics*, 23(8):933, 2007.

M Tompa, N Li, T L Bailey, G M Church, B De Moor, E Eskin, A Favorov, M C Frith, Y Fu, W J Kent, V Makeev, A Mironov, W Noble, G Pavesi, G Pesole, M Régnier, N Simonis, S Sinha, G Thijs, J van Helden, M Vandenbogaert, A Weng, C Workman, and C adn Zhu Z Ye. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23(1):137–144, 2005.

J Torchia, D W Rose, J Inostroza, Y Kamei, S Westin, C K Glass, and M G Rosenfeld. The transcriptional co-activator p/cip binds cbp and mediates nuclear-receptor function. *Nature*, 387:677–684, 1997.

A. Torres and J.J. Nieto. The fuzzy polynucleotide space: basic properties. *Bioinformatics*, 19(5):587–592, 2003.

J. Van Helden. Regulatory sequence analysis tools. *Nucleic acids research*, 31 (13):3593, 2003.

S J Van Laere, I Van Auwera, G G Van Eynden, H J Elst, J Weyler, A L Harris, P Van Dam, E A Van Marck, P B Vermeulen, and L Y Dirix. Nuclear factor-κb signature of inflammatory breast cancer by cdna microarray validated by quantitative real-time reverse transcription-pcr, immunohistochemistry, and nuclear factor-κb dna-binding. *Clinical Cancer Research*, 12:3249–3256, 2006.

T Wang and G D Stormo. Combining phylogenetic data with co-regulated genes to to identify regulatory motifs. *Bioinformatics*, 19:2369–2380, 2003.

Y.P. Wang, M. Gunampally, J. Chen, D. Bittel, M.G. Butler, and W.W. Cai. A Comparison of Fuzzy Clustering Approaches for Quantification of Microar-

ray Gene Expression. *Journal of Signal Processing Systems*, 50(3):305–320, 2008.

W.W. Wasserman and J.W. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *Journal of molecular biology*, 278 (1):167–181, 1998.

W.W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287, 2004.

J. D. Watson and F. H. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

J. West, J. Cogan, M. Geraci, L. Robinson, J. Newman, J.A. Phillips, K. Lane, B. Meyrick, and J. Loyd. Gene expression in BMPR 2 mutation carriers with and without evidence of Pulmonary Arterial Hypertension suggests pathways relevant to disease penetrance. *BMC Medical Genomics*, 1(1):45, 2008.

B J Wilson and V Giguere. Meta-analysis of human cancer microarrays reveals gata3 is integral to the estrogen receptor alpha pathway. *Mol Cancer*, 7(49), 2008.

E. Wingender. TheTRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform*, 9: 326–332, 2008.

A.H. Wright. Genetic algorithms for real parameter optimization. *Foundations of genetic algorithms*, 1:205–218, 1991.

R.R. Yager. On a general class of fuzzy connectives. *Fuzzy sets and Systems*, 4(3):235–242, 1980.

L Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

F. Zare-Mirakabad, H. Ahrabian, M. Sadeghi, A. Nowzari-Dalini, and B. Goliaei. New scoring schema for finding motifs in DNA Sequences. *BMC bioinformatics*, 10(1):93, 2009.

J. Zhao, H.J. Kong, H. Li, B. Huang, M. Yang, C. Zhu, M. Bogunovic, F. Zheng, L. Mayer, K. Ozato, et al. IRF-8/interferon (IFN) consensus sequence-

binding protein is involved in Toll-like receptor (TLR) signaling and contributes to the cross-talk between TLR and IFN-γ signaling pathways. *Journal of Biological Chemistry*, 281(15):10073, 2006.

X. Zhou, M.C.J. Kao, and W.H. Wong. Transitive functional annotation by shortest-path analysis of gene expression data. *Proceedings of the National Academy of Sciences*, 99(20):12783, 2002.

R. Zwick, E. Carlstein, and D.V. Budescu. Measures of similarity among fuzzy concepts: A comparative analysis. *International Journal of Approximate Reasoning*, 1(2):221–242, 1987.

Appendix A

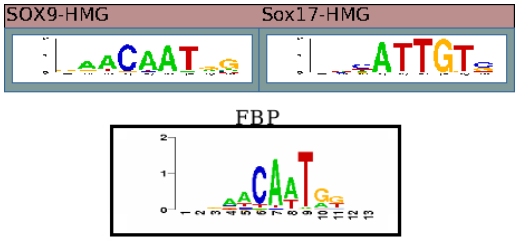

Figure .1: Cluster 1

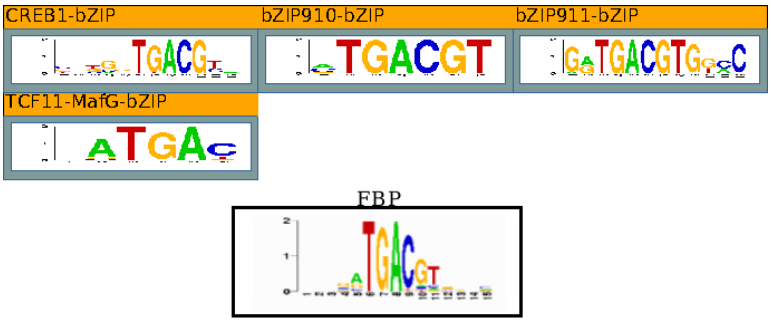

Figure .2: Cluster 2

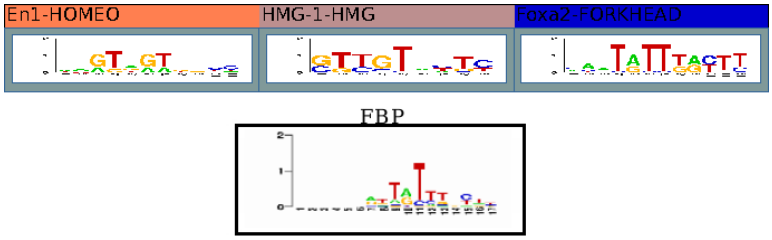

Figure .3: Cluster 3

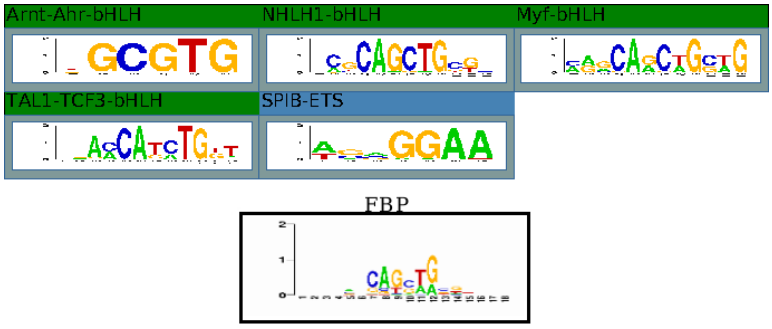

Figure .4: Cluster 4

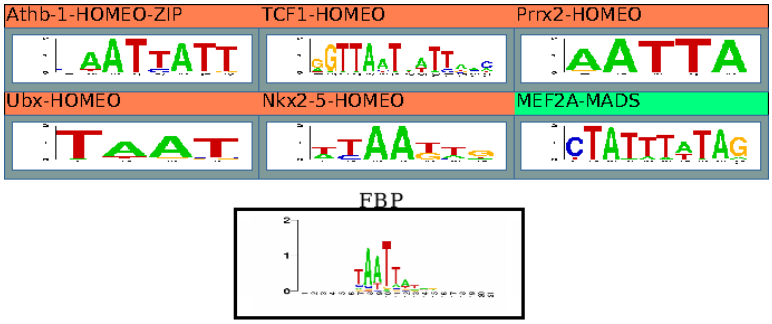

Figure .5: Cluster 5

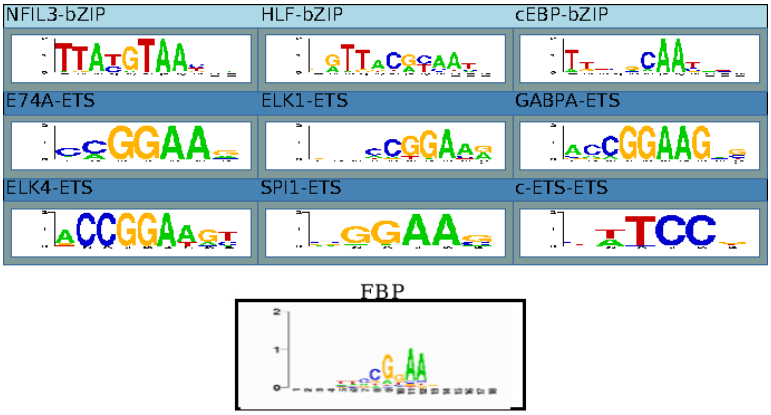

Figure .6: Cluster 6

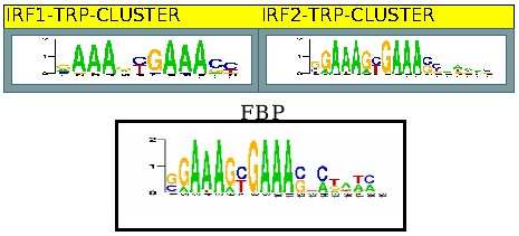

Figure .7: Cluster 7

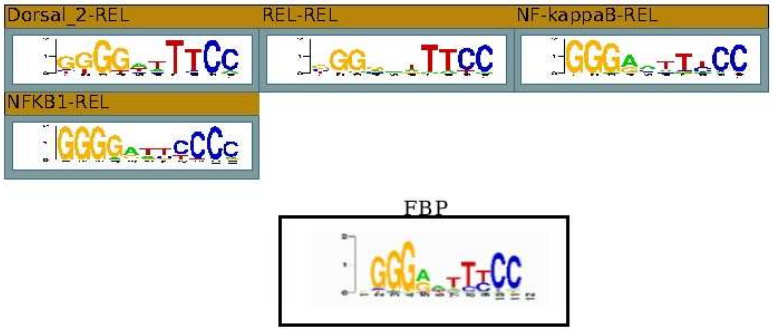

Figure .8: Cluster 8

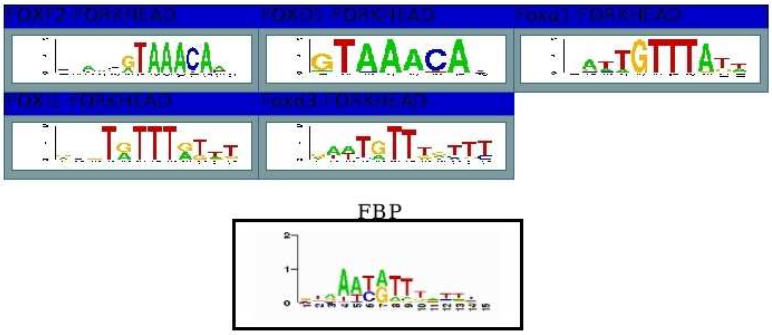

Figure .9: Cluster 9

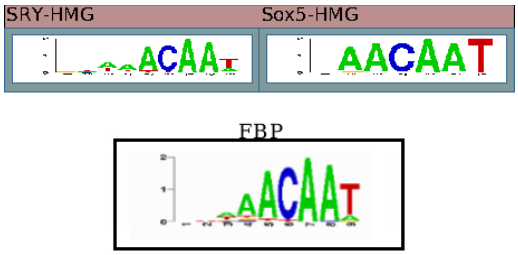

Figure .10: Cluster 10

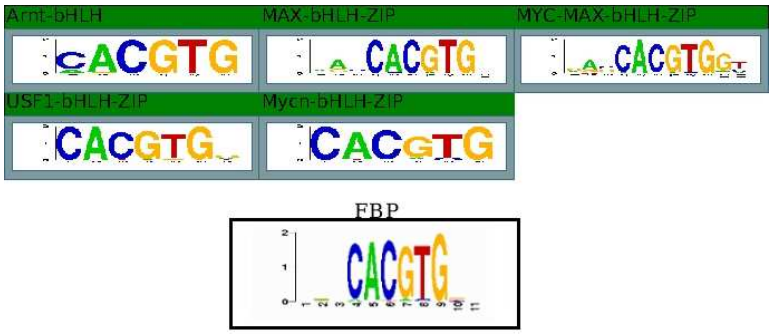

Figure .11: Cluster 11

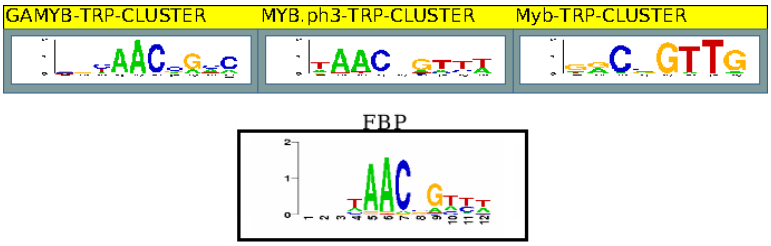

Figure .12: Cluster 12

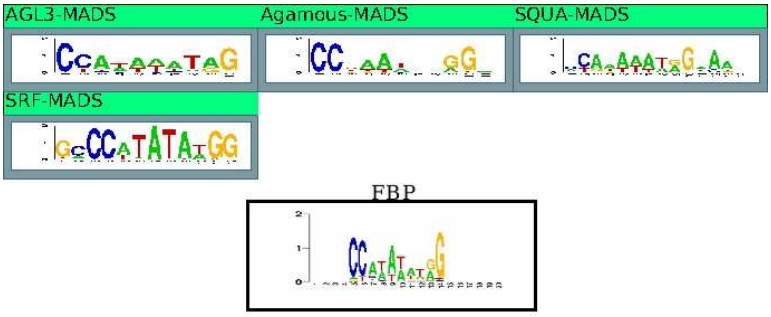

Figure .13: Cluster 13

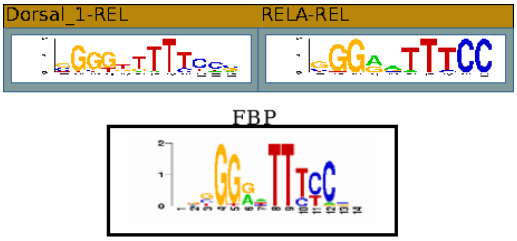

Figure .14: Cluster 14

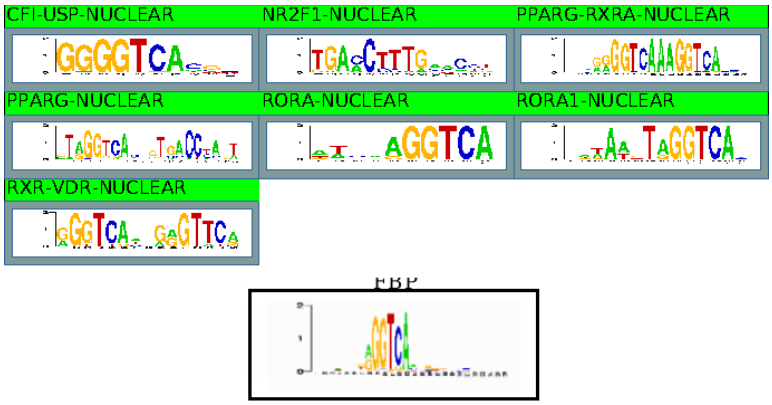

Figure .15: Cluster 15

Appendix B

Retrieved motifs

In this section we show the most significant motifs obtained from the three motif discovery programs used.

MDSCAN

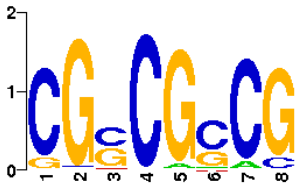

Figure .16: MDSCAN-1

Figure .17: MDSCAN-2

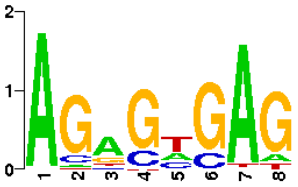

Figure .18: MDSCAN-3

Figure .19: MDSCAN-4

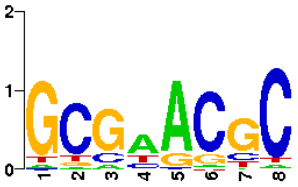

Figure .20: MDSCAN-5

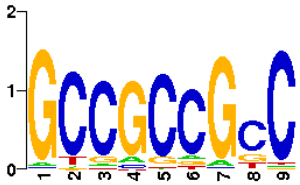

Figure .21: MDSCAN-6

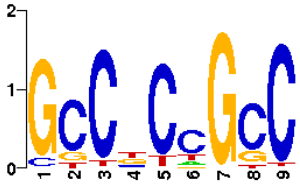

Figure .22: MDSCAN-7

MEME

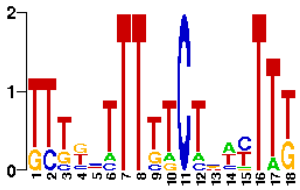

Figure .23: MEME-1

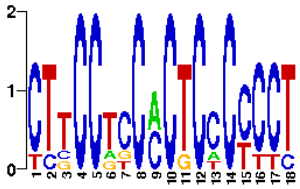

Figure .24: MEME-2

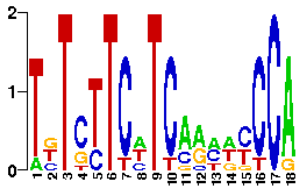

Figure .25: MEME-3

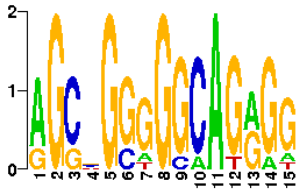

Figure .26: MEME-4

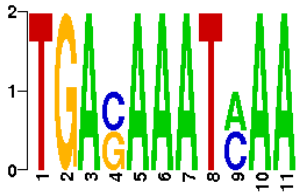

Figure .27: MEME-5

Figure .28: MEME-6

Weeder

Figure .29: Weeder-1

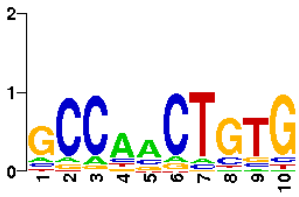

Figure .30: Weeder-2

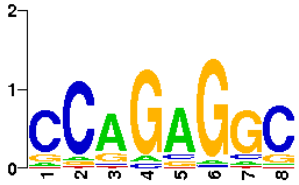

Figure .31: Weeder-3

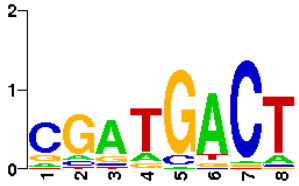

Figure .32: Weeder-4

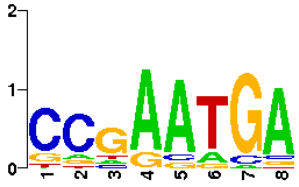

Figure .33: Weeder-5