UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE INFORMÁTICA
Departamento de Ingeniería del Software e Inteligencia Artificial

TESIS DOCTORAL

Machine learning based methods for the study of metabolism and its effect on the behavior of biological systems

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Clara Higuera Cabañes

Directores

Gonzalo Pajares Martinsanz
Federico Morán Abad

Madrid, 2015

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE INFORMÁTICA
Departamento de Ingeniería del Software e Inteligencia Artificial

TESIS DOCTORAL

Métodos basados en aprendizaje automático para el estudio del metabolismo y su efecto en el comportamiento de sistemas biológicos

Machine learning based methods for the study of metabolism and its effect on the behavior of biological Systems

Clara Higuera Cabañes

2015

Documento maquetado con TEXIS v.1.0.

Machine learning based methods for the study of metabolism and its effect on the behavior of biological systems

Memoria que presenta para optar al título de
Doctor en Informática
A thesis submitted in partial fulfillment for the degree of Doctor in Computer Science

Clara Higuera Cabañes

Dirigida por los Doctores
Supervised by

Gonzalo Pajares Martinsanz
Federico Morán Abad

Departamento de Ingeniería del Software e Inteligencia Artificial
Facultad de Informática
Universidad Complutense de Madrid

Madrid, 2015

A mi familia
y a todas las personas que han creído en mí.

Nothing was made of big which wasn't an exaggerated hope.
Jules Verne

Agradecimientos

Explica Rosa Montero, precisamente en un libro sobre la vida de la asombrosa Marie Curie, que el paisaje que atisba cuando empieza a escribir una novela se asemeja a un largo collar de oscuridad iluminado de cuando en cuando por una gruesa perla iridiscente. De esta forma, va avanzando con esfuerzo por el hilo de sombras de una cuenta a la otra hasta llegar a la escena final, que es para ella la última de esas islas de luz, una explosión radiante. Para mí el recorrido de esta tesis doctoral ha sido increíblemente similar a ese proceso de creación que describe Rosa Montero y he de decir que a lo largo de estos cuatro años ha habido muchas personas que me han ayudado a encontrar o al menos vislumbrar la siguiente perla iridiscente de mi camino. Cada una a su manera pero que sin ellas seguro no habría llegado a la última, que es a la vez una perla intermedia en el largo collar que aún me queda por recorrer.

En primer lugar debo darle las gracias a mis directores Gonzalo Pajares y Federico Morán. A Fede por su confianza en mí, pues fue él quien me animó a embarcarme en esta aventura y me insistió en uno de sus primeros e-mails en que si aceptaba el reto no me arrepentiría. Esas palabras se me quedaron muy grabadas y cuatro años después puedo decir que tenía razón. Además, es una de las personas de las que más he aprendido, tanto en el ámbito académico como en el personal. A Gonzalo por su enorme implicación, esfuerzo y por transmitirme tanta tranquilidad en momentos difíciles. Con él también he aprendido que con esfuerzo, creatividad y sosiego se puede conseguir cualquier cosa en la vida. Una vez se asimila esa lección, tu actitud ante la vida cambia y todo es posible.

También quiero darle las gracias a Francisco Montero (Paco), primero por empujarme al mundo de la docencia, permitiéndome impartir algunas de sus clases, así como codirigir con él proyectos fin de grado. Gracias a ello he podido descubrir la enriquecedora experiencia de aprender enseñando, desconocida hasta entonces para mí y con la que he obtenido mucha satisfacción personal y profesional. También le agradezco haber compartido conmigo y mis compañeros toda su cultura literaria, cinematográfica, histórica, folclórica, geográfica culinaria y un largo etc. Creo que todos los que hemos pasado por su laboratorio unos años somos un poquito más cultos gracias a él.

Mis compañeros de laboratorio también han sido un gran apoyo estos años: Sara, Gabriel, Miguel, Jorge, que tanto me ha ayudado discutiendo sobre ratones y proteínas y cuya lucidez asombra. Héctor, que siempre ha ido unos años por delante iluminando el camino, entendiéndome y aconsejándome en muchas ocasiones. Arturo, ejemplo de constancia y persistencia, siempre dispuesto a levantarte el ánimo con palabras halagadoras. Daniela, incansable luchadora, siempre a mi lado repitiéndome: "¡lo que no te mata te hace más fuerte, sigue peleando!" y Laura que nos cuida a todos con un cariño y alegría que parece no acabársele nunca.

También me gustaría darle las gracias a Sole, que conozco casi desde que empecé pero con la que he tenido la oportunidad de trabajar más de cerca en los últimos meses. Siempre me ha hecho sentir muy cómoda y de ella he aprendido mucho, tanto sobre redes auto-organizativas como de su manera de trabajar rigurosa y disciplinada. Espero que en el futuro podamos continuar los proyectos que nos quedaron pendientes.

Me gustaría también mencionar a María y darle las gracias por esos soplos de aire fresco en los últimos momentos, por sus ánimos, por reforzar la confianza en mí y por su vitalidad contagiosa.

Igualmente debo un agradecimiento también a Lukas Käll por acogerme en sus grupo de investigación en el Science for Life Laboratory en Estocolmo, lo que me permitió conocer otras maneras de trabajar y de hacer ciencia. A Matilde Santos por ponerme en contacto con Krzysztof Cios, quien también me acogió en su grupo durante tres meses en la Virginia Commonwealth University. A este último por brindarme la oportunidad de vivir la experiencia americana, dedicarme tiempo y exprimirme para sacar lo mejor de mí. A mis compañeros tanto de Estocolmo como de Richmond que me ayudaron a sentirme como en casa aun estando tan lejos: Stefania, Hugo, Jose, Yue Hu, Xiao, Lumi, Laura, Cissi, Ljiljana, Shaun, Beata, Joseph, Robin, Janus, Silvia Banocy etc.

Sin lugar a dudas y como en la vida de todo doctorando, mi familia ha sido un apoyo imprescindible. Mis padres, Evelio y Cristina y mi hermana Lucía (Chim) han estado siempre a mi lado inyectándome una increíble energía vital en todo momento que he necesitado y en cualquier lugar del mundo en el que me encontrara. Sus esfuerzos por aconsejarme y asesorarme, su inmenso cariño y su filosofía de vida me han guiado y me seguirán guiando siempre para alcanzar cualquier objetivo que me proponga. A Miguel, con quien comparto el frenético modo de vida de no parar nunca de hacer cosas ni de dejar de crecer, le agradezco su paciencia y su amor en este viaje y en tantos otros en los que me ha acompañado desde hace trece años. También por esperarme siempre durante mis ausencias en el extranjero y por saber ayudarme a cambiar de perspectiva en esos momentos en los que me atasco bien cantándome su última canción, llevándome al cine o hablándome de sus miles de proyectos.

Kris, que no sé bien donde colocarla, si más cerca de los compañeros de trabajo o de los compañeros de vida porque pertenece a los dos. Indudablemente mi gemela en esto de la bioinformática. Desde que nos conocimos hace cinco años en el máster y comenzamos las dos nuestras respectivas tesis hemos recorrido el camino juntas. No tengo palabras para describir lo importante que ha sido para mí tenerla cerca, siempre muy cerca, tanto en momentos de confusión, de agotamiento o de indecisión como en momentos de alegría, de éxito y de ocio. No importa cual sea nuestro estado al saludarnos que al despedirnos siempre lo hacemos con escandalosas carcajadas y largos y apretados abrazos, ya sea en Madrid o en el Círculo Polar Ártico.

No puedo acabar sin agradecer a todos mis amigos no relacionados con el mundo de la ciencia que también han tenido un papel importante estos años y que me transmiten cada día tanto afecto y confianza: Xonita, Nachino, Isobel, Mar, Vera, a quien estaré eternamente agradecida porque con su don especial me ayudó a superar un bache que a mí me parecía un 8000. Cris, que también me echa un cable o veinte con sus terapias inventadas. También Choco, Dave, Chalo, Aarón, Bea, Carmen B., Guglielmo, Carmen L., Rodrigo, Mazor, Pablo y todos los grupos que forman y formarán: Incendios, AT, Paracaídas, Autumn Commets, Karen Koltrane... Sin duda les debo mil gracias a Alvarito y a Helios por cuidarme como hermanos en tierras americanas y por hacerme sentir *the authentic princess of Minor Manor*. Mis amigos de Alemania: Lucía, Dan, Pablo, Chloé y la estrella luminosa: Nefeli, *ein Name eine Geschichte*. Por último también quiero darles las gracias a mis amigos de Burgos que siempre hacen que sea un gusto volver a casa: Pili, Vero, Cris, Dani, Borjita, Tito, Gus, Natalia, Blanca, Rubén, Álvaro, Carlos, Alberto y Gonzalo. A este último le debo además una mención especial por su ayuda en el diseño de la portada de esta tesis.

A todas estas personas y muchas más que no he podido mencionar en este espacio les agradezco enormemente su existencia por haber contribuido en mayor o menor medida a hacer este recorrido más ameno y más feliz.

Abstract

The disciplines of bioinformatics and computational biology, that make use of computational methods to solve biological problems, have turned out to be indispensable to conduct a successful research in the life sciences in general and in molecular biology in particular. On one hand, the big amount of complex information generated in the laboratories is no longer possible to be processed, analyzed and visualized by the human eye. On the other hand, certain biological processes are difficult or can not be reproduced in the lab. This problem makes the in silico simulation the only way to study them. For these reasons and because new complex problems arise everyday, the development of new computational methods that assist in the research in molecular biology has become essential.

One of the main subjects in biology is the study of metabolism, which can be defined as an intricate network of chemical reactions that occur inside the cell and by means of which organisms are maintained alive. Precisely, the application and development of novel computational methods, based on machine learning, for the study of metabolism and its effect on biological systems behavior constitutes the central research subject of this thesis. In it we propose computational solutions to solve three specific problems that comprise a dynamic, a structural and a functional study of metabolism. The first work, that corresponds to a dynamic study, explores the regulation at enzymatic level of a metabolic cycle by means of global optimization methods and specifically by the application of multi-criteria optimization. A set of parameters responsible for the dynamic behavior of the system is optimized with the goal of finding a universal pattern of regulation for the system studied. The second work constitutes a structural study with the aim of clustering a complex set of prokaryotic species by their similarity in certain metabolic features. To that end an Expert System adapted to the nature of the data an based on the combination of unsupervised classification methods (Self Organizing Maps, SOM) and clustering validity indices was designed. The system also allows the extraction of underlying information in data imperceptible to the naked eye such as the relation between metabolism and environment. The third work corresponds to a functional study, where a new data mining approach, also based on SOM and combined with a statis-

tical test, is designed. The approach analyzes protein expression data and identifies sets of proteins involved in high-level functional activities such as learning and memory in control and Down syndrome mice. The technique proposed represents a novel way of analyzing protein expression data, which are at the same time the result of the regulation of metabolic networks at the level of expression.

In conclusion, this thesis constitutes an original and multidisciplinary research work in which by means of novel machine learning based methods three problems related to the study of metabolism are solved. The design, development and application of the methodology are based on the fields of artificial intelligence and machine learning while the results obtained through them possess importance and represent advances in the field of molecular biology.

Resumen

Las disciplinas de bioinformática y biología computacional, que se sirven de técnicas informáticas para dar solución a problemas en biología, se han posicionado como piezas clave en la investigación en biología molecular. Tanto por la gran cantidad de información compleja generada en los laboratorios como por la necesidad de simular in silico determinados procesos biológicos para su estudio, actualmente es esencial el desarrollo de nuevos métodos computacionales que asistan en la investigación en biología. Uno de los temas centrales en biología molecular es el estudio del metabolismo, que se define como una red intrincada de reacciones químicas que ocurren dentro de la célula y por medio de las cuales los organismos se mantienen vivos. Precisamente el estudio mediante técnicas computacionales basadas en aprendizaje automático del metabolismo y su efecto en el comportamiento de sistemas biológicos constituye el tema central del presente trabajo de investigación. En él se proponen soluciones computacionales para resolver tres problemas biológicos concretos que comprenden un estudio dinámico, un estudio estructural y otro funcional del metabolismo. El primer trabajo que se corresponde con un estudio dinámico estudia la regulación a nivel enzimático de un modelo de un ciclo metabólico mediante la aplicación novedosa de métodos de optimización, con especial hincapié en la optimización multi-objetivo y con el objetivo de encontrar un esquema de regulación universal para el modelo de estudio. El segundo trabajo se corresponde con un estudio estructural en el que el objetivo es agrupar un conjunto de especies bacterianas por similitud en determinadas características metabólicas. Para ello, se diseñó un sistema experto basado en la combinación de técnicas de clasificación no supervisada con índices de validación. El sistema permite también la extracción de información inapreciable a simple vista, como puede ser la relación entre metabolismo y ambiente. El tercer trabajo consiste en un estudio funcional. En él se desarrolla una nueva estrategia de minería de datos basada también en técnicas de clasificación no supervisada, esta vez combinada con un test estadístico. La estrategia permite identificar proteínas involucradas en actividades funcionales de alto nivel como el aprendizaje y la memoria y constituye una manera novedosa de tratar datos de expresión de proteínas que son a la vez el resultado de la regulación de redes metabólicas a nivel de expresión. En definitiva la presente tesis constituye un trabajo

de investigación multidisciplinar original e innovador en el que se resuelven tres problemas biológicos concretos mediante técnicas novedosas basadas en aprendizaje automático. El diseño, desarrollo y aplicación de estas técnicas tiene relevancia en el campo de la inteligencia artificial y el aprendizaje automático mientras que los resultados obtenidos mediante ellas tienen relevancia y suponen avances en el campo de la biología molecular.

Contents

Index of Figures

Index of Tables

Chapter 1

Introduction

1.1. Historical Review

Since decades computer science has played an important role in the advances in research in life sciences in general and molecular biology in particular. In the early 60s, computers started to be an available resource for researchers in the academic world (Hagen, 2000). The appearance in 1957 of the first high-level programming language, FORTRAN, specially appropriated for scientific applications and relatively easy to learn, in comparison to machine languages of that time, favored the development of what will be later called computational biology. Works such as the ones of Margaret Oakley Dayhoff (Dayhoff and Ledley, 1962; Dayhoff, 1965, 1969; Eck and Dayhoff, 1966) or Walter Fitch (Fitch, 1966; Fitch and Margoliash, 1967) laid the foundations of this new discipline. Even some of their techniques or updated versions are still used nowadays. Dayhoff, whose work has been reviewed in (Hunt, 1983; Strasser, 2010), wrote programs to aid to determine the primary structure of small proteins (Dayhoff and Ledley, 1962; Dayhoff, 1965) in minutes in contrast to other traditional methods that used to take months. She also created the first database in Molecular Biology creating a library of sequences of aminoacids from proteins known at that time, the Atlas of Protein Sequence and Structure (Dayhoff et al., 1965), that began to be used in studies of comparative biochemistry or molecular evolution. That work turned out to be a starting point for many other computational biologists that began to create their own databases. At that time the aid of computers was of great importance, specially because it was possible to automate some tasks and compute them faster than manually. Fitch (1966) designed a method that searched for nonrandom alignments by comparing two sequences of protein molecules and calculated the mutations required to transform one sequence into the other one. His method was later improved by Saul Needlman and Christian Wunsch and is currently one of the standard methods for sequence alignment (Needleman and Wunsch, 1970).

It also stimulated later scientists to develop more refined methods. In 1970 of last century already many computational biologists had developed several computational techniques useful for analyzing molecular structure, or studies related to evolution.

During the early 70s scientists Paulin Hogeweg and Ben Hesper, experts on theoretical biology, thought there should exist a discipline that studied how living organisms gathered, processed and used information. They knew that one of the properties that defined life was the processing of information in various forms (e.g. accumulation of information throughout evolution or transmission of information in DNA to intra and intercellular processes (Hogeweg, 2011) so the same way disciplines like biophysics or biochemistry existed, they started to use the term bioinformatics to denominate a new research field. Although this term changed with the time, and specially after the boom of massive sequencing, they originally denominated it as the study of the informatics processes in biotic systems (Hesper and Hogeweg, 1970; Hogeweg and Hesper, 1978; Hogeweg, 1978). They also thought that this new field should combine pattern analysis and dynamic modeling of biological systems by means of computational techniques. Firstly, they were interested in analyzing patterns of variation at different levels. Secondly, they wanted to detect emergent phenomena in models and compare the results with real data. Thirdly, they believed that the relation between genotype, phenotype, behavior and environment could be studied by searching patterns and their transformation and the understanding of these processes formed the core of the investigation in bioinformatics (Hogeweg, 2011). Hogeweg, Hesper and other contemporary authors began to use and develop pattern recognition techniques (Lance and Williams, 1966; Macnaughton-Smith et al., 1964) as well as clustering methods (Hogeweg, 1976) in their research. Also cell automats as a formalism of modeling in ecology (Hogeweg, 1988; Boerlijst and Hogeweg, 1991) were introduced. Event-based and individual-oriented simulation methods as well were developed, currently called agent based methods. All these techniques and many more used in these first years belong to the artificial intelligence (AI) paradigm, what allows us to establish the implications that this field has had and still has in molecular biology.

Nevertheless, not only biology has obtained a benefit from AI, but the research in AI started in the 60s to find inspiration in biology as well in order to design new representations of processing information systems. Models based on neural networks and the brain functioning for pattern recognition (Rosenblatt, 1962) started to appear. Also genetic and evolutionary algorithms, that were firstly designed to simulate evolution and natural selection (Goldberg, 1989; Holland, 1992; Rechenberg, 1973; Schwefel, 1977), ended up being useful and powerful tools for solving optimization problems (Crosby et al., 1973; Fraser and Burnell, 1970; Fraser, 1957).

During the 80s and 90s more novel methods based on AI were developed

to solve different biological problems. However, it was from year 2000 on that these methods acquired special importance. That year, the first complete draft of the human genome [1] was published. This event was considered an unprecedented milestone in the already consolidated molecular biology. From that moment genome sequencing technologies began to evolve very rapidly and now the current ones allow the sequencing of a complete genome from an individual in hours. Other current high throughput techniques provide different kind of information, for instance techniques for transcriptomics analysis not only give information about the genes present in a cell but also about which ones are expressed in a certain moment of time. Also proteomics and metabolomics can determine which proteins and which metabolites are present respectively. Finally, fluxomics deduces fluxes of transformation of one metabolite into others. All these techniques are denominated "omics" and allow scientists to have a complete insight of a certain cell, tissue or organism. Systems biology and integrative biology are fields that try to integrate all this information. Figure 1.1 depicts the variety of the omics high-throughput techniques that generate data and the set of mathematical and computational techniques needed to analyze and combine the data fundamental for the research in systems biology. For the first time we are close to be able to determine the phenotype or behavior of an organism from its genotype under particular conditions.

One of the biggest problems lies on the treatment and analysis of the huge amount of data generated from these techniques. This fact makes very difficult to process and integrate all the information. At this moment the field of biology has joined the group of "Big Data". Figure 1.2 shows the ratio of model predictability according to data size in comparison to other disciplines. In molecular biology the ratio is not accordingly correlated with data size. Precisely bioinformatics has turned out to be a key to conduct a successful research in molecular biology. In order to be able to process and interpret complex and varied data generated in the laboratory it is necessary to develop new computational techniques capable of analyzing all these information.

In the past few years a great number of methods based on artificial intelligence and specifically in machine learning have been developed with the main goal of building useful tools that, whether provide aid in the analysis, extraction and interpretation of biological data or are capable of simulating biological processes and building models that facilitate the understanding of such processes. Larranaga et al. (2006) make an exhaustive review about machine learning methods used in bioinformatics and divide the main biological problems in two types: modeling and optimization problems.

In the first type, the learning process consists of running a program that induces the construction of a model based on a training dataset in order to

[1] The genome is the genetic material of an organism. It includes genes and non-coding sequences of DNA.

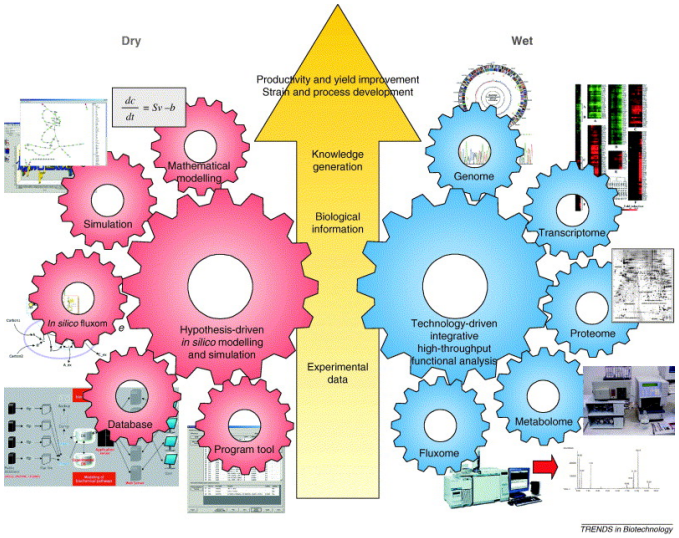

Figure 1.1: **In silico (Dry) and experimental (Wet) techniques for the research in systems biology.** In blue: High-throughput techniques that provide amounts of data (genome transcriptome, proteome, metabolome and fluxome data). In red: Mathematical and computational techniques needed for analyzing and integrating the data (Lee et al., 2005).

later on infer information from it. To this type belong mainly classification problems where in some cases work with a labeled training dataset of samples from which the correct class is known (supervised classification) and it is used to build a generalized model that allows afterwards the classification of similar new unlabeled data. In other cases, the classes to which the samples belong is unknown (unsupervised classification) and the main goal is to identify interesting patterns in data. Examples of the different kinds of methods in classification revised in (Pajares and de la Cruz, 2010) can be seen in Figure 1.3

Some examples of application of supervised classification methods in bioinformatics are (Bao and Cui, 2005; Carter et al., 2001; Cypess et al., 2013; Jagga and Gupta, 2014; Kim, 2004; López-Bigas and Ouzounis, 2004; Salamov and Solovyev, 1995; Sø rlie et al., 2001; Yi and Lander, 1993). In these works neural networks, support vector machines and hierarchical clustering algorithms are used mainly for the identification of functional genes related to genetic diseases. Also the strategy of the nearest neighbor has been used in protein structure prediction.

Unsupervised classification methods, also called clustering methods, have been mainly used in bioinformatics for the analysis of gene expression data with the aim of finding a relation between genes with a similar expression profile and some kind of functional or regulation similarity. Some examples

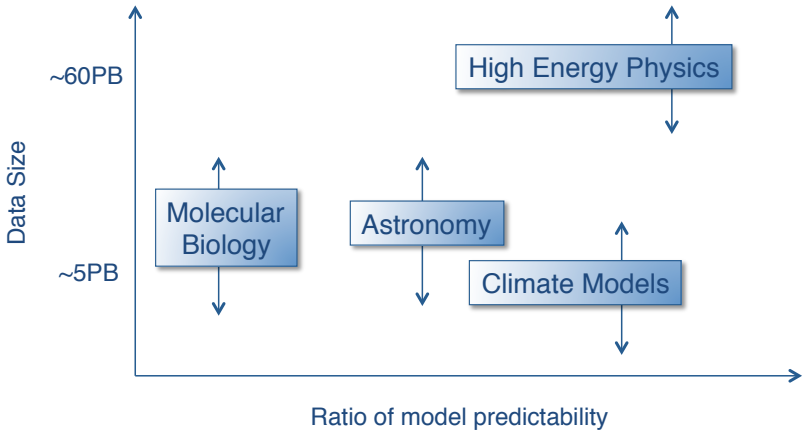

Figure 1.2: **Ratio of model predictability with data size in different fields.** Adaptation from Dr. A. Valencia´s oral communication.

are (Spellman et al., 1998; Tamayo et al., 1999; Herrero et al., 2001; Bohlin et al., 2009; Sheng et al., 2003; Brohée and van Helden, 2006; Lorenzo-Redondo et al., 2014).

The second type of problems (optimization) are the ones that require finding an optimal solution within a space of multiple possible solutions. There exist two main types of methods to solve these problems: local and global optimization methods. The first ones, as for example gradient-based methods, are useful when the solution space does not contain many local optima. They are able to find a solution in the region close to the starting point, what makes them very dependent on the initial values of the algorithm and prone to get stuck in local optima. In biology many problems are nonlinear, multimodal and often NP-complete. For this type of complex problems global optimization methods are advantageous because they are able to explore a wider solution space, escaping from local optima. These methods can also be divided into deterministic and stochastic. Stochastic global optimization methods are based on probabilistic strategies and have therefore certain random component. Because of this, they can not offer an absolute guarantee of having found the global optimum. Although deterministic methods can guarantee global optimality in certain problems none of them can solve any kind of problem in finite time. The computational cost of applying these algorithms increases very fast (often exponentially) with the problem size. Stochastic methods are capable of at least locate the vicinity of the global optimum in a reasonable time but can not guarantee the global optimality of the solution. Examples of global stochastic optimization methods applied in bioinformatics are genetic algorithms (GAs), evolutionary

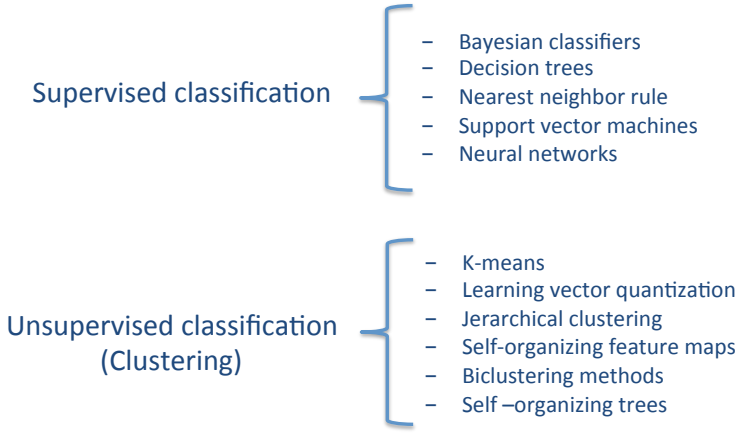

Figure 1.3: **Examples of supervised and unsupervised classification methods.**

programming (EP), simulated annealing (SA), which was first invented by (Kirkpatrick et al., 1983) and extended in (van Laarhoven and Aarts, 1987) and optimization methods based on ant colonies (Colorni et al., 1991). These kind of methods have been successfully used to solve several biological problems, for example GAs were used in the problem of multiple sequence alignment (Lee et al., 2008; Nguyen et al., 2002) and the study of protein folding in simplified models (Krasnogor et al., 2002; Smith, 2004). Also they have been very important in the modeling of genetic networks (Kikuchi et al., 2003) and the estimation of critical parameters in bioprocesses (Park et al., 1997). A recent review of GAs, SA and EP on this later field can be found in (Sun et al., 2012). Particularly, SA has been relevant in the identification of consensus sequences between several sequences of DNA (Keith et al., 2002) or in order to align experimental transcription profiles with a set of reference experiments (Wren et al., 2004). Lastly, some examples of problems in which EP has been applied in biology are the discovery of structural elements in RNA (Fogel et al., 2002), clustering of microarray data (Falkenauer and Marchand, 2002) or the estimation of parameters in metabolic networks (Rodriguez-Fernandez et al., 2006). In (Moles et al., 2003) authors review and compare different global optimization methods applied to this last type of problems.

It should also be mentioned a last type of optimization strategy that has started recently to be used in biological problems, and that has been used for many years in other fields; multi-criteria optimization. It consists of simultaneously optimize two or more objective functions. This type of

optimization, instead of obtaining a unique solution, finds a set of solutions called Pareto front of optimal solutions, in such a way that it is not possible to improve any of them in one objective without worsening in another objective. Each solution represents a trade-off solution between the different objectives and a decision making process is necessary to chose a solution as the optimal candidate. The Non-dominated Sorting Genetic Algorithm-II (NSGA-II) (Deb et al., 2002) based on evolutionary programming has been widely used in many fields, however other authors have implemented new strategies with successful results (Sendín et al., 2009). (Handl and Knowles, 2007) also published a revision of the application of multi-objective optimization (MOO) in bioinformatics and computational biology. An example of application of MOO strategies in these fields are: (Cutello et al., 2006; Day et al., 2002; Lanning et al., 2000), for the prediction of protein structure or the study of optimality in biochemical processes (Andrés-Toro et al., 2004; Halsall-Whitney et al., 2003; Ierapetritou et al., 2004; Sendín et al., 2010).

Taking the aforementioned into account, it is clear the relevance of machine learning methods in the advances in molecular biology. In the previous review several biological problems faced with machine learning approaches have been discussed. However, as briefly mentioned before, inside the wide field of molecular biology there exist many research areas like genomics, proteomics, metabolomics or fluxomics. From the information generated in many of those research fields the determination of metabolism is conducted. In essence metabolism is the phenotypic expression of genotype (genomics, proteomics, metabolomics etc.) and it is of great importance because it is present at very different biological levels, from the transformation of substrates into energy or products necessary for the cell to the adaptation of species to certain environments or the development of learning deficits in patients with intellectual disorders. Precisely metabolism, and its effect on the behavior of biological systems, constitutes the main research subject of this thesis. By means of the application, design and development of novel computational methods based on machine learning techniques, three biological problems in which metabolism is involved will be solved.

1.2. Identification of the research problem

Metabolism consists of an intricate network of chemical reactions that occur inside the cells and by means of which organisms are maintained alive. In general terms, metabolism has the role of whether carrying out the transformations from an initial product like food into end products like energy, or molecules to form new structures (membrane lipids, proteins, genetic material like DNA, etc.) or degrade biomolecules (catabolism). Each of these transformations is called “metabolic pathway”, but at the same time metabolic pathways are connected to each other by reactions that transform one

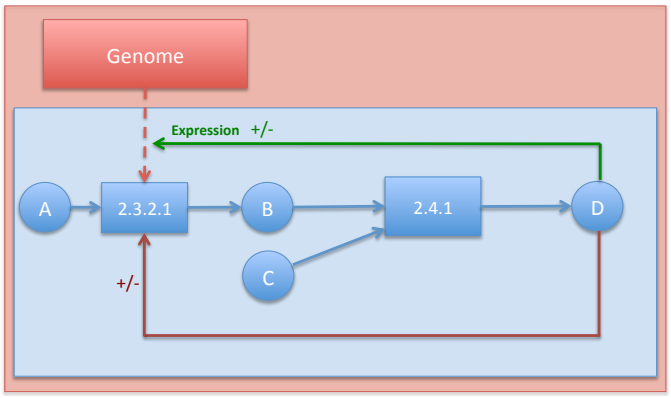

Figure 1.4: **Scheme of a metabolic network represented as a bipartite graph.** Alternate nodes (A, B, C, D) represent metabolites and intermediate nodes (2.3.2.1, 2.4.1) reactions

metabolite [2] of one pathway into a metabolite of a second one. This way, metabolism rather than a set of metabolic pathways can be conceived as a net of reactions, for this reason we will refer to it from now on as metabolic network. Alternatively and in a more exact and rigorous way, a metabolic network can be represented as a bipartite graph where alternate nodes represent metabolites and intermediate nodes represent reactions that transform one metabolite into another one. These reactions are mediated or catalyzed by proteins called enzymes that are activated or inhibited according to the cell needs.

Figure 1.4 represents a metabolic network where circles A, B, C and D correspond to metabolites (substrates and products of reactions) and squares correspond to reactions catalyzed by the corresponding enzymes. The result of the expression of the genome is, among other things, a set of proteins (indicated in the figure inside the square) that have a certain catalytic activity (enzymes). These enzymes are associated with a number according to the reaction that each one is able to catalyze. These activities relate certain metabolites with others: for instance enzyme 2.3.2.1 is capable of catalyzing the conversion of metabolite A into B and enzyme 2.4.1 catalyzes B and C into D. As a result it appears a metabolic network that can be represented as a bipartite graph where alternative nodes (A, B, C and D) are metabolites and intermediate ones (2.3.2.1 and 2.4.1) reactions. Over the metabolic network represented inside the blue background square a regulation network is superimposed (red square). This way, a metabolite (for example D) can exert a regulation over reaction 2.3.2.1. Such regulation can take place at

[2] A metabolite is any kind of molecule that participates whether as substrate or product in a metabolic reaction.

two levels: at the level of enzymatic activity, acting directly on one enzyme (brown arrow) forcing it to increase or decrease the production of one metabolite (this case B) or regulation at the expression level inhibiting (-) or activating (+) the formation of a protein that has certain enzymatic activity (green arrow). This is performed by inducing or repressing the expression of one or several specific genes that are functionally related. The result of gene expression is the production of functional proteins that carry out a specific activity.

Currently, thanks to the sequencing techniques, complete genomes of a big amount of organisms are available. However, in order to really be able to understand these organisms it is necessary to know the patterns of activation of individual genes. Specifically, it is necessary to be able to know and predict at which level, in which moment and under which circumstances specific genes of certain organisms are expressed. The mechanism under these gene expression patterns consists of a complex interaction of DNA, RNA, proteins, metabolites that from an abstract point of view can be seen as an interaction network of inhibition or activation of genes. A gene regulation network (Handl et al., 2005).

At the time of studying metabolism and its influence in the behavior of a cell or an organism and taking into account the scheme previously explained there exist several aspects or perspectives:

1. Analysis of the dynamic behavior of metabolic networks
2. Study of the structure of the network and its relation with phenotypic features or environmental preferences
3. Identification of specific expression patterns related to certain metabolic or functional activities or behavior.

Through the application and development of novel computational methods based on machine learning, the work presented in this thesis tries to give answers to three different kinds of problems inside the three perspectives previously listed. The methods used are mainly focused on optimization and unsupervised learning techniques.

The first work corresponds to a dynamic study of metabolism. In it, by means of global optimization methods we explore the regulation at enzymatic level that metabolites have over its own metabolic network, as described in Section 1.2. Under the hypothesis that metabolism and metabolic pathways have undergone a process of optimization through time, in this work we study the regulation of a metabolic cycle [3] using a novel approach in this kind of studies: multi-criteria optimization. Using this kind of optimization we estimate and optimize the values of a set of parameters that regulate

[3] A metabolic cycle is a metabolic network where the end product of the reaction is also the substrate that starts the pathway

and determine the dynamic behavior of the metabolic cycle. Depending on the value of these parameters the direction of the flux[4] is directed in one direction or another in order to meet cell needs under different conditions. Consequently the evolution of concentration of metabolites through time will vary. The main goal of using multi-criteria optimization to this problem is the finding of a universal scheme of regulation (an optimal set of parameters) that allows the system to behave optimally under different varying conditions, which is not possible to find by means of mono-objective optimization.

The second work corresponds to a structural study, where an expert system is developed. The goal of the system is to cluster a complex dataset of bacterial species by their similarity in the structure of their metabolic network (metabolism), specifically by the similarity in the absence or presence of a set of metabolic pathways. The task of applying clustering techniques to biological data is generally very complex for three main reasons: 1) the existing number of classes is in many cases unknown 2) the absence of methods to validate the resulting partition specially from the biological point of view and 3) frequently biological data have a complex nature or are incomplete. The system proposed combines a clustering method (Self Organizing Maps, SOM) and clustering validity indices in a hierarchical strategy to face these problems. Its first goal is therefore adequately clustering the different species in metabolically similar groups. The second goal consists of identifying common phenotypic characteristics or environmental preferences among the resulting groups. The purpose is to search for underlying information in data that may help to relate metabolism with certain behaviors, as can be the adaptation of species to the environment. The results of this work can be of great help in the understanding of communities of bacterial species, which are responsible of many natural and artificial processes.

The third work corresponds to a functional study, in which a data mining approach, also based on unsupervised classification, is designed. The main goal is the identification of proteins involved in high-level functional activities such as learning and memory in control and Down Syndrome (DS) mice. The strategy proposed exploits the functionalities and advantages of Self Organizing Maps over other clustering methods providing a novel analysis on experimental protein expression data. It combines SOM with a statistical test and is designed to discover new informative patterns in this kind of data not possible to find with standard statistical analysis or other clustering methods.

Protein expression data provide information about which proteins are present in certain tissue. Because proteins are the consequence of the expression of specific genes, as previously explained, in this work we explore the second type of regulation present in metabolism: regulation at expression level. Depending on the cells or organism needs and the tasks or activities

[4]The flux is the rate of turnover of molecules through a metabolic pathway.

that they develop at each time, the proteins involved in such tasks are less or more expressed. The proposed data mining approach firstly determines if the protein expression information can be used to cluster individuals in groups according to their response to learning. In our case case control (healthy) mice that learn normally, DS mice unable to learn and DS mice that recover the learning ability after having been injected with a drug. Secondly, the method identifies reduced subsets of proteins that best discriminate between classes of mice and therefore define changes in the level of expression due to genetic or treatment causes. The identification of these proteins has great relevance in the field because it can help to recognize which proteins should be altered by drugs in order to decrease the learning and memory deficits in Down syndrome patients.

Taking all the aforementioned into account, it is clear that the current research is focused on bioinformatics and particularly in the machine learning field and its application to biological problems. The development of new methods and the application of classical methods that as far as we know have never been used in the problems in question constitute the main contribution of the research. The results obtained in this research work have implications in the field of molecular biology, whereas the methodology used has relevance in computer science research. The result is therefore a clearly multidisciplinary project. Figure 1.5 shows a scheme of the thesis presented. In it are displayed the different kind of studies performed, the specific biological problems faced and the machine learning approaches proposed.

1.3. Motivation

Firstly, one of the personal motivations at the time of accomplishing this thesis, halfway between molecular biology and artificial intelligence, was the challenge of building a piece that belonged to the machinery that joins these two disciplines so different one another but at the same time capable of complementing each other so much. AI is a field that advances at a very fast pace and its applications to other disciplines are frequent. However, many times these advances remain unknown to other fields where they would be very useful. Thanks to the relatively recent disciplines of bioinformatics and computational biology this disconnection between biology and AI is being partially solved. Nevertheless it is necessary to train professionals with the technical knowledge and proper skills to understand the biological problems and at the same time capable of deciding which computational methodology is the most adequate to solve them. Currently, in these disciplines professionals with very different backgrounds such as computer scientists, mathematicians, statisticians, physicists, biologists, biochemists, physicians each of them experts in their own fields work together. For this reason professionals who work in bioinformatics or computational biology should possess an open

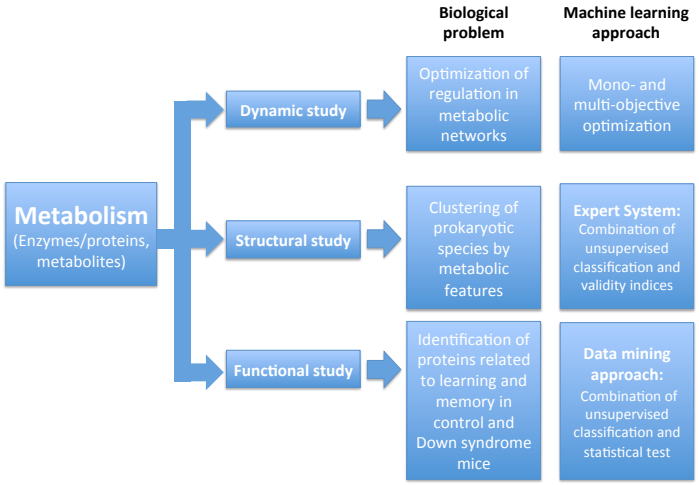

Figure 1.5: **Scheme of the presented thesis.** Box on the left-hand side represents the case of study: metabolism. Second column displays the three different perspectives from which metabolism is studied: dynamic, structural and functional. Third column specifies the name of each biological problem faced within each perspective and fourth column displays the machine learning approach proposed to solve each problem.

and transdisciplinar mentality in order to be able to establish collaborations between different research groups from different disciplines and set the proper goals for the common benefit. Because of these reasons, it is patent that one of the main motivations of the work presented in this thesis is to establish a common nexus between molecular biology and artificial intelligence with a clear benefit for the scientific communities in both fields.

Secondly, the research has also been motivated and possible thanks to a FPI scholarship for doctoral studies from the Spanish Ministry of Economy and Competitiveness granted to the biophysics group from the department of Biochemistry and Molecular Biology I in the School of Chemistry of the University Complutense of Madrid. The scholarship was associated to the project BFU2009-12895-C02-02 named: *A Systems Biology approach to bacterial interactions in insects: genomic analysis, functional and evolutionary studies, and constraint-based modeling.* One important part of the project consisted of the study of the genome and metabolism of bacterial species being of special relevance the implementation of new computational techniques for its analysis. This latter fact fostered the collaboration with the ISCAR group (from its initials in Spanish of Systems Engineering, Control, Automation and Robotics) formed by members from the department of Software Engineering and Artificial Intelligence and Architecture of Computers

and Automatics from the School of Informatics of the same University. ISCAR group has recognized expertise in computational methods based on machine learning. This way, the present thesis is defined as a coordinated multidisciplinary project in which the expertise of each group is exploited with one main goal: The resolution of novel problems within the field of molecular biology and specifically within the study of metabolism by means of the development and application of also novel techniques within the field of applied machine learning. The main contribution of the work presented in this thesis to the BFU project consisted of the design and application of such methodology for the study of metabolism in bacterial species, which can additionally help to associate genotypic characteristics to phenotypic behaviors.

Thirdly, during the four years of doctoral studies, two stays abroad in foreign research groups as visiting PhD student were made, both funded by the FPI scholarship from the Ministry of Economy and Competitiveness and individually approved by the ANECA. The first stay took place in the Science for Life Laboratory in Stockholm, Sweden under the supervision of Dr. Lukas Käll. The second one took place in the Virginia Commonwealth University, Richmond, Virginia, USA in the department of Computer Science and under the supervision of Dr. Krzsytof Cios within the group of Biomedical Informatics. This last stay motivated the third work presented in this thesis, which is at the same time a collaboration with the research group of the biologist Dr. Katheleen Gardiner from the Linda Crnic Institute for Down Syndrome and the University of Colorado, School of Medicine, USA.

1.4. Objectives

The objectives proposed during the research are based on the problems previously described, to which a set of solutions by means of computational approaches have been proposed. The general objective of this thesis consists of the study, application and development of computational methods based on machine learning for the study of metabolism and its effect on the behavior of biological systems. The specific objectives are the following:

1. Finding an optimal enzymatic regulation pattern of a metabolic network by means of optimization methods.

2. Development of an expert system based on unsupervised classification for clustering bacterial species by metabolic features.

3. Development of a data mining approach based on unsupervised classification for the analysis of experimental protein expression data from control and Down Syndrome mice.

1.5. Contributions

The main contributions and results of the research are gathered in the following publications:

1.5.1. Publications in journals

1. Higuera C. Villaverde AF. Banga JR. Ross J. Morán F. Multi-Criteria Optimization of Regulation in Metabolic Networks. PLoS ONE (2012) 7(7): e41122. doi:10.1371/journal.pone.0041122.

   In this contribution, a multi-criteria approach has been used to optimize parameters for the allosteric regulation of enzymes in a model of a metabolic substrate-cycle and to find a universal regulation scheme for the model studied. This has been carried out by calculating the Pareto set of optimal solutions according to two objectives in different "environments" (specific time courses of end product concentrations). Using multi-criteria optimization we were able to calculate a consensus set of parameters that worked optimally in the different environments, which is an indication on the existence of a universal regulation mechanism for this system. Mono-objective optimization methods were also applied but no universal pattern was find. The implications from such a universal regulatory switch are discussed in the framework of large metabolic networks.

   The complete description of this work can be found in Chapter 2 of this thesis.

2. Higuera C., Pajares G., Tamames J., Morán F. (2013) Expert System for Clustering Prokaryotic Species by their Metabolic Features. Expert Systems with Applications (2013)

   In this work we propose an expert system (ES) to cluster a complex data set of 365 prokaryotic species by 114 metabolic features. The ES is inspired by the human expert reasoning and based on hierarchical strategies and the unsupervised classification method Self-Organizing Maps (SOM). It clusters the data in stages and makes use of a new validity index adapted to the dataset that identifies relevant clusters and monitors the process by using the well known validity index Davies Bouldin (DB). DB asses the validity of the partition obtained in each step. The results prove that the use of metabolic features combined with the ES is able to handle a complex dataset that can help in the extraction of underlying information, gaining advantage over other existing approaches, that may relate metabolism with phenotypic, environmental or evolutionary characteristics in prokaryotic species.

The complete description of this work can be found in Chapter 3 of this thesis.

1.5.2. Communications in conferences

The following works propose optimization and data mining methods to solve the three problems presented in this thesis. All of them constitute preliminary versions of the works published afterwards or submitted to international journals.

1. Higuera C., Villaverde AF., Banga J.R., Ross J., Morán F. Multicriteria Optimization of Regulation in Metabolic Networks. Conference on Research in Computational Molecular Biology (RECOMB) Systems Biology Conference, October 16th-17th, 2011, Barcelona, Spain.

   The communication in this conference described a preliminary version of the work described in Chapter 2 of this thesis.

2. Higuera C., Pajares G., Morán F., Tamames J. A Machine Learning Approach to explore the Correlation between Metabolism and Environment in Prokaryotes. XI Spanish Symposium on Bioinformatics, January 23rd-25th, 2012, Barcelona, Spain.

   The communication in this conference described a preliminary version of the work described in Chapter 3 of this thesis.

3. Higuera C., Gardiner J.K., Cios J.K. Self Organizing Maps based approach to identify protein patterns related to learning in control and mouse models for Down syndrome. XII Spanish Symposium on Bioinformatics, September 21st-24th, Seville, 2014, Spain

   In this communication the strategy and first results of the research described in Chapter 4 were presented.

1.6. Thesis layout

This thesis is structured in chapters and they are organized according to the natural evolution of the research. Their distribution has been organized as follows:

- Chapter 2 describes a dynamic study of metabolism by means of the application of optimization methods with a strong focus on multi-criteria optimization for the study of regulation in metabolic networks corresponding to a dynamic study of metabolism.
- Chapter 3 describes a structural study of metabolism by means of the design and development of an expert system designed for clustering prokaryotic species by their metabolic features.

- Chapter 4 describes a functional study of metabolism by the design and development of a data mining approach for the identification of critical proteins in learning and memory in control and Down syndrome mice.
- Chapter 5 sums up the general conclusions of the thesis and describes the future work.

The three problems faced in this thesis are slightly different from the biological and methodological perspective. For this reason chapters 2 ,3 and 4 describe the specific problematic, proposed computational methodology, results and conclusions of each problem. Chapter 5 summarizes the main conclusions of each of the three works that comprise this thesis and describes the general conclusions of the work as a whole.

Chapter 2

An approximation to the dynamic regulation of metabolic networks by means of multi-criteria optimization

2.1. Introduction

As explained in the previous chapter, one way of studying metabolism and metabolic networks is the study of the evolution of the concentration of metabolites throughout a period of time and more importantly the estimation of parameters that regulate their dynamic behavior. In this chapter we present the successful results published in (Higuera et al., 2012) where multi-criteria optimization was applied for the estimation of such parameters in a metabolic substrate-cycle with the goal of finding a universal pattern of dynamic regulation for the model studied. This study is considered therefore a dynamic study of metabolic networks by means of optimization methods. The particular application of multi-criteria optimization presents a novel approach in the field.

2.2. Metabolic pathways as a result of a natural evolutionary process

For decades the regulation of metabolic networks at genome scale and its mechanisms has been studied to further our understanding of this process, especially after the massive increase of sequencing data during the post-genomic era. Cell regulation can be accomplished through two complementary strategies. Genetic regulation (genetic circuits) occurs at genome level, controlling the expression of certain genes. This regulation affects the pre-

sence or absence of enzymes in the metabolic network. On the other hand, post-transcriptional regulation operates in two forms: RNA mediated regulation and the dynamic control of enzyme activities. The latter is achieved by the activation or inhibition of certain enzymes by means of controlling metabolites, as is the case with allosteric regulation.

The idea that the metabolic pathways and regulation strategies that take place inside a cell are the result of an evolutionary optimization process is widely accepted (Ebenhöh and Heinrich, 2001; Meléndez-Hevia et al., 1994). Optimality principles have also been used to explain the structure of genetic networks (Tkacik et al., 2009; Walczak et al., 2010). However, when it comes to defining the objective function that characterizes such evolutionary optimization, many uncertainties remain (Banga, 2008; Mendes and Kell, 1998; Nielsen, 2007; Schuetz et al., 2007). Depending on the case in question, different criteria must be satisfied. Generally, in studies concerning metabolic networks the most frequently chosen objective is the maximization of metabolic reaction rates, or steady-state-fluxes. However, other criteria such as the maximization of the concentration of metabolites (Goodacre, 2005; Sendín et al., 2010), enzymes, or other metabolic performances could be considered. A more realistic alternative is to take more than one criterion into account, an approach that may be closer to the way in which nature has acted in the evolutionary process of optimization. In this way multi-criteria optimization plays an important role since it considers the simultaneous optimization of several objectives. Multi-objective optimization has already been used in different biological contexts. Handl et al published in 2007 an exhaustive review (Handl and Knowles, 2007) about the application of multi-objective optimization in fields such as supervised and unsupervised classification of biological data, gene regulatory networks inference, sequence and structure alignment, protein structure prediction or optimization of biochemical processes among others. Several authors have performed preliminary research on the application of multi-objective optimization methods to reverse-engineering gene networks (Esmaeili and Jacob, 2009; Guo et al., 2009; Van Someren et al., 2003). More specifically, this kind of optimization has also been used to search patterns or unique optimal solutions. In (Shoval et al., 2012) the authors find that in different organisms the best-trade-off phenotypes were weighted averages of phenotypes specialized for single tasks. Furthermore Chubukov et al. (2012) found a pattern which relates the regulatory architecture of several yeast metabolic pathways to the gene expression response by searching a trade-off between two objectives: the cost of making a protein and the benefits of making it (its cellular function). This kind of works reveal that multi-objective optimization can, on the one hand contribute to find such patterns, and on the other hand to provide a closer approximation to natural evolutionary processes.

Unfortunately, finding a regulation design of a metabolic system as a

result of an optimization process is an NP-hard problem in the majority of cases (Banga, 2008; Goodacre, 2005). The complexity and non-linearity of metabolic systems make the task of obtaining global optima, in reasonable computational time, impossible in many cases. In these situations the so-called stochastic global optimization methods, such as genetic algorithms or simulating annealing among others, can at least locate a near globally optimal solution, although they do not offer a full guarantee that the global optimum has been achieved (Banga, 2008).

In their work Gilman and Ross (1995) proposed a genetic algorithm (GA) to optimize the parameters governing a post-transcriptional regulation model. Their model studied the dynamic regulation of allosteric enzymes and idealized an animal cell that metabolizes blood glucose for energy as long as the glucose concentration in the blood is adequate, but synthesizes glucose for export if the glucose concentration in the blood drops too low. The end goal of Gilman and Ross was to find a regulation pattern that could perform optimally in different time-varying courses of concentrations of glucose inside and outside the cell. However, after running the GA on different courses no global winner was found. Their work showed the presence of "generalist" solutions, which performed well on one or several courses, and "specialist solutions", which performed well on a single course but poorly on the others (Gilman and Ross, 1995).

In this chapter, we exhaustively describe the work proposed in (Higuera et al., 2012) in which we take up again the challenge of finding a universal pattern of post-transcriptional dynamical regulation for this kind of model, set out by Gilman and Ross. We accomplish this goal, first trying to reproduce the study of (Gilman and Ross, 1995) with different mono-objective global optimization techniques and afterwards, within the context of multi-criteria optimization. The latter has been carried out by calculating the Pareto-optimal (Miettinen, 1999) set of solutions according to two objectives. This set of solutions is considered to be a family of optimal solutions in the sense that it is not possible to improve one of the objectives without worsening the other; any choice of a unique solution would be a trade-off between both objectives. By means of this kind of optimization we try to simulate the natural evolutionary process of optimization and achieve a universal pattern of regulation. A pattern that allows the system to behave optimally under different varying conditions.

2.3. Substrate-cycle model

In order to achieve the above mentioned goal, the model examined by Gilman and Ross (1995), depicted in Figure 2.1A, has been used. It idealizes an animal cell that metabolizes glucose in blood for energy as long as the glucose concentration in the blood is adequate but synthesizes glucose

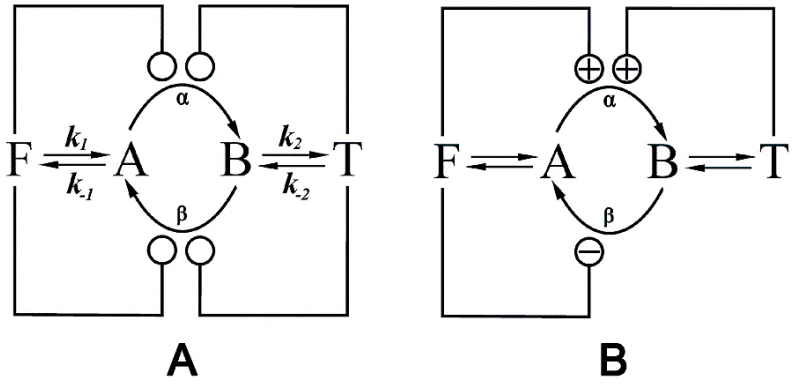

Figure 2.1: **Diagram of the model.** Substrate-cycle where enzymes α and β interconvert A into B (Figure A), both regulated by external effectors F and T. Arrows indicate reactions, knobs indicate regulation. The kinetic constants are: $k_1 = 10^{-2}s^{-1}; k_{-1} = 8x10^{-3}s^{-1}; k_2 = 10^{-2}s^{-1}; k_{-2} = 4x10^{-3}s^{-1}$. For enzyme α, V_{max} = 1,6 $mM\ s^{-1}, Km$ = 1,5$x10^{-3}mM$. For enzyme β, V_{max} = 3,5 $mM\ s^{-1}, Km = 2x10^{-3}mM$. Figure B shows an example of a regulation scheme of the model where symbol '+ índicates activation and '-ínhibition. In this case is activated by F and T because $R_{\alpha,F}$ and $R_{\alpha,T}$ are greater than 1 for the set of parameters taken as an example, and β is inhibited by effector F because $R_{\beta,F}$ is lower than 1. T has no effect on enzyme β because $R_{\beta,T}$ is 1.

for export when the glucose concentration in blood decreases too much. The model consists of a simple substrate-cycle where two metabolic intermediates (A and B) are interconverted by a pair of enzymes (α and β). These enzymes are regulated by two external reservoirs of metabolic species, and their concentrations are specified externally at any time (these variations of concentrations in a certain period of time are named "courses" from now on). An example of a time course is given in figure 2.2.

Since α catalyzes the conversion of A into B with rate v_α and β catalyzes the conversion of B into A with rate v_β the kinetic equations describing the temporal variation of these metabolic intermediates are described by the following differential equations:

$$\begin{aligned} dA/dt = &\quad k_1F + v_\beta - k_{-1}A - v_\alpha \\ dB/dt = &\quad k_{-2}T + v_\alpha - k_2B - v_\beta \end{aligned} \tag{2.1}$$

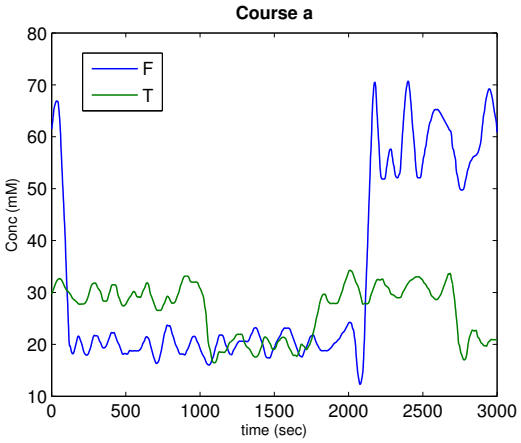

Figure 2.2: **Example of a time course of varying concentrations of *F* and *T*.** Y axis represents the varying concentrations of F (green) and T (blue). X axis represents the time.

$$\begin{aligned} v_\alpha &= \frac{V_{max,\alpha}A}{K_{M,\alpha}+A} R_{\alpha,F} R_{\alpha,T} \\ v_\beta &= \frac{V_{max,\beta}A}{K_{M,\beta}+A} R_{\beta,F} R_{\beta,T} \end{aligned} \tag{2.2}$$

Where K_M is the Michaelis-Menten constant and $Vmax$ the maximum velocity of the corresponding enzyme. The factors modifying the intrinsic Michaelis-Menten rate expression are:

$$\begin{aligned} R_{\alpha,F} &= \frac{K_{\alpha,F}+r_{\alpha,F}F}{K_{\alpha,F}+F} \\ R_{\alpha,T} &= \frac{K_{\alpha,T}+r_{\alpha,T}T}{K_{\alpha,T}+T} \\ R_{\beta,F} &= \frac{K_{\beta,F}+r_{\beta,F}F}{K_{\beta,F}+F} \\ R_{\beta,T} &= \frac{K_{\beta,T}+r_{\beta,T}T}{K_{\beta,T}+T} \end{aligned} \tag{2.3}$$

The parameters $K_{\alpha,F}$ and $K_{\alpha,T}$ are the dissociation constants for the complex of enzyme α, and $r_{\alpha,F}$ and $r_{\alpha,T}$ are the ratios of the catalytic rate constants for the enzyme for the effectors T and F respectively. Similar notation is used for the enzyme β. Depending on whether the resulting expression of R, Eq.2.3, is greater or less than 1 the corresponding enzyme, α or β, is activated or inhibited. A regulation diagram can be drawn from these statements. For example, if $R_{\alpha,F}$ is greater than 1 the enzyme α will be activated by the effector F and the connection between F and α in the diagram will have a '+'symbol. However, if $R_{\alpha,F}$ is less than 1 the enzyme α

will be inhibited by F and the connection between F and α in the diagram will have a '-'symbol, while if $R_{\alpha,F}$ is 1 the connection will not be shown since F has no effect on α. The same reasoning is applied to enzyme β. An example of a regulation diagram can be seen in Figure 2.1B.

2.4. Optimization of flux response

The regulation of the system, through the activation or inhibition of the enzymes α and β, is determined by the values of the set of these eight parameters ($K_{\alpha,F}$, $K_{\alpha,T}$, $K_{\beta,F}$, $K_{\beta,T}$, $r_{\alpha,F}$, $r_{\alpha,T}$, $r_{\beta,F}$, $r_{\beta,T}$). In order to optimize the flux response of the system the proper values of these parameters need to be selected. The main criterion for such optimization is the proper direction of the flux according to the system's need. The response of the system should be able to provide an appropriate flux of both F and T, in response to a given external condition. For example, the system metabolizes blood glucose for energy as long as the concentration in blood is adequate but synthesizes glucose for export if the glucose concentration in blood is too low. In order to evaluate the system response (Gilman and Ross, 1995, Eq. 4) formulated the following equation:

$$f = \xi_T(K_2B - K_{-2}T) + \xi_F(K_{-1}A - K_1F) \tag{2.4}$$

Where the terms $(K_2B - K_{-2}T)$ and $(K_{-1}A - K_1F)$ represent the net fluxes into the reservoirs T and F respectively, and ξ_F and ξ_T represent their need state (equations for ξ_F and ξ_T are fully described in Figure 3 of Gilman and Ross (1995)). If the concentration of F is below a specific target concentration, considered optimal, due to external variations, there will be a positive need state (ξ_F=+1), and the flux should flow from B to A in order to produce F. However, if the concentration of F is above the target concentration a negative need state will be induced (ξ_F=-1) and the flux should flow in the opposite direction (from A to B). The same applies to ξ_T.

If the algebraic sign of both the net flux and the need state into a reservoir is the same, the flux will be directed in the proper direction, so in this equation a positive value of f is considered to be a good response. However the concentrations of F and T may vary during the time-course and therefore their need states. In order to know how the network behaves during a whole time-course of concentrations of F and T, the integral of f over a period of time is calculated. Eq.2.5 gives some indication of the fraction of the period of time during which the flux was directed properly.

$$f_1 = \int_0^\tau f \, dt \tag{2.5}$$

The energy "cost" for performing this operation during the period of time τ was calculated by Gilman and Ross (1995) as a function of the operation of enzyme α, defined as:

$$f_2 = \int_0^\tau v_\alpha \tag{2.6}$$

2.5. Mono-objective global optimization

The nonlinearity and frequent multimodality of this kind of model make the optimization of its parameters a difficult task for traditional optimization methods, which are very sensitive to the initial values. Such problem models can contain several local optima, hence if the initial values are far from the global optimum it is difficult to assure a convergence towards it (Mendes and Kell, 1998). A robust alternative for solving complex-process optimization problems is to use global optimization methods (Banga, 2008). These kinds of methods can be roughly divided into two classes: deterministic and stochastic. Deterministic methods guarantee finding the global optimum under certain conditions. Their drawback is that the computational effort they require increases very fast with the problem size (Sendín et al., 2009). On the other hand, stochastic methods are based on probabilistic algorithms and do not offer the guarantee of finding the global optimum; however, it has been proved that they provide excellent results in solving complex-process optimization problems in reasonable computation time (Egea et al., 2010; Sendín et al., 2009).

In (Gilman and Ross, 1995) a GA, which belongs to the class of global stochastic optimization methods was used. The authors combined the flux response, Eq.2.5, and a weighted cost, Eq.2.6, by means of a single objective function:

$$OF = f_1 - m \cdot f_2 \tag{2.7}$$

A high value of OF is obtained not only when the network responds properly to changes of external concentrations but also when it does so at a low biological cost. Therefore a set of parameters must be found that maximizes f_1 and minimizes f_2, resulting in an optimal solution which would be a trade-off between a proper performance of the network (f_1) and the cost (f_2), merging these two concepts into one equation. As asserted in (Gilman and Ross, 1995), the GA procedure did not always find the global optimum, indeed for each run of the method a different value of OF was found making it difficult to assure the convergence towards an optimum. In this work we have performed a mono-objective study of the system using three different stochastic global optimization methods: an implementation of the GA used by Gilman; the enhanced scatter search SSm method described in (Egea et

al., 2010); and the multistart clustering method GLOBALm (Sendín et al., 2009). The scatter search method uses a relatively small population size, partially chosen by a quality criterion from an initial set of diverse solutions. It also performs systematic combinations among the population members. It is interesting to note the similarities and differences between scatter search and the original genetic algorithm (GA) framework. Both can be regarded as "population based" or "evolutionary" approaches, since both incorporate the idea that a key aspect of producing new elements is to generate some form of combination of existing elements. However, GA approaches are based on the idea of choosing parents randomly to produce offspring, and on using randomization to determine which components of the parents should be combined. In contrast, the scatter search approach does not place so much emphasis on randomization. Instead, the approach is designed to incorporate strategic responses, both deterministic and probabilistic, that take account of evaluations and history of the search. These components result into a more efficient search than GAs. On the other hand, GLOBALm is an extension of the multistart clustering algorithm for global optimization, incorporating new key features, including an efficient mechanism for handling constraints and a robust derivative-free local solver. The multistart clustering framework is based on starting with the generation of a uniform sample in the search space (the region containing the global minimum, defined by lower and upper bounds). After transforming the sample (e.g., by selecting a user set percentage of the sample points with the best function values), the clustering procedure is applied. The aim of the clustering step is to identify points from which the local solver will lead to already found local minima. Then, further local searches are started from those points which have not been assigned to a cluster, and the process is repeated until a stopping criterion is satisfied. The three methods have been applied to eight different time courses of external variations of concentrations of F and T, which are pictured in Figure 2.3. The first four courses (a-d) have been taken from (Gilman and Ross, 1995) and the other four courses (e-h) are sinusoidal periodical variations of F and T with different frequency, amplitude and phase.

2.5.1. Comparison of mono-objective optimization methods

Following (Gilman and Ross, 1995), a value of $m = 10^{-3}$ has been used for the objective function OF, Eq. 2.7. Ten optimization runs were repeated using the three methods on each of the courses. To allow a fair comparison between the three methods, we ran them with equivalent setting parameters: in the case of GA a population of 100 individuals and 100 generations was used, which represents a total of 10,000 evaluations. The same number of evaluations was set for SSm and Globalm. The optimum was achieved in less than 1000 evaluations, a relatively small number. In each of the courses a different optimal OF value was obtained, and remarkably the three opti-

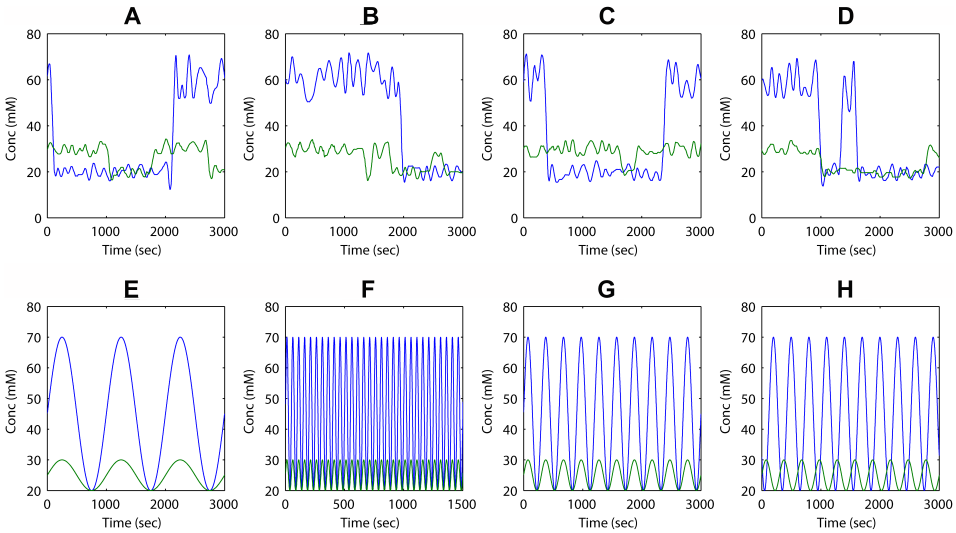

Figure 2.3: **Time courses of external variations of concentrations of *F* (blue) and *T* (green).**The first four courses (a-d) are taken from figure 4 of (Gilman and Ross, 1995) labeled as I, II, III and IV. The other four courses are obtained through the following sinusoidal equations: $F = a_1 sin((2\pi/T)t + \varphi) + (a_1 + min_F)$ and $F = a_2 sin((2\pi/T)t) + (a_2 + min_T)$, where a_1 and a_2 are the amplitudes, t is the time, T the period, φ the phase and $minF$ and $minT$ the minimum values of F and T. The first two (e and f) differ in their period but have the same phase ($\varphi = 0$); course e presents a high period ($T = 1000$) while course f presents a lower one (T=50). The two last sinusoidal courses (g and h) differ from each other in phase (for course g, φ=0 and for course h, $\varphi = 10$) and from the other two in period (T=300). The concentrations of the reservoir species F and T vary within two regimes. For F centred at 60 mM and 30mM and for T at 30 mM and 20 mM.

mization methods obtained the same optimal value of OF in each course. No significant differences were found between the three methods in terms of computation time or convergence towards the optimum. Figure 2.4 shows the evolution of the value of OF obtained by the three different methods for course *a*. It is not possible to decide which one is better because they all get to the optimum value in the first evaluations. It should be noted that in the figure the values of OF are negative because SSm and GLOBALm implementations were designed for minimization, therefore OF was multiplied by -1. The results of GA that was designed to maximize were also multiplied by -1 afterwards in order to compare the results with the other methods.

As stated above, stochastic global optimization methods do not generally guarantee convergence to the global optimum. However, the fact that the

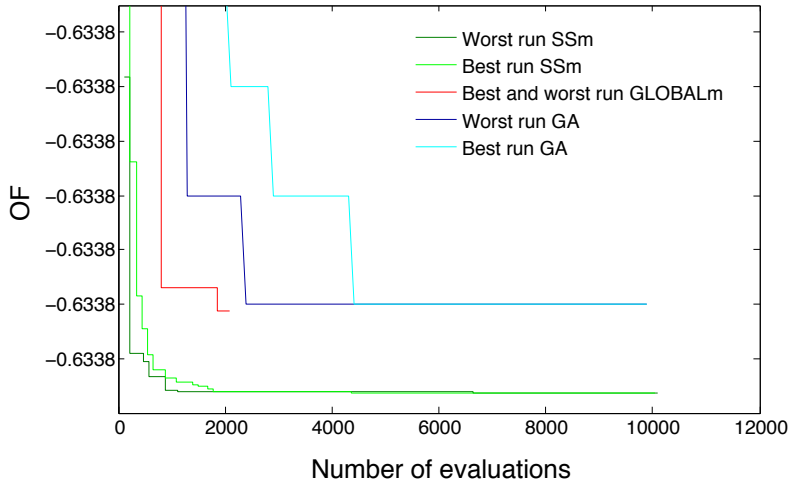

Figure 2.4: **Evolution of the value of *OF* for course *a* with the three mono-objective optimization methods SSm, GLOBALm and GA** Best and worst run of each method. The x axis represents the number of evaluations made by the optimization methods and the y axis represents the value of *OF*.

three different methods reached the same value of the objective function in several runs strongly suggests that in this case the global optimum was achieved.

In terms of regulation, due to the nature of the system, certain degeneracy in the solutions could be expected since different schemes of regulation could achieve the global optimum. However it is worth mentioning that after running any of the algorithms ten times in a particular course the regulation scheme corresponding to the optimal solution obtained for that course was in most of the cases the same. Nevertheless, among the different courses the resulting regulation scheme was different, as shown in Figure 2.5. In order to obtain these schemes, since R, Eq.2.3, depends on the values of F and T which vary during time, we compute their averaged values $\bar{R}_{\alpha,F}, \bar{R}_{\alpha,T}, \bar{R}_{\beta,F}, \bar{R}_{\beta,T}$ for maximum and minimum concentrations of F and T, (60mM and 30mM, and 30mM and 20mM respectively) these values are also shown in Fig. 5. The optimized parameters of the different courses are shown in Table 2.1. At this point, using mono-objective optimization we have reproduced the original results of (Gilman and Ross, 1995) using a GA and, additionally, two state of the art global optimization methods: SSm (Egea et al., 2010) and GLOBALm (Sendín et al., 2009). The difference with Gilman results is that in their case they obtained different OF values in different

	$K_{\alpha,F}$	$K_{\alpha,T}$	$K_{\beta,F}$	$K_{\beta,T}$	$r_{\alpha,F}$	$r_{\alpha,T}$	$r_{\beta,F}$	$r_{\beta,T}$
a	$9{,}4 \cdot 10^{5}$	$3{,}9 \cdot 10^{1}$	$1{,}3 \cdot 10^{4}$	$1{,}2 \cdot 10^{-10}$	$4{,}3 \cdot 10^{7}$	$4{,}2 \cdot 10^{9}$	$6{,}9 \cdot 10^{-7}$	$6{,}2 \cdot 10^{-9}$
b	$2{,}4 \cdot 10^{2}$	$4{,}5 \cdot 10^{4}$	$2{,}7 \cdot 10^{-9}$	$3{,}2 \cdot 10^{-10}$	$6{,}7 \cdot 10^{6}$	$9{,}9 \cdot 10^{9}$	$1{,}8 \cdot 10^{-10}$	$3{,}0 \cdot 10^{-3}$
c	$1{,}4 \cdot 10^{-8}$	$3{,}0 \cdot 10^{-6}$	$3{,}2 \cdot 10^{-8}$	$6{,}5 \cdot 10^{4}$	$4{,}8 \cdot 10^{6}$	$3{,}0 \cdot 10^{4}$	$7{,}2 \cdot 10^{-10}$	$1{,}6 \cdot 10^{7}$
d	$2{,}4 \cdot 10^{4}$	$1{,}4 \cdot 10^{5}$	$1{,}8 \cdot 10^{-10}$	$2{,}3 \cdot 10^{9}$	$5{,}1 \cdot 10^{9}$	$3{,}4 \cdot 10^{9}$	$4{,}4 \cdot 10^{-10}$	$4{,}4 \cdot 10^{7}$
e	$1{,}8 \cdot 10^{-2}$	$3{,}6 \cdot 10^{3}$	$1{,}1 \cdot 10^{-10}$	$1{,}5 \cdot 10^{-5}$	$1{,}3 \cdot 10^{4}$	$1{,}4 \cdot 10^{8}$	$2{,}0 \cdot 10^{-10}$	$3{,}9 \cdot 10^{-8}$
f	$3{,}2 \cdot 10^{-7}$	$5{,}4$	$5{,}3 \cdot 10^{-9}$	$5{,}4 \cdot 10^{-10}$	$3{,}2 \cdot 10^{-1}$	$3{,}6 \cdot 10^{8}$	$3{,}4 \cdot 10^{-7}$	$8{,}0 \cdot 10^{-2}$
g	$9{,}3 \cdot 10^{-3}$	$2{,}0 \cdot 10^{6}$	$1{,}0 \cdot 10^{-10}$	$1{,}0 \cdot 10^{-10}$	$1{,}0 \cdot 10^{5}$	$1{,}0 \cdot 10^{10}$	$1{,}0 \cdot 10^{-10}$	$2{,}5 \cdot 10^{-4}$
h	$4{,}9 \cdot 10^{2}$	$1{,}7 \cdot 10^{5}$	$3{,}5 \cdot 10^{-8}$	$4{,}3 \cdot 10^{-6}$	$3{,}4 \cdot 10^{8}$	$1{,}3 \cdot 10^{9}$	$1{,}0 \cdot 10^{-7}$	$5{,}2 \cdot 10^{-6}$

Table 2.1: Resulting optimized parameters of the eight courses (*a, b, c, d, f, g* and *h*) after running the mono-objective optimization.

runs of the AG for a certain course, actually the authors in their work affirmed that in some courses their genetic algorithm did not reach the global optimum. In our case the three methods reached essentially the same solutions in every run, strongly suggesting that the results presented here are very likely global optima.

2.6. Multi-objective global optimization

The performance of the network in terms of flux response and energy cost can be analyzed independently, therefore a more desirable and realistic approach to simulate the biological evolutionary process of optimization would be to consider the simultaneous optimization of these two criteria. In this case the result would not be a unique solution, but a set of solutions representing the trade-off between both objectives (Sendín et al., 2009). This approach is called multi-objective (or multi-criteria) optimization (MO), and despite being better able to cope with complex models, few applications are found in the systems biology literature in comparison with other scientific and engineering fields (Sendín et al., 2006).

The simultaneous optimization of multiple objectives differs from traditional mono-objective optimization in that if the objectives are in conflict with each other, the solution to the optimization problem will not be unique; instead, there will be a family of solutions known as a Pareto-optimal set (Miettinen, 1999). For the case in which there are two objectives, f_1 and f_2, the Pareto optimal set is a set of solutions in which no improvement can be obtained for f_1 without making f_2 worse, and vice versa. In this sense,

no point from this set can be said to be better than another; hence, in the absence of any further information about the problem, all Pareto-optimal solutions (which may be an infinite number for continuous problems) are mathematically equivalent. If one is interested in achieving only one final solution, there is a need for a decision-making process that allows one of the solutions in the set to be selected, using additional information. The choice of a particular solution is often subjective or difficult to express in mathematical terms, and it is therefore difficult to obtain systematically. However, the Pareto front of some multi-objective optimization problems shows a solution that can be considered to be the best compromise, i.e. the optimum of the front. These solutions are called "knee" points (Deb and Gupta, 2011). They are characterized by the fact that even a small improvement in one of the objectives (say f_1) would come at the cost of a much worse value of the other objectives (in this case f_2).

One of the advantages of the Pareto front perspective is that it allows the representation of the solutions within a diagram. Since the present model has two objectives, the solutions can be displayed in a 2D diagram dividing the graph in different regions. For instance, a set of solutions which control the flux properly but at a high cost will be situated together on one side of the diagram while solutions which do not control the flux so well but minimize the cost optimally will be placed on the opposite side, leaving the solutions which represent a trade-off between the two objectives in the middle. Since the solutions are laid out in regions along the Pareto front, with this approach it is possible to organize them in a graphical and more visual way, thus obtaining a wider perspective in the study of optimization applied to biochemical systems.

2.6.1. Calculation of Pareto fronts

Optimizations were carried out with the NBIWT weighted Tchebycheff method presented in (Sendín et al., 2010). NBIWT is a multicriteria optimization method that ensures an even spread of solutions in the Pareto front without the need of user-specified weights. It is based on the normal bounday intersection (NBI) method (Das and Dennis, 1998) with extensions based on the weighted Tchebycheff method (Koski and Silvennoinen, 1987). NBIWT also incorporates several stochastic local and global optimization solvers so it is able to handle both convex and non-convex Pareto fronts. Overall, it provides the user with a robust and efficient method of computing Pareto fronts without the trying of weights or other tuning parameters for the different objective functions.

An optimal (Pareto) set was computed for each of the eight courses for the cost, Eq. 6, and flux response, Eq. 2.5, simultaneously. The NSGA-II method (Deb, 2001; Deb et al., 2002) was also used initially but resulted in worse results than NBIWT. After applying the NBIWT weighted Tchebycheff

method on each of the eight different time courses we obtained eight fronts of solutions. Strikingly, each of them exhibited the same kind of Pareto front characterized by containing a clear knee point, which represents the ideal trade-off between the two objectives: in this case that solution combines a high flux response (f_1) at a low cost (f_2). Figure 2.6 shows the eight Pareto fronts and the corresponding knee point. Each of the solutions (points) of the Pareto fronts corresponds to a set of parameters (four K's and four r's). The regulation diagrams corresponding to the different solutions of the front are shown in the small diagrams of Figure 2.7 for courses c (Figure 2.7A) and d (Figure 2.7B). A certain similarity in terms of regulation schemes between the knee points of the different courses can be expected, since they are optimal solutions. In this way it would be possible to find a scheme (i.e., a set of parameters), which would be optimum for every course, however, although it can be noticed that the regulatory schemes of the knee points maintain some basic similarities, they are not identical.

There are some remarkable similarities among certain regions of the Pareto fronts within different courses that can be observed in Figure 2.7 A and B. For instance, on the right-hand side of the fronts many of the solutions presented a scheme where both enzymes α and β were inhibited. This would explain the fact that these solutions have a very low value of f_2, since f_2 (ecuation 2.6) is directly related with v_α, and since this enzyme is inhibited the value of f_2 will be low. In contrast to this, on the top left-hand side the solutions of the fronts presented a high value of f_2 and the most frequently found scheme was the one in which α was activated by the two effectors.

Interestingly, it was found that if the different knee points were interchanged within the different time courses, performing a cross-course comparison, the resulting behavior was also optimal in each of them. For example, the set of parameters corresponding to the knee point of course a also yielded optimal values of f_1 and f_2 for the other seven courses, and that happened with every knee point. This result suggests the existence of an underlying universal regulation pattern. In order to find such pattern, ten runs of the method NBIWT were carried out for each course and the regulation scheme of the knee point of each run was studied. It was observed that between different runs of a course the regulation scheme corresponding to the knee was slightly different. However it was noticeable that certain regulation pattern was more frequent than others. We considered this a consensus regulation scheme for this system, see Figure 2.8. The consensus set of parameters was calculated averaging the parameters belonging to the knee points which presented this scheme. To this end we calculated the average of each of the eight control parameters individually (α and β, is determined by the values of the set of these eight parameters ($K_{\alpha,F}$, $K_{\alpha,T}$, $K_{\beta,F}$, $K_{\beta,T}$, $r_{\alpha,F}$, $r_{\alpha,T}$,

$r_{\beta,F}$, $r_{\beta,T}$).

$$\left(\frac{\sum_{i=1}^{n} K_{\alpha F,i}}{n}, \frac{\sum_{i=1}^{n} K_{\alpha T,i}}{n}, \frac{\sum_{i=1}^{n} K_{\beta F,i}}{n}, \frac{\sum_{i=1}^{n} K_{\beta T,i}}{n}, \frac{\sum_{i=1}^{n} r_{\alpha F,i}}{n}, \frac{\sum_{i=1}^{n} r_{\alpha T,i}}{n}, \frac{\sum_{i=1}^{n} r_{\beta F,i}}{n}, \frac{\sum_{i=1}^{n} K_{\beta T,i}}{n} \right) \quad (2.8)$$

where n is the number of times that the consensus scheme has been observed. The resulting consensus set of parameters resulted to be as good as the optimal as well proving that a universal set of parameters for this model can be achieved.

Figure 2.9 represents the value of flux response (f_1) and cost (f_2) for course *f*, evaluated with the different knees of all the courses (optimal solution obtained for each course) and also with the consensus set of parameters. Similar results were obtained for the other seven courses. Remarkably, these values are very similar to each other, indeed the deviation of the different values of flux response for a single course evaluated with its optimal set of parameters, the other knee solutions, and the consensus set, is always less than 0.006 (that corresponds to a maximum deviation of 0.7 %). In the case of the cost the deviation is always less than 0.019 (maximum deviation of 1.38 %). Figure 2.10 shows a comparison of the flux response and cost obtained for each course run with its optimal set of parameters and the values obtained with the consensus set of parameters. It is noteworthy that there are practically no differences.

The regulation scheme corresponding to the consensus set of parameters is depicted in Figure 2.8. Enzyme α is activated by both effectors F and T, favoring the production of B, and enzyme β is inhibited by F. These schemes, where the enzymes are regulated by products and substrates of the reaction, are frequent in metabolism (Berg et al., 2006; Morán and Goldbeter, 1984). Specifically, the universal pattern obtained in this paper corresponds to several examples of substrate-cycles found in literature. One relevant example is the conversion of fructose 6-phosphate (F6p) into fructose 2,6-bisphosphate (F2,6BP) described in textbooks, e.g. (Berg et al., 2006). In this cycle of transformation ($F6P \rightleftarrows F2,6BP$) the kinase activity of the phosphofructokinase (PFK2) is activated by its substrate (F6P) and the activity of the phosphatase is inhibited by the product (F6P). Finally, the PFK2 is activated by the product F2,6BP. A similar behaviour is also observed in the regulation of gluconeogenesis and glycolysis in the liver (Berg et al., 2006, figure 16.28)), where there is an activation by substrate of the PFK mediated by AMP and F2,6BP and also an inhibition of FBPase by the same metabolites. This result reinforces the natural appearance of reciprocal feedback seen in multiple instances of biochemical networks.

2.7. Multicriteria optimization as a novel successful approach for studying regulation in metabolic networks

The starting point of the work presented in this chapter was the hypothesis that the regulation mechanisms of metabolic networks are the result of an evolutionary process of optimization. The idea that nature carries out optimizations in terms of metabolic regulation led us to search for existing universal regulatory patterns. In this work we have investigated the existence of a global optimal solution in a substrate-cycle previously presented in (Gilman and Ross, 1995) which was optimized using a GA. Since GAs are stochastic global optimization methods, they do not provide guarantees of convergence to the true global solution. When Gilman and Ross (1995) performed optimization runs on different environments (i.e., time courses of end point species concentrations), they were surprised that they did not find a global winner. Instead, solutions found to be optimal for one of the courses were not optimal for the other courses: they were "specialists" but not "generalists".

A first objective of the research reported here was to investigate this aspect further by, on the one hand, reproducing the original results of (Gilman and Ross, 1995) using an implementation of Gilman's GA and, additionally, two state of the art global optimization methods: SSm (Egea et al., 2010) and GLOBALm (Sendín et al., 2009). All these methods reached essentially the same solutions, strongly suggesting that the results presented here are very likely global optima. In contrast to what Gilman and Ross stated in their article: *"The GA in this study does not always find the global optimum for a given course"*.

As a second objective, we wanted to find a unique regulatory scheme by means of multi-criteria optimization. Here we have been able to find a generalist solution by switching from mono- to multi-criteria optimization. Instead of optimizing with respect to an objective function consisting of a fixed combination of the performance and cost terms (f_1 and f_2) we applied a multi-objective strategy (NBI-based weighted Tchebycheff, NBIWT) as presented in (Sendín et al., 2010). Essentially, this technique generates an even spread of points on the Pareto front, which correspond to the different relative weights of the two objective functions. As a result, instead of a single solution we found, for each environment, the set of Pareto-optimal solutions (set of optimal compromises for the two costs considered). Then, realizing that the Pareto fronts of all the environments exhibited a clearly defined knee point, we identified those solutions as the ones providing the best trade-offs between the two objectives. It should be noted that these solutions can not be found systematically using a classical mono-objective optimization scheme.

Interestingly, we found that although these solutions corresponded to

different regulation schemes, they performed optimally not only in the environment for which they were optimized, but also in the other environments. Repeating several times the multi-criteria optimization for each course we found a frequent optimal pattern of regulation, a regulation scheme that balances performance and cost optimally in every environment for the system considered. This can be seen as an indication on the existence of a universal regulation mechanism for substrate-cycles which are very frequent in metabolism. Several examples can be observed in the literature (Berg et al., 2006; Morán and Goldbeter, 1984); of special relevance is the PFK2-FBPase2 cycle (Berg et al., 2006), which has exactly the same regulation pattern that we have obtained by means of multi-criteria optimization.It is worth mentioning that resulting optimal trade-off solution (knee point) presented multiple global solutions (different regulation schemes with the same trade-off in the space of cost functions). This multiplicity is typical of multi-criteria problems where the cost functions are of the integral type.

This approach can be easily scalable to larger networks composed of more than one regulatory unit, such scalability poses no major problems other than increased computational requirements. The scatter search method scales quite well with problem size and has been successfully used in optimizations of several hundreds of decision variables. The increased computational cost can be handled by exploiting parallelization strategies. Versions of the scatter search and NBIWT solvers exploiting high performance computing hardware are being developed and therefore will enable the application to larger networks which could allow a more systemic optimization of metabolic systems. It should be noted as well that the approach presented is general in the sense that can be applied to other contexts (such as e.g. different individual cost functions, additional constraints, etc.) and can therefore be tailored to arbitrary multi-criteria optimization problems. Besides, the implications of the work presented in this paper go beyond the analysis of regulation based on optimality principles. For example, we can use a similar multi-criteria optimization scheme for the optimal design of biological circuits, as considered in synthetic biology. Optimization methods have recently been used for such designs, as discussed in e.g. the review by Marchisio and Stelling (2009). We suggest increasing the robustness and feasibility of these designs by adopting a multi-criteria framework similar to the one presented here.

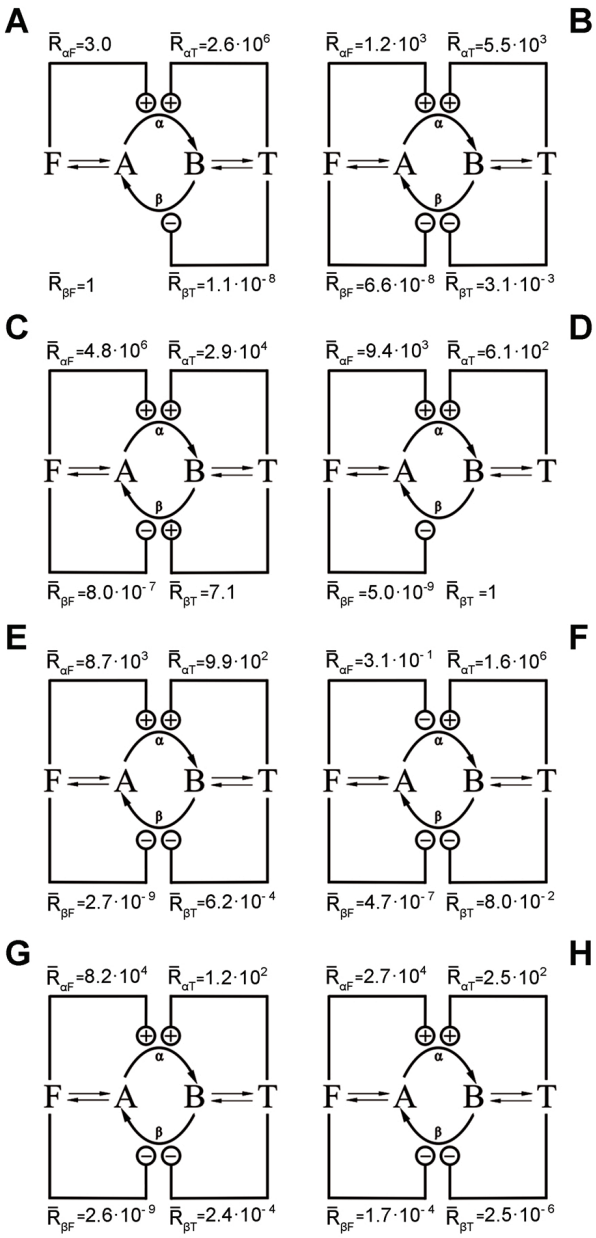

Figure 2.5: **Regulation schemes obtained by the mono-objective methods for all the courses.** Each diagram is drawnn based on the values of $\bar{R}_{\alpha,F}, \bar{R}_{\alpha,T}, \bar{R}_{\beta,F}, \bar{R}_{\beta,T}$ calculated with the resulting optimized parameters presented in Table 2.1.

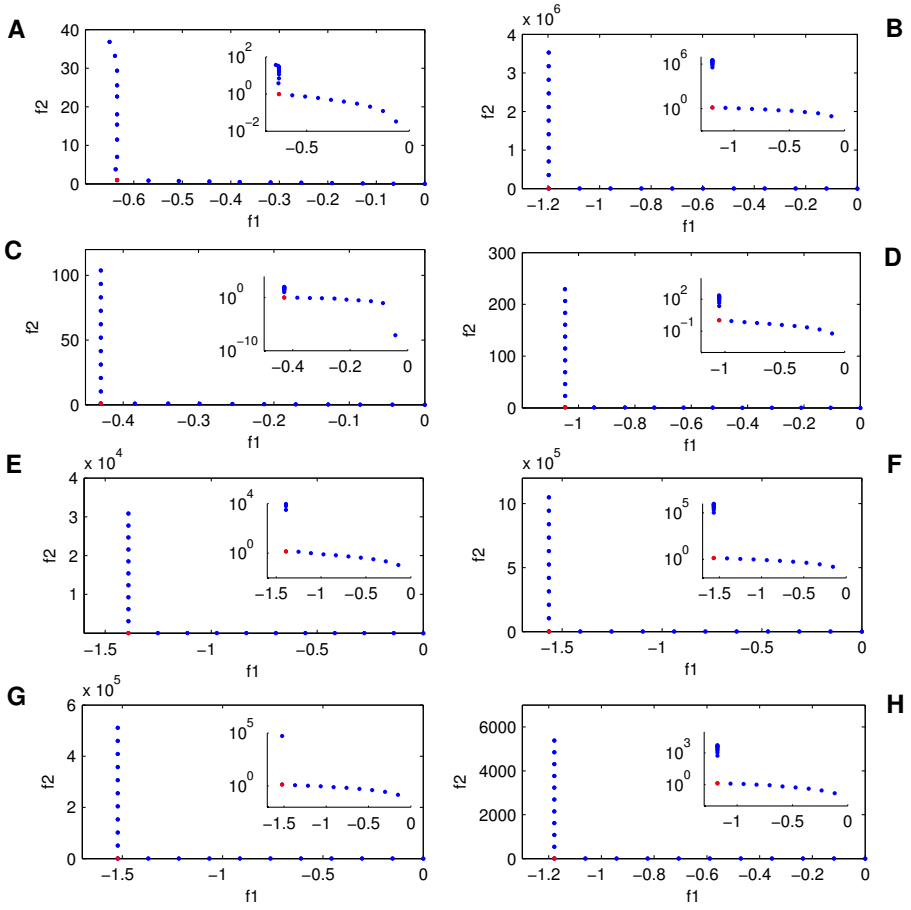

Figure 2.6: **Pareto fronts obtained for the two objective functions in the six courses** Flux response (f_1) vs. cost (f_2), for the eight courses (a-h). The insets are semi log plots. Each point corresponds to a set of the eight parameters. Red points indicate the knee point of each front. f1 has negative values because the method NBIWT was implemented to perform as a minimization method, in order to use it for maximization we converted the result of the objective function into negative values.

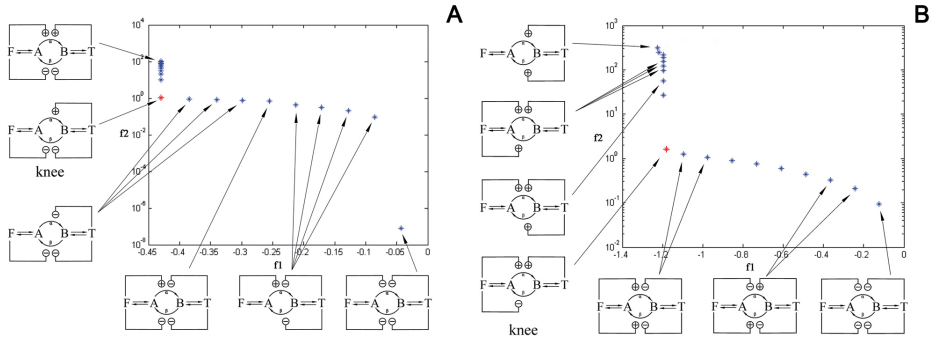

Figure 2.7: **Semi log plots of the Pareto fronts for courses *c* (A) and *d* (B).** Each point in the front corresponds to an optimal solution for f_1 and f_2 given by the estimated set of the eight control parameters. The small diagrams represent the corresponding regulation schemes deduced from each set of parameters. The red point corresponds to the knee point.

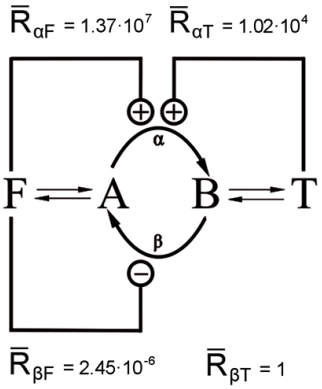

Figure 2.8: **Resulting regulation scheme drawn from the consensus set of parameters.** Enzyme α is activated by both effectors F and T ($\bar{R}_{\alpha,F} > 1$ and $\bar{R}_{\alpha,T} > 1$) whereas β is inhibited by F ($\bar{R}_{\beta,F} < 1$) and and T has no effect on it ($\bar{R}_{\beta,T} = 1$). The corresponding optimized parameters are: $K_{\alpha,F} = 1{,}5x10^{-2}, K_{\alpha,T} = 5{,}5x10^{3}, K_{\beta,F} = 9{,}2x10^{-8}, K_{\beta,T} = 3{,}7x10^{7}, r_{\alpha,F} = 1{,}9x10^{7}, r_{\alpha,T} = 2{,}3x10^{9}, r_{\beta,F} = 1{,}6x10^{-7}, r_{\beta,T} = 3{,}65$.

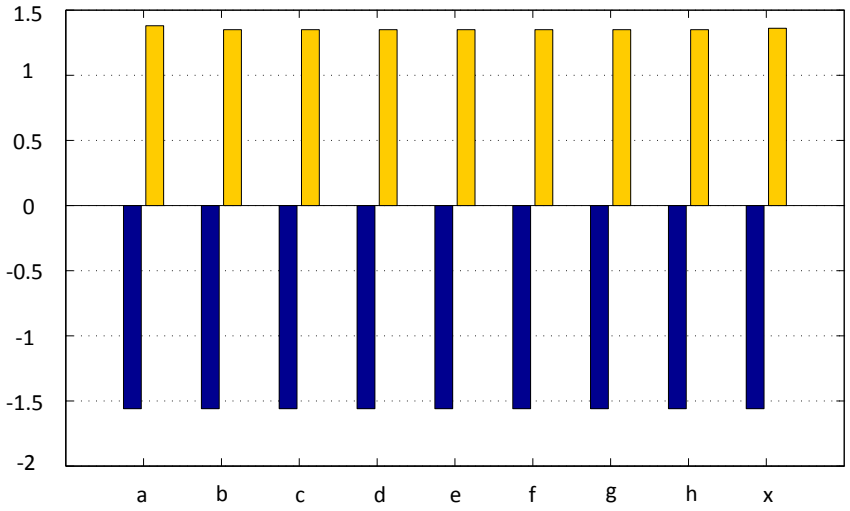

Figure 2.9: **Evaluation with knee parameters and consensus parameters.** Values of flux response (f_1 in blue) and cost (f_2 in yellow) for course f evaluated with the sets of parameters corresponding to the knees obtained for the different courses and with the consensus set of parameters, represented as x.

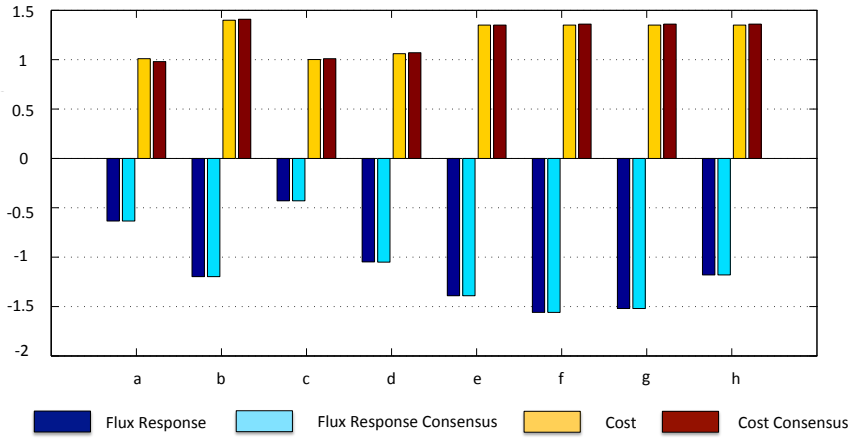

Figure 2.10: **Cross-course comparison.** Flux response (f1) and cost (f2) for the different courses evaluated with their optimal parameters (in dark blue and yellow) and with the consensus set of parameters (in light blue and brown).

Chapter 3

Expert system for clustering prokaryotic species by metabolic features

3.1. Introduction

The phenotypic response of certain species and organisms is strongly conditioned to their metabolic structures, which are at the same time the consequence of the genetic expression, also called genetic transcription. Relating metabolic structures of organisms at the level of metabolic pathways with their multiple phenotypic properties is a hard task. In this chapter we propose the design of an Expert System (ES) to cluster a set of 365 prokaryotic species by their similarity in the presence or absence of 114 metabolic pathways. Inspired by the human reasoning and based on the clustering method Self Organizing Maps (SOM), validity indices and hierarchical strategies, the ES finds relevant clusters at different stages. The resulting clusters prove that the use of metabolic features combined with the ES allows to handle a complex dataset and can help in the extraction of underlying information that may relate metabolism with phenotypic, environmental or evolutionary characteristics in prokaryotic species. The design and results of the ES have been published in Higuera et al. (2013).

3.2. Classification of prokaryotic species

Trying to understand the communities of microbial species is highly important because many natural and artificial processes are mediated by groups of microbes rather than by isolated entities. In order to create artificial communities or manipulate the existing ones it is necessary to comprehend the specific requirements of the individual species and to be able at a long term

to predict in which conditions are they able to survive.

One kind of microorganisms that are important in life are the prokaryotes and a way of studying them have been since many years trying to categorize the huge variety of prokaryotic organisms which is itself a challenging task (Hong et al., 2004). One of the reasons is the lack of a globally accepted concept of species for prokaryotes and the fact that their taxonomy is continuously being influenced by the advances in microbial population genetics, ecology and genomics (Gevers et al., 2005) . When it comes to assign an unknown bacteria to a species the experts usually do it identifying phenotypic or genome similarity.

The traditional method to classify prokaryotes has been since many years the identification of the 16s rRNA(Jain et al., 2009), a sequence highly conserved through evolution. It allows finding differences among microorganisms and building evolutionary trees, also called phylogenetic trees that show the evolutionary relationships among species that are believed to possess a common ancestor. Although the analysis of 16sRNA has been widely and successfully applied, experts have started to look for other kinds of information which may shed some light into the differentiation of prokaryotic species. One of them is the search of common metabolic characteristics, which some authors suggest to be not only a potential measure for the classification or differentiation of closely-related organisms (Lee et al., 2012) but also that their study may allow the finding of common functional properties that traditional methods such as the analysis of 16s rRNA is not able to find (Jain et al., 2009).

As explained in Chapter 1, in biochemistry a metabolic pathway consists of a set of reactions that take place inside the cell. Metabolic pathways involve the transformation of substrates into different products necessary for maintaining its life. These reactions are intermediated by molecules called enzymes, which are responsible for the proper performance of the pathway. Several different pathways can co-exist inside the cell. The collection and spatio-temporal organization of the pathways in the cell, together determines its metabolism. However, among species, neither do all present the same pathways nor do all enzymes have equal importance in their metabolism. This can be caused because different types of metabolism are adapted to the specific conditions of the environment in which a species live including the interactions with other species.

One of the main problems in separating prokaryotic species by their metabolic features is the wide diversity of possible representations and their comparison, such as the number of common enzymes between two organisms, the presence or absence of various metabolic paths and so on (Clemente et al., 2005). Another problem lies on the lack of information from some species in relation to others. For instance while *Escerichia Coli* is a very studied bacteria and its metabolic pathways have been deeply analysed, there are others

like *Rhodococcus Ruber* from which there exist little information. These are two problems addressed on this work based on a computational intelligent approach.

Unsupervised classification techniques, more commonly called clustering techniques, are machine learning methods that automatically group or separate elements by their similarity in a set of features. Examples of these techniques are hierarchical clustering, K-means, SOM or neural networks, which have proven to obtain succesful results in fields such as medicine, biology, biochemistry, ecology and microbiology (Park et al., 2003; Rabow et al., 2002; Stegmayer et al., 2012; Szaleniec, 2012). Moreover, in many cases these techniques allow the discovery of underlying information or patterns in data. In the case of prokaryotic species these techniques have been useful to infer inherent information from groups of species and also to predict certain behaviors or life styles (Bohlin et al., 2009; Larsen et al., 2012; Suen et al., 2007). For example, Larsen et al. (2012) utilized artificial neural networks to predict microbial community structure as a function of environmental parameters and microbial interactions.

When the number of samples and features involved is large, SOM can be very appropriate for cluster analysis and knowledge discovery. However, few has been done in the field of clustering prokaryotic species by their metabolic features. To date, the clustering methods which use these kind of features have been, (to our knowledge), above all hierarchical clustering methods applied to the reconstruction of phylogenetic or evolutionary relationships among species. The output of a hierarchical clustering is a tree called dendrogram which reflects the possible hierarchical clustering structure of the data. One of the advantages of the method is that the dendrogram has shown to result significantly similar to an evolutionary hierarchy when applied to prokaryotic and eukaryotic bacteria. Hong et al. (2004) use a complete-linkage hierarchical clustering to construct a phylogenetic tree that represents the similarity of metabolic profiles of a set of 43 microorganisms. The method helped them to study the changes in metabolism as a result of the evolutionary process. (Clemente et al., 2005) utilize an average-link hierarchical clustering to reconstruct phylogenetic relationships from other metabolic features. Casasnovas et al. (2006) enhanced a year later the method by using fuzzy clustering. Other authors like Jain et al. (2009) found functional similarities among clustered species such as adaptation to cold environments or ability to suppress agriculture pathogens by using a novel method based in hierarchical clustering to generate comparison trees based on characteristics collected from metabolic networks of bacteria. Nevertheless, the authors of this work use a very reduced sample of only twenty species.

Even though hierarchical clustering has shown good results in terms of reconstructing phylogenetic trees, in order to find inherent common functional characteristics different from metabolism among elements of a group, the

structure of a dendrogram does not seem very helpful. The reason is that the expert must decide the appropriate level or scale of clustering to start considering groups. Hierarchical clustering does not actually create clusters, but compute only a hierarchical representation of the data set (Sander et al., 2003). Another problem of hierarchical clustering is that for big data sets, of hundreds of elements, it is extremely difficult to identify and visualize relationships between elements.

An alternative are other classical unsupervised clustering algorithms, also called partitioning algorithms. They divide unlabeled data (sets of samples from which its belonging to a specific class is unknown) into defined groups (clusters) of similar elements. The fact that the output of these methods is a set of clusters makes easier the biological interpretation of the results and the possible extraction of common underlying information in data. However, clustering complex datasets is a very hard and arduous task. When applying clustering methods many problems arise. On one hand, algorithms are usually very sensitive to data which may be noisy or incomplete. On the other hand, similar algorithms can result in much worse performance than others when applied to the same dataset. This leads to the need of selecting the best method for the data being used. Moreover, there are cases in which the optimal number of clusters that best describes the topology of the data is unknown. This is specially common when dealing with biological data, that in contrast to other fields there is not any a priori information. A later stage of the clustering process consists of the assessment of the quality of the final partition. In the case of biological data it is also particularly difficult, especially in cases of lack of experimental scientific corroboration or expert supervision.

In the work presented in this chapter, published in (Higuera et al., 2013), we face the problem of clustering a complex dataset of 365 prokaryotic species by a set of metabolic features. This amount of species is considerably big and hierarchical clustering approaches become ineffective because of the problems described above. Concerning other classical clustering approaches, after different experiments, the most promising method was SOM. SOM clusters multidimensional data into a 2D map composed of n rows by n columns of neurons. Figure 3.1 depicts and example of a Self Organizing Map of dimensions 7x7 used to cluster 1080 samples.

Because of its unsupervised nature, one way of testing its performance on unlabeled data is by means of quantitative indices that measure the quality of the clustering. A review of different indices can be found in (Halkidi et al., 2001). Also Handl et al. (2005) review the use of these indices with special orientation to researchers in bioinformatics, where they have been less applied than in other fields. One very well known index is the Davies Bouldin (DB) (Davies and Bouldin, 1979). As mentioned before, we have tested different clustering strategies for our dataset, including Fuzzy Cluste-

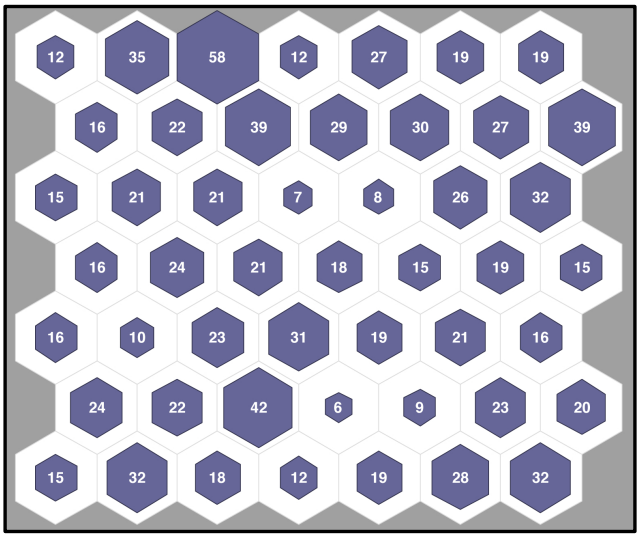

Figure 3.1: **Example of clustering with SOM.** Resulting 2D map after clustering 1080 samples with a SOM of size 7x7 equivalent to 49 neurons. Each cell is called neuron and gather similar samples according to certain features. Numbers indicate how many samples are clustered in each neuron.

ring (Bezdek et al., 1984; Bohlin et al., 2009; Larsen et al., 2012; Suen et al., 2007) and Learning Vector Quantization (Kohonen, 1988). We have verified that such methods obtained not only a worse distribution of the species in clusters than SOM, confirmed by the experts, but also worse values of DB.

Thus, because of the performance of SOM together with the idea of hierarchical strategies, we design an expert system (ES) to cluster a set of 365 prokaryotic species by 114 metabolic features with the aim of proving its validity for large datasets of microbial species. The ES is adapted to the nature and complexity of the dataset in question and capable of tackling the problems of clustering biological data mentioned before, as for example the estimation of the number of clusters that best represent the topology of the data or the use of an strategy to mathematically validate the clustering process. On the other hand, the design of the ES is based on what would be common in human reasoning with such complex data. When the human expert is faced with this type of data, a common practice is to begin separating the data with less difficulty making progresses towards higher levels of difficulty. At each step, a minimum degree of confidence is required about the appropriate progress. In the following sections the dataset used, the design of the system and its tested effectiveness clustering microbial species will be described.

	Metabolic Pathways				
Species	Glycolysis	Citrate cycle	Pentose Phosphate Pathway	Fructose and mannose metabolism	Galactose metabolism
Acaryochloris marina	35.44	23.08	45.45	21.21	8.77
Acetobacter pasteurianus	24.05	30.77	31.82	12.12	5.26
Bacillus amyloliquefaciens	35.44	34.62	50.00	33.33	24.56
Bacillus anthracis	36.71	40.38	47.73	21.21	12.28
Caldicellulosiruptor bescii	25.32	30.77	29.55	27.27	17.54
Caldicellulosiruptor hydrothermalis	22.78	26.92	29.55	22.73	17.54
Dechloromonas aromatica	30.38	46.15	29.55	13.64	8.77
Deferribacter desulfuricans	21.52	38.46	25.00	19.70	7.02

Table 3.1: Example of a reduced part of the data set. Rows represent species, columns metabolic pathways and values percentages of annotated enzymes of each pathway.

3.3. Determining the metabolic features of a set of prokaryotic species

Enzymes are key elements in metabolic pathways and determine their structure. The metabolic features selected for the clustering are the percentages of annotated enzymes that a set of species possess of certain metabolic pathways. This way, each species will have a vector of pathways assigned, which contains that percentage. This value is an indicator of how complete the pathway is in a certain species. A high value would mean that a species contains all or most of the enzymes of a pathway, while a zero value would mean that the species does not possess any enzyme of that pathway, therefore, that pathway is absent in that species. Consequently, the clustering will be made in terms of the similarity of the metabolic vector among species. This information was obtained from the KEGG Pathway Database (Kanehisa, 2002), which is currently one of the most comprehensive databases for studying metabolic pathways. A matrix of 365 species and 114 metabolic pathways was chosen as dataset. Table 3.1 shows an example of the structure of the data, where each row represents a species, each column a specific metabolic pathway and values represent the percentages of annotated enzymes that belong to a pathway in each species.

It must be taken into account the complexity of the dataset. On one hand it contains information from many different species and many metabolic pathways. Usually, the greater the number of features, the harder the clustering process becomes. In this case we are considering 114 features, which is already a significant number. On the other hand, it is also possible that the information is incomplete for some species. In order to know the number of enzymes that one species has of a certain pathway it is necessary firstly to perform an experimental procedure, which consists of sequencing its genome, followed by a computational supervised analysis called annotation. In this later analysis the locations of genes are identified and associated with relevant information. This information can be for example the metabolic enzymes that are encoded by genes. The process of annotation is very complex and errors during the process are sometimes inevitable, but they are corrected as the knowledge about the biology of the species in question is extended. Some of these errors can lead for instance to missing enzymes. This way, the information we find in databases about highly studied species like *Escherichia coli* will be more consistent and complete than information about others recently discovered or less studied like *Rhodococcus*. This determines the difficulty of applying classical clustering methods. For this reason it seems reasonable to face the clustering problem for this dataset from a hierarchical perspective, finding firstly clusters of species more clearly defined and later clusters less clear, obtaining in the last stages the elements more difficult to assign to a group. These will be, whether the ones with lack of information or the ones which are very different to the others and do not maintain fundamental metabolic similarities.

3.4. Brief introduction to Self Organizing Maps, SOM

SOM was invented by the Finish scientist Teuvo Kohonen in 1982 and it is considered a special type of artificial neural network. Basically, like other clustering methods, SOM first carries out a training phase and then a mapping phase. Each neuron of the SOM is associated to a weight vector that has the same dimensions as the input data. Firstly, random values are assigned to the weight vectors but during the training phase their values are updated. The update in SOM is performed by what is called competitive learning. SOM calculates for each sample the distance to all the weight vectors of the neurons of the map and selects the one with the minimum distance. That neuron is called the best matching unit (bmu) and its weight vector is updated with the values of the sample. This is repeated iteratively several times with all the samples of the dataset until the values of the weight vectors converge or a fixed number of iterations have been achieved. Equation 3.1

shows how the weight vectors are updated:

$$w_j(t+1) = w_j(t) + \alpha(t) \cdot h(j, w_{bmu})[x(t) - w_j(t)], j = 1 \dots k \tag{3.1}$$

Where x is the data sample provided to the network that is going to update the weight vector of neuron j. $w_j(t)$ is the weight vector in the iteration t of neuron j and $w_j(t+1)$ is the new one after being updated. k is the number of neurons and $\alpha(t)$ is called the learning rate, which decreases with the number of iterations. This is performed in order to allow the weights to vary more in early phases of the algorithm that is when weight vectors are more similar to the initial random values. As iterations pass the weight vector of neuron j will converge to values more similar to the samples for which j was the bmu. $h(j, w_{bmu})$ is called the neighborhood function. In SOM each time a data sample enters the network, not only the weight vector of the best matching unit is updated but also the weight vectors of the neighbor neurons. Depending on how close a neuron is to the best matching unit its weight vector will be more or less affected by the data sample. $h(j, w_{bmu})$ determines according to the distance between the neuron j and the bmu in which proportion the weight vector of j is affected by the data sample x. In the end of the training phase, when all the weight vectors are updated, the neurons are arranged/organized in a 2D map of nxn neurons. In this map, thanks to the neighborhood function, close neurons will contain close elements in the input space. Because of this particular property of SOM it is said that it preserves the topology of the data. This property can be exploited and it can result very useful depending on the data and the problem, as will be shown in Chapter 4

Once the training is finished and weight vectors have a fix value, each data sample is assigned to its best matching unit. This is called mapping and the result is shown in Figure 3.1. Each neuron gathers certain amount of input samples according to their similarity in the features selected for the clustering.

3.5. Combining SOM with validity indices in a hierarchical strategy: ES approach

Inspired on the human expert reasoning and based on hierarchical clustering strategies the ES is designed consisting of several stages, where at each stage different levels of complexity are considered. When humans tackle a problem, it is usual to start with the easiest parts to progress to the more difficult parts, this is the philosophy applied in this approach.

At a first stage, our proposed ES estimates the optimal size of SOM to start clustering the data. As explained in the previous section, SOM clusters

the input data in a two-dimensional map of n columns by n rows of neurons. In the work described in this chapter we will use the terms neuron and cluster indistinctly, however in other works as will be seen in the next chapter of this thesis a cluster can be considered a group of neurons. This difference depends on the use of SOM in each specific problem.

Therefore, at this first stage the ES estimates the optimal number of neurons/clusters that configure the SOM to properly divide the dataset. This will be performed calculating the DB index for different SOM sizes. Afterwards the ES starts an iterative process of clustering, where it finds relevant clusters at different steps. By relevant clusters it is meant clusters where their elements are close to each other but far from the elements of other clusters. To be able to identify such clusters we have defined a new validity index inspired on the DB index, but conveniently modified. The elements of these clusters are removed in each step from the dataset.

In order to verify that the clustering is being coherent, the partition obtained in each step is validated with the DB validity index, which acts as a quantitative measure, that allows to monitor the process and assess the behavior of the ES. This way, at the final stage we manage to have all the elements distributed in well-defined clusters validated with the DB index. The results were analysed finding whether phenotypical or functional significant similarity among elements of the same group, such as pathogenecy or non pathogenecy, similar environmental preferences or tolerance or production of certain substances. This proves that the use of metabolic features combined with a novel machine learning method adapted to a complex dataset can shed light into the understanding of microbial communities and help in the extraction of underlying information that may relate metabolism with phenotypic, environmental or evolutionary characteristics in prokaryotic species.

The above justifies the appropriateness of designing an ES that exploits the potential of a combined approach of a classical unsupervised method (SOM) and a hierarchical strategy. In order to build such system capable of clustering this set of data with reliability, we face a two-fold problem, on one hand the estimation of the optimal number of clusters that best describes the topology of the data and on the other hand, the clustering itself. Moreover it is necessary to evaluate the partition with a qualitative measure (DB), which gives us certain guaranty that the data have been correctly clustered into groups that share metabolic similarities. Unlike in many other fields where the assessment of the clustering may be clear, here it is hard to assure that the elements grouped together are biologically close. Clustering methods usually use distance measures to make comparisons between elements, in this case if we compare species by common percentage of enzymes it may occur that two of them are grouped together because their similarity distance is minimum, but we cannot guarantee that they are biologically similar. Our ES addresses conveniently these two problems. Figure 3.2 displays the scheme

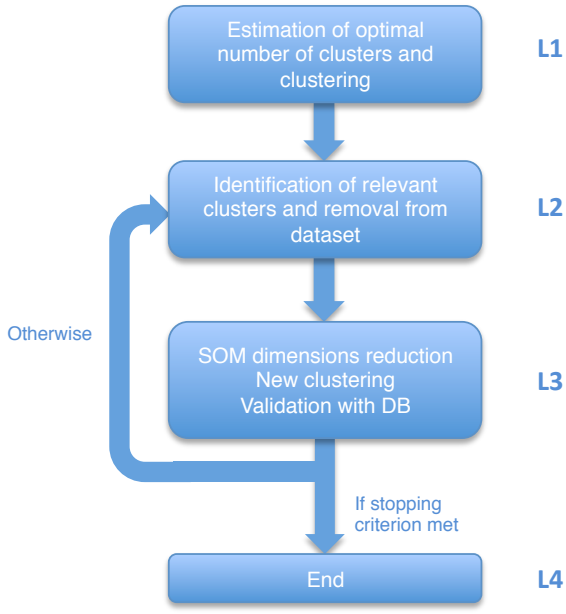

Figure 3.2: **Flowchart of the Expert System.**

of the proposed ES where each process is identified with its label (*L1, L2, L3* and *L4*). First of all the optimum number of clusters for the dataset is estimated (*L1*) and afterwards we take a step forward into the clustering process by applying iteratively the unsupervised method SOM and the *DB* validity index to cluster the data in stages using a hierarchical strategy (*L3*).

At each stage we separate the less problematic elements, those species that are grouped in clusters where their elements are very close to each other and far from the elements of other clusters. Since the dataset is reduced at each stage, the remaining data are clustered again but using a SOM of smaller dimensions than the previous one and again validated with DB. This process is repeated until a stopping criterion is met (*L4*).

The idea behind this strategy lies in finding step by step groups of elements that are significantly different to others and validate each step with a quantitative measure. This way we can verify if the clustering is being coherent and monitor the performance of the method. The following paragraphs describe in detail each stage of the proposed ES. As already mentioned, this implements the common human expert reasoning as expected.

3.5.1. Estimation of the optimum number of clusters and clustering (L1)

In order to estimate the optimum size of SOM we use the DB validity index, which has been successfully applied in many works of different fields to validate a partition of data previously clustered. Several validity indices have been also tested in our experiments without improvements with respect to DB in the dataset tested such as Silhouette, Calinski-Harabasz or Dunn's validity index. The aim is to validate the "goodness" of a partition without the need of a manual exploration process. DB performs this validation in terms of intra and inter-clustering distance and is defined as follows:

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{i \neq j} \frac{S_n(Q_i) + S_n(Q_j)}{S(Q_i, Q_j)} \tag{3.2}$$

Where n is the number of clusters, S_n the intra-cluster distance, which consists of the distance from the elements of a cluster Q to their center [1] and S the inter-cluster distance, the distance between centers. The first is calculated averaging the Euclidean distance from each element of the cluster to its center and the second one represents the distance between the center of cluster Q_i and cluster Q_j. Figure 3.3 shows a graphical example of both concepts. An optimal partition will be the one that maximizes S and minimizes S_n, in other words, the elements in the clusters are very close to each other, and far from other clusters, hence low values of DB would lead to better clustering.

The main goal of this part of the ES consists of finding out which configuration of SOM (SOM size) is the most appropriate to represent the topology of the data. In order to identify the optimal number of neurons/clusters to be used with SOM we calculate the DB index of several clusterings resulted from previously trained SOMs of different dimensions. The configuration that obtains the minimum value of DB is considered the optimum. The resulting clustering obtained with the optimal DB is considered the first partition of the dataset. From this moment, it starts the iterative part of the clustering process.

3.5.2. Identification of relevant clusters and removal from dataset (L2)

At this step we try to discriminate relevant clusters that are compact (their elements are very close) and well separated. In order to identify them, we introduce a new validity index inspired by the DB index. Using the

[1] The center of a cluster is a vector of the same size of the input data that is representative of the samples grouped in it. In this case the center is equivalent to the weight vector of the neuron.

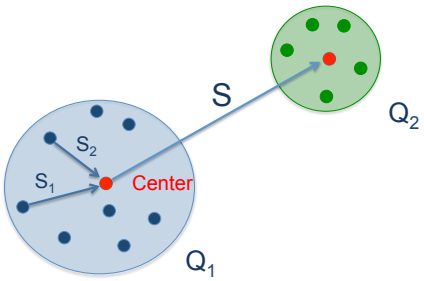

Figure 3.3: **Inter-cluster and Intra-cluster distances.** Q_1 and Q_2 represent two clusters. Blue and green points represent data samples while the red point represent the center of the cluster. Intra-cluster distance (S_n) is calculated by averaging the distances of each data sample (S_1, S_2, etc) to the center of the cluster. Inter-cluster distance is the distance from the center of one cluster (Q_1) to the center of another cluster (Q_2).

concepts of intra- and inter-cluster distances the proposed DB' is defined as follows:

$$DB' = \frac{1}{n}\sum_{i=1}^{n}\frac{S_n(Q_i) + S_n(Q_j)}{S(Q_i, Q_j)} \tag{3.3}$$

DB' describes how separate is one cluster from all the others and how compact its elements are. For instance, the value of the term: $\frac{S_n(Q_i)+S_n(Q_j)}{S(Q_i,Q_j)}$ for cluster Q_j represents how distant it is from cluster Q_i by means of S, and how close are the elements grouped in it by means of S_n. The value of this term is calculated for Q_j and the rest of the clusters and the average (DB') is the indication of how dissimilar is Q_j with respect to the majority of the other clusters. The same is done for the rest of the clusters. The result is an array of n elements, being n the number of clusters, where each position contains the value of DB' for a specific cluster. The difference between DB' and DB is that DB gives an overall indication on how well the data is clustered whereas DB' gives specific information on each cluster.

A low value of DB' indicates that a cluster has little similarity with the others, consequently it could be said that the clusters with a minimum DB' are the "purest" because the elements grouped together are very close to each other but far from the elements of other clusters. The clusters are then sorted by their DB' value in ascending order and the ones with the lowest values of DB' are separated from the data set. The criterion established to select how many clusters should be removed in each iteration will be explained in Section 3.6 because it highly depends on the data and on the optimal number of clusters estimated.

3.5.3. Reduction of SOM dimensions (L3)

After having removed some of the clusters and therefore several elements, the clustering is repeated with the remaining data; however instead of maintaining the SOM dimensions of the previous hierarchical level (L2), these are reduced. The reason is trying to compact the data, this way the method is able to find new clusters with more than one element and which could be discriminated as pure clusters in the next step. The criterion followed to reduce the SOM size consists of decreasing the SOMs dimensions of the previous stage from nxn to $(n-1)$x$(n-1)$. This way if the optimal number of clusters is estimated to be a SOM of 12x12 neurons in the next step the SOM size will be 11x11, then 10x10 and so on. This criterion will be further discussed in Section 3.6.

Considering that this is an iterative algorithm, processes L2 and L3 are repeated until a stopping criterion is met; therefore, in order to measure in each step the performance of the method, DB is calculated for the partition obtained in L3. A comparison of DB among the different clusterings allows us to monitor the process and assess the behavior of the method. If DB decreases, the quality of the clustering will be improving, otherwise it will be worsening.

3.5.4. Stopping criterion (L4)

Processes L2 and L3 must be repeated until a stopping criterion is met. There exist three possibilities to define this criterion. One possibility is that DB decreases progressively in each step, what would mean that the clustering is improving. In this case, steps L2 and L3 are repeated until the data set is empty or until SOM size reaches the dimension 2x2 and can not be further decreased. SOM 2x2 is the minimum size of SOM because size 1x1 would only consist of one cluster. The second possibility is that DB increases, in that case the algorithm would stop because that means that the clustering is worsening. The third possibility occurs when DB decreases until certain point and then it increases. This means that the decrease of the size of whether the SOM or the data set has been excessive and does not improve anymore the clustering. In that case the algorithm will stop and the final partition will be the one obtained with the last SOM and the clusters previously discriminated. Figure 3.4 displays an illustrative example of the proposed ES starting with a SOM with 10x10 neurons, i.e. 100 clusters. The system progresses once certain clusters are removed, achieving a SOM with less number of neurones each time until the convergence criterion is met. We can see at each stage the SOM is reduced in one dimension. The expert system stops the iterative clustering when the value of DB starts to increase, what happens when a SOM of dimensions 7x7 is applied, as can be observed in the graph on the right side of Figure 3.4.

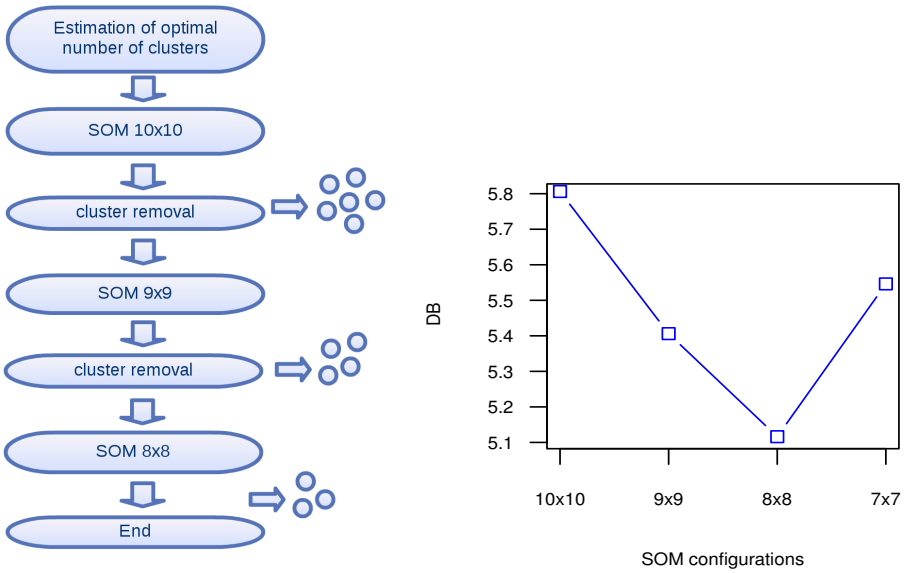

Figure 3.4: **Illustrative example of the ES.** On the left: stages of the ES for the case in which DB decreases until certain point (SOM 8x8) and then increases; On the right: graph of the evolution of DB after each clustering step with SOM.

3.6. Applying the ES to the dataset of prokaryotic species

The methodology explained in the previous section has been applied to the data set, determining firstly the optimum number of clusters and then proceeding with the different processes of the new hierarchical strategy with different levels. We trained eleven SOMs of different dimensions (from 7x7 to 17x17) with the same data set, clustered it and calculated the DB index value for each configuration. This process was repeated ten times. It was observed that the minimum value of DB was obtained with a SOM of dimensions 10x10. Table 3.2 displays for the eleven SOM configurations the values of DB for all the executions. Each column represents a SOM of a certain size and the rows represent the different runs for each SOM. The last row exhibits the minimum DB value obtained for each of the ten runs performed for each configuration of SOM. This is also displayed in 3.5 for clarity. It is remarkable that on one hand DB tends to decrease from configuration 7x7 to 10x10 but it increases from configuration 10x10 on. Taking this into account we are able to assume the configuration of SOM 10x10 as the most appropriate.

Once the optimal number of clusters has been estimated, the strategy progresses as explained in Section 3.5. First of all, the clustering obtained

		SOM configurations										
		7x7	**8x8**	**9x9**	**10x10**	**11x11**	**12x12**	**13x13**	**14x14**	**15x15**	**16x16**	**17x17**
Runs	**1**	6.00	6.39	6.36	6.51	6.36	6.75	6.88	6.89	6.87	7.03	7.23
	2	6.72	6.19	6.64	6.38	6.16	6.39	6.59	6.94	6.89	7.04	7.27
	3	6.14	6.35	6.39	6.22	6.61	6.64	6.80	6.94	6.79	7.25	7.32
	4	5.96	6.36	5.91	6.69	6.45	6.43	7.25	6.67	6.94	6.82	7.21
	5	6.02	6.13	6.17	6.47	7.17	6.49	6.81	7.14	6.87	7.53	7.00
	6	6.42	6.24	6.19	6.29	6.61	6.38	6.65	6.69	6.77	7.06	6.96
	7	6.27	5.93	6.49	6.96	6.86	6.80	6.96	7.09	7.00	7.21	6.99
	8	6.32	6.14	6.43	6.25	6.43	6.44	7.11	7.14	6.89	7.23	7.04
	9	6.05	5.96	6.11	6.65	6.43	6.55	6.96	6.96	6.98	7.19	6.97
	10	6.23	6.45	6.23	5.81	6.29	6.71	6.58	6.36	7.18	7.21	6.97
	Min DB	5.96	5.93	5.91	**5.81**	6.16	6.38	6.58	6.36	6.77	6.82	6.96

Table 3.2: Values of DB obtained for different SOM configurations. In bold the minimum value of DB obtained.

with the best value of DB with SOM 10x10 is considered the first partition of the data (L1). From this point the iterative process begins until the stopping criterion is met (L4). In the process L2 (Section 3.5.2) the elements grouped in relevant clusters are removed from the data set. To this end, instead of removing a constant number of clusters at each iteration, we establish a logical criterion, which consists of reducing the rate of clusters removed at each iteration. The reason for this is that the clusters removed in the first iteration will be the ones more clearly different, and in each following iteration (L3) this difference will have decreased. Therefore, after several experiments the starting removal rate we have estimated as adequate is 30 %, that is to say, from a SOM of 100 clusters the elements corresponding to the first 30, arranged in ascending order of DB', are deleted from the data set. For the next iterations this rate will be reduced in 33.33 %. For three consecutive iterations of the algorithm the rates will be 30 %, 20 % and 10 % respectively. It should be mentioned that the optimal behavior of the ES was obtained applying this criterion. We observed that in that situation DB generally tended to decrease in each iteration, what did not happen in other experiments as for instance when a constant rate of clusters was eliminated.

Regarding the criterion to reduce SOM dimensions, explained in Section 3.5.3, because the optimal SOM size has been estimated to 10x10, in the following steps of the system the SOM size will be 9x9, 8x8, and so on. It should also be mentioned that other experiments were performed not reducing the size of SOM, what resulted in worst performance of the ES.

Since SOM is a stochastic method, because of its initialization process, the resulting clustering at each step may slightly vary among different runs. To avoid misunderstanding of the results we run the ES 100 times (phases repeated were L2-L4 of the ES, since L1 contemplates the estimation of the optimal number of clusters) and analyse the results. It has been observed that starting with a SOM 10x10 in 31 % of the cases DB decreases in the

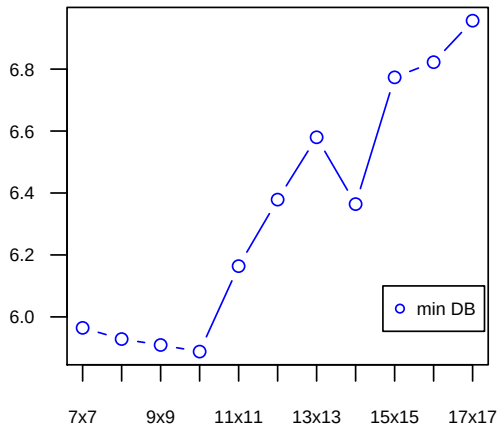

Figure 3.5: **Minimum DB values obtained for each configuration of SOM.**

following stages with SOM 9x9, SOM 8x8 and afterwards, and in the 45 % of the cases the ES reaches the stopping criterion at the size of 7x7, what means that DB presents a lower value in each next stage with a SOM 9x9 and 8x8 and it increases for SOM 7x7. This evidence shows us that the 76 % of the cases present a decrease of DB until SOM 8x8. Therefore an excessive reduction of the SOM size does not improve the clustering. This may also be caused because the reduced data set used in the fourth iteration of the clustering contains residual or complex elements hard to cluster, what would explain the variability of the behavior of the ES at that point.

Consequently, in order to guarantee a good performance of the proposed ES for this kind of data, the ES should halt when DB stops decreasing, what occurs in a general basis when the ES reaches a SOM 8x8. The final clustering is then formed by the clusters eliminated in every previous iteration of the ES and the clusters obtained in the last partition of SOM. Figure 3.6 represents the resulting behavior of the ES after being run 100 times. In Figure 3.6A are displayed the percentages of cases when DB decreases progressively until SOM 8x8 and afterwards, when it decreases only until SOM 8x8 and then it increases, and when other behaviour occurs. In Figure 3.6B are displayed those cases when DB decreases at least until SOM 8x8. Figure 3.7 exhibits an example of the two main behaviors of the ES when applied to our data. The graph on the left displays an example of a run of the ES where DB decreases until SOM 8x8 and then increases. The graph on the righ shows an example of a run where DB decreases also at SOM 7x7. These are the two main behaviours observed after applying the ES to the data set.

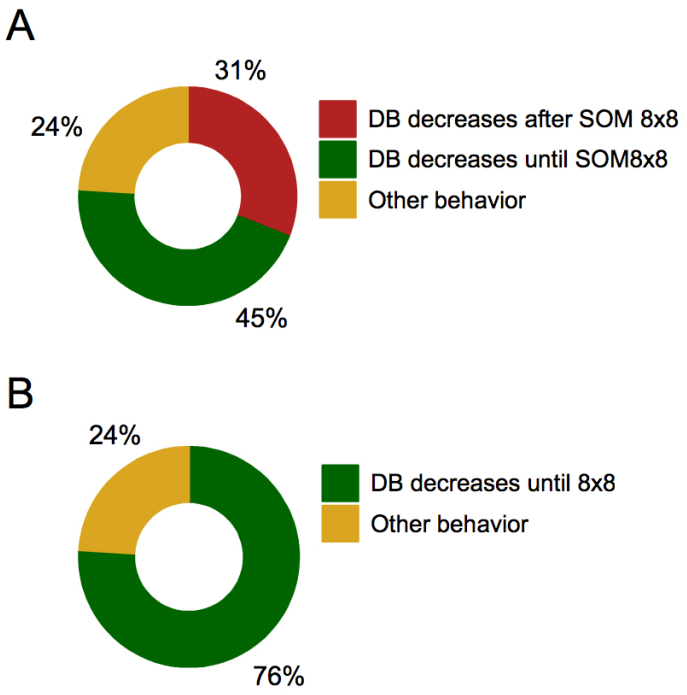

Figure 3.6: **Resulting behaviors after running the ES 100 times with the dataset.**(A) Percentage of cases when DB always decreases, when it decreases only until SOM 8x8 and when other behaviour occurs; (B) cases when DB decreases at least until SOM 8x8.

3.7. Biological significance of the obtained clusters

After the application of the ES to the data set at different stages, we have a guarantee that the data have been clustered with a sufficient level of confidence, based on the use of the validity indices. The ES obtains a set of clusters of prokaryotic species similar in their metabolic profile. In this section, the clusters obtained iteratively by the ES in one of the cases when DB decreases until SOM 8x8 are analyzed. The results obtained were very similar in comparison to other runs of the ES in which DB also decreased until 8x8.

In order to determine the biological significance of the resulting clusters we analyzed their content and identified also other important common similarities among elements of the same group. We found that many species in the same cluster shared not only metabolic similarities but also phenotypic or environmental properties, information that was not used in the clustering process. As explained in Chapter 1, phenotype is known as the observable characteristics of an organism such as morphology, development, physical

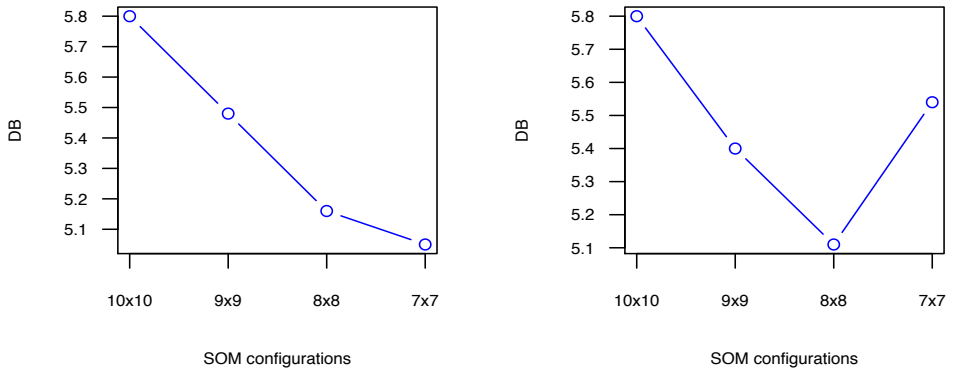

Figure 3.7: **Example of the two main behaviours of ES applied to our dataset.** SOM configurations against DB.

and biochemical properties, physiology or behavior. In microbes some of these characteristics are for example pathogenesis, the ability to grow in harsh environments or the capability to form spores to protect themselves. We say that several species have environmental similarities when they are often found in the same environments, however it should also be mentioned that clusters do not necessarily have to correlate totally with environmental preferences because the same metabolism is possible in different environments and in different environments are possible different metabolisms as well.

The fact of finding these similarities among elements clustered together by their metabolic features can be very helpful to explore to what extent metabolism may be related with those characteristics and specifically to find out if certain metabolic pathways are involved in the development of such features. Since the process of clustering (L3) is performed three times, one with a SOM of dimensions 10x10 afterwards with 9x9 and finally with 8x8, in order to properly analyze the biological results we will name from now on each run of L3: *clustering stage* as the three phases where different partitions of the data set have been performed.

In this section the content of the most relevant clusters obtained at the three different clustering stages will be analyzed. The individual characteristics about the species grouped in the clusters were consulted in the current literature about the species in question. Firstly it is remarkable that clusters obtained at the first two clustering stages contain more elements than the ones obtained at the last one. While at first stage we can find groups of 10, 12 elements at the last stage we found mostly groups of one or two species. This is a consequence of the strategy applied since elements at the beginning

of the process are easier to cluster and the ones which remain at the end are more difficult to assign to a group.

Figure 3.8 exhibits five representative clusters obtained at the first clustering stage. The circles stand for the clusters and the legend describes the taxonomy and common characteristics among the species grouped in it. The first cluster, Figure 3.8A, groups species with the same taxonomy, they all belong to the Lactobacillaceae family (one of the bacterial taxonomic ranks) and are gram positive which is a characteristic related to the cell wall. Other important common characteristic is that all of them are typically found in the intestinal environment. Other interesting cluster is depicted in Figure 3.8.B which consists of eight members of the Enterobacteriales order, all of them pathogenic. At this stage there have been obtained more clusters also of species which belong to the same taxonomy and share common phenotypical features, however we found three clusters particularly interesting (Figure 3.8 C, D and E) which gather species taxonomically different. The fact of belonging to different taxonomy category means that they are phylogenetically distant. Phylogenetic proximity is a criterion based on how evolutionary far species are from each other to assign a certain species to a taxonomic rank. For example, cluster C contains a mix of twelve bacteria that belong to different phylums, which are taxonomic ranks very distant in the tree of life. The twelve of them usually live in geothermal environments or submarine volcano and they have the quality of surviving in habitats rich in sulfur and high temperatures. The fact that these species are grouped together means that although they are taxonomically different, that is to say they are phylogenetically distant, they are metabolically close and this similarity is correlated with an environmental proximity. The same metabolism appears in species that live in similar environments but are phylogenetically distant. The same metabolic adaptations are observed in distant branches of the tree of life. This finding reflects that somehow the metabolism may be implicated in the adaptation of such species to the specific environment.

Two more clusters with this peculiarity have been found at the first stage. Cluster D contains seven species, which belong also to different phylums (*Firmicutes, Actinobacterias, Bacteroidetes* and *Spirochaetes*) and the seven of them share their main environment: oral. The same happens with cluster E that gathers eight species of two different phylums: *Proteobacteria* and *Actinobacteria* commonly found in soil.

At stage 2, we found smaller clusters but not less interesting, four have been selected and depicted in Figure 3.9 A, B, C and D. Cluster A contains four species of *Actinobacteria* which are gram positive and also the four of them have the quality of being pathogenic. Cluster B groups species of *Deltaproteobacteria*, gram negative and usually found in habitats such as freshwater aquifer, marine and subsurface sediment. Clusters C and D gather *Gammaproteobacteria* and *Firmicutes*, the first ones usually live in marine

habitats and the second ones have the ability to tolerate hard environmental conditions, such as temperatures between 35 and 55°C and are predominantly found in soil. In addition they all possess an endospore to protect themselves against harsh environments.

As the system moves forward to next stages it becomes more difficult to extract common biological information from clusters. As stated in previous sections, due to the complexity of the data, species less studied or species hard to assign to a group are swept along till late stages of the system. It is remarkable that at the third stage the number of elements per cluster decreases, one cluster contains six species but most of them are mainly formed by 1 or two elements, indeed the 32 % of the clusters contain two species and the 28 % just one. This reflects that at this stage the system finds difficult to cluster many species together because on the whole, the remaining species are very different among themselves in terms of metabolism and they tend to be clustered by the system in many separate groups.

However it is worth mentioning that most of the clusters are formed by phylogenetically close bacteria such as cluster A in Figure 3.10 but we also find a cluster formed by six species from the *Actinobacteria* and *Firmicutes* phylums usually found in gut, Figure 3.10 B.

In conclusion, the fact of finding several species which share metabolic and other common characteristics may be decisive in terms of finding out to what extent is the metabolism responsible in the development of these characteristics, and moreover to be able to find which exact pathways may be implicated in each of them.

In addition, it should also be mentioned that most of the clusters contained species from the same taxonomic category, what induces to think that on the whole, metabolic similarities among species are related to their phylogenetic proximity. However, some clusters contained species taxonomically different but with similar metabolic and environmental preferences. This finding is particularly interesting because it suggests that the metabolism may be related not only to the phylogeny but also in some cases to the adaptation to the habitat where a species lives. This adaptation through time may have consisted of acquiring metabolic pathways that their ancestors did not have or that they had lost.

If several species of different taxonomic rank are gathered in the same cluster is because they share similar metabolic profiles. If they also share environmental preferences, those pathways that have been relevant to cluster them together, could be considered environment-specific pathways and may shed some light into the understanding of microbial communities and their relation to their habitat. This application could be especially important in fields such as microbial ecology, which concerns the study of the relationships among microorganisms and with their environment or metagenomics that study the genetic material recovered directly from environmental samples.

3.8. Success of the expert system in clustering complex biological data

We have presented an ES capable of clustering a complex biological dataset composed of approximately three hundred different microbial species by their metabolic features. The system has been designed according to what a human expert would perform in such case, finding at first clusters easier to identify and discriminating problematic elements to be clustered later. The system groups, using SOM, species in stages following a sort of hierarchical strategy that makes also possible the evaluation of the clustering at each stage, by calculating the DB validity index. The resulting clusters were coherent biologically speaking. It has been observed that many of them contain species which not only are metabolically similar but at the same time they share other common phenotypical functional features such as pathogenecy, ability of adaptation to certain environments or capability to form spores to resist external threads. Individual classical clustering or hierarchical methods were not able to achieve the performance of our ES.

With this ES we manage on one hand to cluster with reliability a complex set of biological data and on the other hand to be able to infer valuable information from the resulting clustering that may help biologists to further study the relation between metabolism, environment and other functional characteristics. Furthermore, it opens the possibility of designing a new utility of the expert system to behave as a predictor of phenotypic characteristics from metabolic features. Regarding the strategy used for this dataset, it is also worth mentioning that it can be applied to other complex datasets composed of heterogeneous or incomplete data, what gives the system a general validity to cluster complex sets of data.

One possible future improvement for this system could consist of exploiting the property of SOM of preserving the topology of the input data, which has not been used in this work. As explained in section 3.4 closer neurons in the map tend to cluster similar samples of the input space. Using this characteristic, an option at the time of selecting relevant clusters would be to consider groups of neurons within certain proximity, instead of single and isolated neurons in each of the stages described. In that case the system could improve because bigger clusters would be obtained. However, it can also result a complex task at the time of automating the process because in this specific problem there exists no information about how biologically similar samples in close neurons are or where to set the border of vicinity. These problems could make the system fall into grouping neurons not biologically similar.

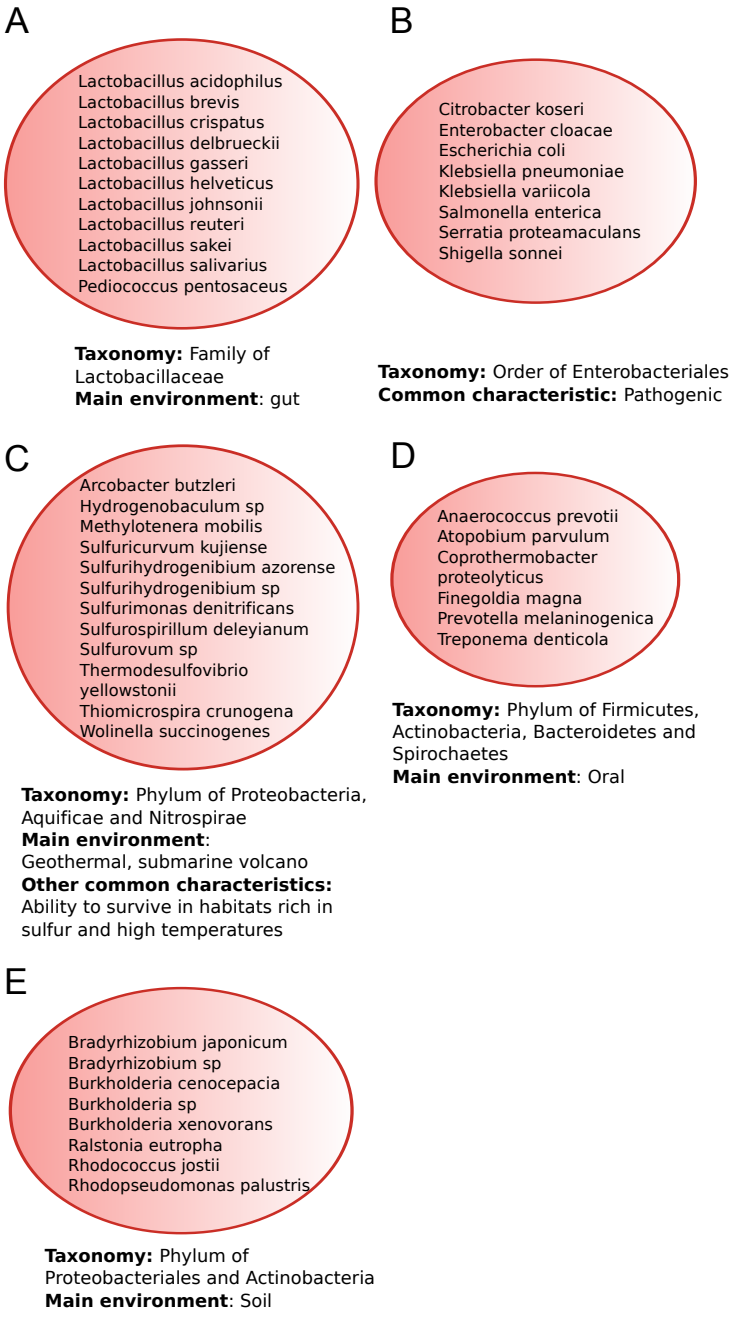

Figure 3.8: **Representative clusters obtained in the first stage of the clustering with the ES.** Each circle represents a cluster. The names of the species clustered together are displayed inside each cluster. Below each cluster, information about common phenotypic characteristics among the species of the cluster is presented.

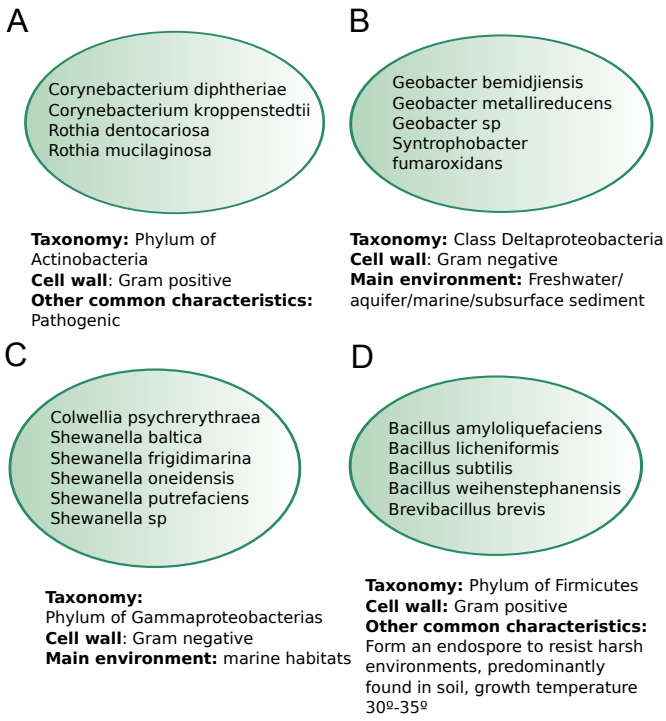

Figure 3.9: **Representative clusters obtained in the second stage of the clustering with the ES.** Each circle represents a cluster. The names of the species clustered together are displayed inside each cluster. Below each cluster, information about common phenotypic characteristics among the species of the cluster is presented.

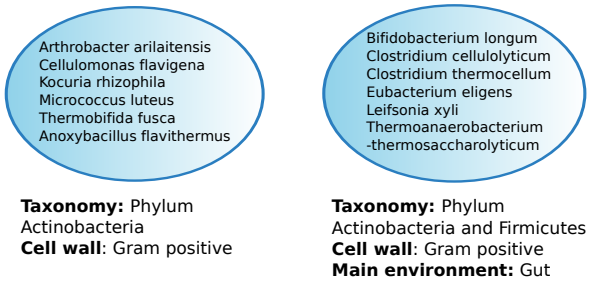

Figure 3.10: **Representative clusters obtained in the third stage of the clustering with the ES.** Each circle represents a cluster. The names of the species clustered together are displayed inside each cluster. Below each cluster, information about common phenotypic characteristics among the species of the cluster is presented.

Chapter 4

Novel data mining approach based on unsupervised classification for the analysis of protein expression data

4.1. Introduction

Depending on cells or organisms needs and the activities they carry out at each time, specific genes are expressed. This expression is the consequence of gene regulation networks that activate or inhibit certain genes based on external or internal signals to the cell or organism. Gene expression is the process by which the information in the genes is translated into functional specific molecules. The central dogma of molecular biology is based on this concept and establish that the DNA of the genes is transcribed into messenger RNA and this is later translated into proteins that carry out specific functions. As explained in Chapter 1 , the regulation of metabolic networks can take place at two levels, at enzymatic level, as was the case of the work described in chapter 2 and at expression level. In a very simple way, depending on the expression or not expression of certain genes, the formation or synthesis of the corresponding proteins will be activated or inhibited, and therefore the performance of the metabolic networks in which they take part will differ. This synthesis can occur at a bigger or smaller scale, producing more or less amount of proteins depending on the task they need to perform. This is what is called protein expression and it is measured in a specific region of a tissue. It provides information about the levels of abundance of a specific protein in that region in the moment of measurement. Proteins are vital parts of living organisms, they are involved in many different and complex tasks and in some cases their function is unknown. Also many diseases

are related to an anomalous level of expression of certain proteins.

In this chapter we describe the design of a novel data mining computational approach to analyze protein expression data from healthy and Down syndrome mice. This chapter constitutes a functional study of metabolism by identifying critical proteins involved in high-level behaviors like learning or memory.

4.2. Learning and memory deficits in Down syndrome

Down syndrome (DS), also called Trisomy 21, is the most common genetic cause of learning/memory deficits (Irving et al., 2008). It is due to an extra copy of the long arm of human chromosome 21 (Hsa21) and the consequent increased level of expression of some subset of the genes it encodes(Wiseman et al., 2009). Although no pharmacotherapies for learning deficits in DS are available, because the incidence is high, one in 1000 live births worldwide (Irving et al., 2008), there is considerable interest in their identification. DS is a genetic perturbation of considerable complexity. Hsa21 encodes >500 genes/gene models (Sturgeon and Gardiner, 2011) and an unknown subset of these contribute to the learning deficits. Functional information is available for fewer than half of Hsa21 genes, and even for these information is limited. However it is known that overexpression of these genes, as is predicted to occur in DS, will perturb many different biological processes and pathways, including many affecting brain development and function. Because of its complexity, for choosing drug targets, it is logical to look for perturbations in pathways that are critical to learning and memory and then to consider drugs that would correct the observed perturbations.

For preclinical evaluation of drug effects, multiple mouse models of DS, each carrying an extra copy of a subset of Hsa21 orthologous genes, have been created (reviewed in (Rueda et al., 2012)). Katheleen Gardiner and her research group recently investigated normal responses to learning. They experimentally measured levels of expression of 80 proteins in brain regions of healthy mice exposed to context fear conditioning (CFC) (Ahmed et al., 2014), a task commonly used to assess associative learning (Radulovic et al., 1998). In those mice, approximately half the proteins responded to learning in CFC in at least one fraction/brain region. They also examined the effects of memantine on protein expression, with and without stimulation to learn in CFC. Memantine is currently in use for treatment of moderate to severe Alzheimer's Disease (AD) (Olivares et al., 2012) and has been proposed for treatment of learning deficits in DS (Boada et al., 2012; Costa et al., 2008). Little is known about the effects of memantine on protein expression, either alone or with learning paradigms. However, it is known that treatment with memantine does not affect learning in control (healthy) mice (Costa et

al., 2008) but does alter initial protein profiles and modulates molecular responses to CFC (Ahmed et al., 2014).

They have now completed a similar analysis of protein responses in the partial trisomy[1] mouse model of DS, Ts65Dn. They observed that untreated trisomy mice fail to learn in CFC but if they were first injected with memantine, they learnt successfully, i.e., learning is rescued (Costa et al., 2008). Comparing protein profiles between trisomy mice when they fail in CFC, trisomy mice when their learning is rescued with memantine and similarly treated control mice, revealed statistically significant changes in protein levels associated with normal, failed and rescued learning, changes in protein levels caused by memantine treatment alone, and differences in responses between control and trisomic mice.

The complexities of these new data are novel and they provide a more realistic view of the molecular consequences of learning than is seen in the more common types of studies that examine effects of single gene mutations and measure the levels of fewer than 5-10 proteins. The standard statistical analyses employed in the Ts65Dn CFC studies do not, however, identify several important features, for example, which of the changes seen in control mice are required for successful learning, which of the abnormalities in the Ts65Dn directly contribute to failed learning and which are compensatory responses to the perturbations caused by trisomy. Lastly, which changes induced by memantine are critical for rescuing successful learning in the Ts65Dn.

To begin to answer these types of questions, we have designed a novel data mining approach based on the unsupervised classification technique of Self-Organizing Maps (SOM) (Kohonen, 1982). The approach is used to analyze a fraction of the experimental protein expression data obtained in Gardiner's lab from eight different classes of control and trisomy mice exposed to CFC. Dealing with experimental biological data is often a complex task, due to many reasons. First, problems during the experiment can arise and they result in incomplete data. However, such experiments are usually very expensive and time consuming, therefore there is a great need of handling this problem in order to make the most out of the available data. Also the nature of data and the novelty of questions researchers want to answer often require the design of new data mining approaches.

Our SOM based data mining approach faces these problems and is specially designed to accomplish three main goals. Firstly, to find out if mice can be automatically clustered together with other mice of the same class based on similarities on their protein expression levels. Secondly, identify the most critical proteins that discriminate between the different classes. Lastly, being able to discover underlying knowledge in data that can be useful to

[1]Down syndrome is also known as Trisomy 21; consequently we will often refer to Down syndrome mice as trisomy or trisomic mice. Several mouse models of DS have been developed however the Ts65Dn is the most commonly used.

understand the differences in terms of expression levels of the diverse types of learning in control and Down syndrome mice.

SOM was first used to cluster control and trisomic mice. As mentioned in Chapter 3, SOM clusters multidimensional data into a two-dimensional map of neurons where each data element is clustered in one neuron. In this work we use several functionalities of SOM not previously exploited in chapter 3, where an expert system based on SOM was developed. These functionalities are the preservation of topology, the visualization of the map of neurons and the possibility of labeling the map with information different to the one used to cluster the data. Preservation of topology in SOM means that close data elements in the input space are clustered in close neurons of the map. This can be very helpful in order to identify regions of neurons in the map that gather similar data elements, in this case mice with similar protein expression levels.

Once mice have been clustered with SOM according to their similarities in protein expression levels, the map of neurons was labeled by color-coding each neuron by the majority class of mice grouped in it. This was performed in order to identify and visualize the degree of separation and/or overlap of classes of mice in the map with similar protein profiles and that exhibited successful learning, failed learning or no learning, and responses to treatment with memantine. Afterwards groups of neurons, called clusters in this work, that contained the same class of mice were selected. Then the Wilcoxon rank-sum test was used to identify the subsets of proteins that differed significantly between clusters/classes. Additional runs of SOM were carried out to validate the subsets of proteins most critical to the separations between classes of mice. The goal was to evaluate if the separation of the different classes of mice could be performed with the reduced subsets of proteins found.

Using protein expression data obtained from the nuclear fraction of cortex in control and trisomy mice, we show that this approach: 1) helps to corroborate that the protein expression profiles used effectively describe the different classes of mice by automatically clustering mice of the same class in close regions of the SOM. 2) it manages to identify subsets of proteins that are most significant in discriminating the different classes. Therefore, as will be explained later, proteins will be found that discriminate between successful learning and no stimulation to learn, between failed and rescued learning, and between memantine treatment and saline treatment in trisomy mice. 3) the possibility that offers SOM of visualizing and labeling the resulting clustering represents a novel approach to interpret this kind of data and makes possible the extraction of underlying information such as biological patterns in the data important to corroborate or to build new hypothesis.

In short, the proposed approach describes the following achievements:

1. Clustering the different classes of control and trisomy mice based on their similarities in protein expression levels $\longrightarrow$ Corroborates that

protein expression levels describe the different kinds of learning.

2. Automatic labeling of SOM $\longrightarrow$ Visualization of the structure of the data and identification of informative biological patterns.

3. Identification of discriminant subsets of proteins between classes of mice $\longrightarrow$ Helps in the detection of protein abnormalities in DS mice and proteins that need to be altered by drug treatments to facilitate the rescue of learning deficits.

4. Validation of results repeating clusterings with SOM using the subsets of proteins found as features $\longrightarrow$ Corroborates the biological relevance of the subsets of proteins found.

The results suggest that the novel data mining approach applied to additional datasets can help to identify those protein abnormalities in DS mice that most critically need to be altered by drug treatments to facilitate the rescue of learning impairments. In the following sections the dataset, the experiment and techniques by which the dataset was obtained and the design of the proposed approach will be described in depth.

4.3. Protein samples and groups of mice

All mice, protein samples and protein expression levels have been reported previously (Ahmed et al., 2014). Briefly, in the experiment were used brains of 3 month old male Ts65Dn Down syndrome model mice and their male littermate wild type controls, after training in context fear conditioning (CFC) with and without injection with the drug memantine. The basic CFC protocol requires two groups of mice (Fanselow, 1990). The context-shock (CS) group, stimulated to learn, are placed in a novel cage, allowed to explore for several minutes and then given a brief electric shock; normal mice learn to associate the novel context with the aversive stimulus and will freeze/paralyze upon re-exposure to the same cage. This behavior shows that mice recognize the context and associate it with the shock. To control for the effects of the shock alone, a second group of mice called the shock-context (SC) group, not stimulated to learn, are placed in the novel cage, immediately given the electric shock and then allowed to explore; with these conditions, normal mice do not learn to associate the novel cage with the shock and do not freeze upon re-exposure to the same cage (Fanselow, 1990). Unlike, the trisomy CS group of mice fail to learn and do not freeze; this learning impairment can be corrected, however, if they are injected with memantine prior to training (Costa et al., 2008). To control for the effects of injection alone, an additional CS group injected with saline (no drug) must be included and to control for effects of injection and memantine alone, separate SC groups must also be injected with saline and with memantine.

Control mice	**Learning**	**#Mice**	**#Measurements**
c-SC-s	No	9	135
c-SC-m	No	10	150
c-CS-s	Normal	9	135
c-CS-m	Normal	10	150
Total control		38	570
Trisomic mice			
t-SC-s	No	9	135
t-SC-m	No	9	135
t-CS-s	Failed	7	105
t-CS-m	Rescued	9	135
Total trisomic		34	510
Totals		72	1080

Table 4.1: **Learning outcome in CFC, numbers of mice and measurements in each class of mice.** From each mouse 15 measurements of the 77 proteins were registered. SC, shock-context; CS, context-shock; c, control mice; t, trisomic mice; s, saline injected; m, memantine injected.

Thus, with a drug treatment included in CFC, four groups of trisomy were required: CS-saline, CS-memantine, SC-saline and SC-memantine. The same four groups of control littermates were also required. Thus, a total of eight groups of mice were generated and protein expression levels were measured for each. Names of the eight classes are simplified for clarity in 4 letters and separated by dashes: first one indicates genotype: c for control and t for trisomy, the second and third indicate stimulation or not stimulation to learn: CS for context-shock and SC for shock-context and the last one indicates type of treatment: m for memantine and s for saline. Therefore, a trisomy mouse not stimulated to learn injected with saline will be assigned to the class t-SC-s. Figure 4.1A summarizes the classes of mice according to the CFC experiment and Table 4.1 lists the learning outcomes and the numbers of mice in each group.

4.4. Dataset of expression levels of relevant proteins

The dataset used here consists of the expression levels of 77 proteins that produced detectable signals in the nuclear fraction of cortex, from a total of 72 mice (7-10 mice in each of the eight groups; see Table 4.1). Measurements were made using reverse phase protein arrays (RPPA), a high throughput

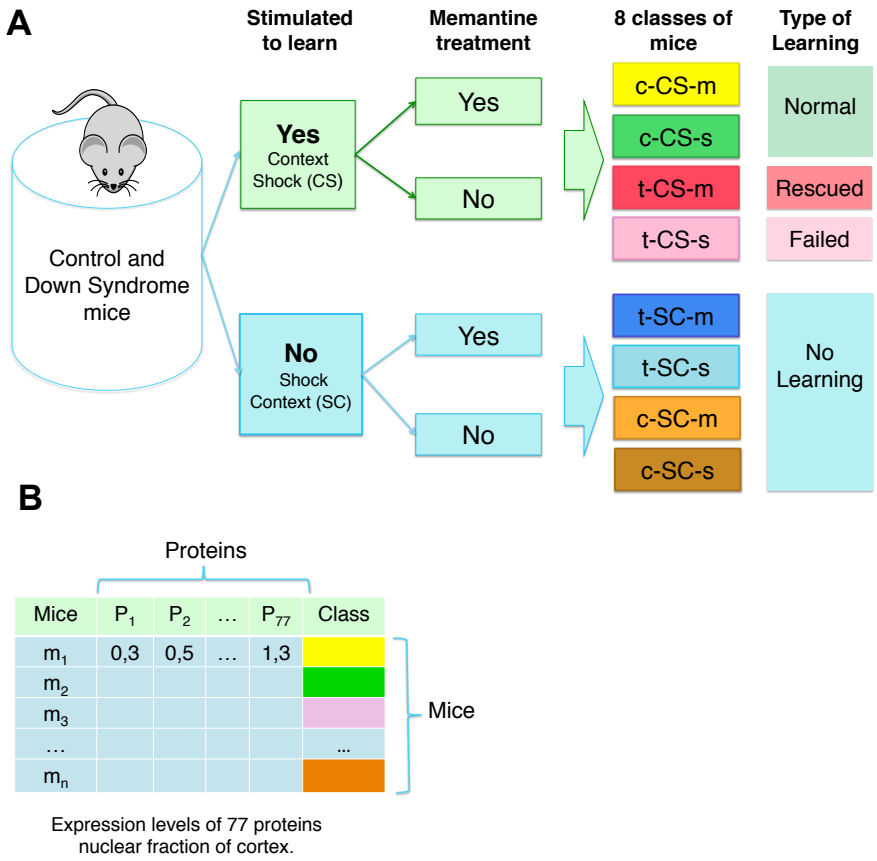

Figure 4.1: **A. Explanatory diagram of the classes of mice and types of learning in CFC experiment.** The eight classes of control mice (yellow:c-CS-m, green:c-CS-s, orange:c-SC-m, brown:c-SC-s) and trisomic mice (red:t-CS-m, pink:t-CS-s, dark blue:t-SC-m, light blue:t-SC-s) and their corresponding learning response. **B. Matrix representing the dataset.** Rows represent mice and columns the values of expression level of the 77 proteins measured in each mouse. Last column (color coded) indicates the class of each mouse.

technique in which protein samples from individual mice are robotically spotted onto nitrocellulose-coated microscope slides (Nishizuka et al., 2003). For experiments here, a single slide contained 20 spots per sample: three replicates of a five-point dilution series, plus replicate buffer controls, i.e., 15 measurements of protein level per sample, from the 72 samples, for a total of 1460 spots. RPPA is highly sensitive and reproducible (Ahmed et al., 2012)) but technical artifacts can occur which require the elimination of data from

individual spots. Similar to other high throughput techniques, it is not possible to repeat experiments for individual measurements and therefore the final dataset contains missing values, i.e., there were <15 measurements for a small number of samples/proteins measured (see below). Protein expression data were generated from additional subcellular fractions from both cortex and hippocampus of the same mice. However, the cortex nuclear fraction was chosen for use here because it was the most complete of the datasets.

4.5. SOM based data mining approach

In the work described in this chapter we count with expression levels of 77 proteins from eight different classes of control and trisomy mice. The 77 proteins were selected by experts because of their relevance in learning and memory. However it is unknown if the proteins and their expression levels can be used to describe the different kinds of learning associated with the mice classes present in the data. The first goal of this work is to answer that question: Are the levels of expression of those 77 proteins sufficient to describe and discriminate the different classes of mice? And second, are there subsets of proteins among those 77 more critical for the discrimination of the different classes? We have designed a data mining approach based on the machine learning clustering method Self Organizing Maps (SOM) and the statistical Wilcoxon Rank-Sum test to answer these questions.

As briefly explained in section 4.2, SOM is used here to cluster the data from the eight classes of mice. SOM has several advantages over other existing clustering methods that make it the best option for these data. They will be discussed in depth along the next sections but the two main are the following: first, it does not require the user to specify the number of classes present in the dataset, like for instance in k-means. This unsupervised characteristic is helpful in our case because, it is not our intention to force the data to be clustered in a fix artificial number of classes, but to observe the natural structure of data. Second, the result of clustering multidimensional data (in our case 77 proteins or 77 dimensions) is projected into a two-dimensional map of neurons preserving the topological properties of the input space, which provides a big help in the interpretation of results and knowledge discovery.

The proposed strategy is hierarchical, and it can be summed up in 4 steps: 1) preprocessing of the data, 2) determination of the optimal size of SOM, clustering and labeling, 3) identification of class-specific clusters of neurons and use of the Wilcoxon Rank-Sum test to identify proteins that discriminate between classes, and 4) validation of results by repeating SOM clustering with different subsets of proteins as input. Note that the class information (about the 8 mice classes) was not used by SOM algorithm; it was used only later to label the neurons of the map. Figure 4.2 shows a

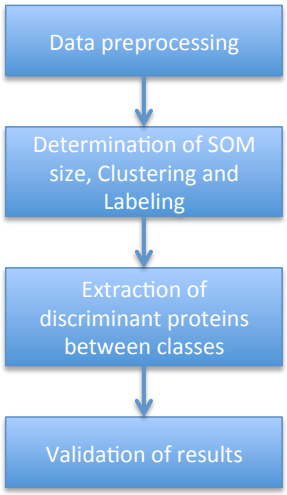

Figure 4.2: Scheme of the four phases of the approach designed for the analysis of the dataset.

scheme of the approach.

4.5.1. Data preprocessing

Data for seven of 77 proteins had one or more mice with missing values. Instead of removing these proteins, we replaced the missing values with the average value of the expression level of that protein in the same class of mice. One mouse in the t-CS-m (trisomy-CS-memantine) group had missing values for the majority of proteins and values that were very different from other mice of its same class. We therefore considered this mouse an outlier and removed all its registered data.

Application of SOM requires that all features (protein expression levels) have a similar range of values. If some proteins have values in the range 0-3 and others between 0 and 0.6 (as is the case in our data), then the proteins with higher values would have more influence on the clustering outcome, possibly leading to erroneous results of clustering. Thus all measurements (comprising a matrix where samples/mice measurements are the rows and proteins are the columns) are normalized to the range 0-1, column by column, using Equation 4.5.1

$$Normalized(e_{ij}) = \frac{e_{ij} - E_{j,min}}{E_{j,max} - E_{j,min}} \tag{4.1}$$

where e_{ij} is the value of expression level of mouse i for protein j. $E_{j,min}$ and $E_{j,max}$ the minimum and maximum of all the values of protein j (min

and max of the column). SOM clusters mice measurements using protein expression profiles as features.

4.5.2. Determination of SOM size, clustering and labeling

SOM allows visualization of the clustering result in the 2D space. Neurons in the SOM are arranged in a grid of n rows by n columns of neurons. Because SOM preserves the topology of the original data in the projected 2D space, by means of a neighborhood function, it groups data items within neurons (over the entire 2D map) so that those close in the original space are also close in the projected 2D space. Here, mice will be clustered in neurons according to similarities in their protein levels. Hence, if protein levels differentiate between classes, we expect to find two or more zones on the map, e.g., one zone containing c-CS-s (control-CS-saline) mice that learn and another containing c-SC-s (control-SC-saline) mice that do not learn.

Application of SOM requires estimate of the size of the network, i.e., deciding the number of neurons into which the input data are to be clustered. No general rule still exists for the automatic selection of the number of neurons in SOM. Depending on the data set and the purpose of the analysis, the number of neurons that best fits the data can vary. In some cases it is interesting to have the data elements very disperse over the map and in other occasions it is more interesting to have them arranged in compact groups. Another possibility is the use of clustering validity indices, as performed in the problem of the prokaryotic species, described in Chapter 3. This is particularly useful when there exist no a priori knowledge of the number of classes in the dataset, however sometimes these indices do not give sufficient information when applied to certain data sets. In the case of the protein expression dataset we select the number of neurons based on data properties. As explained in the Section 4.4 and Table 4.1, 15 measurements were registered for each of the total 72 mice under consideration. We selected a number of neurons that allowed each mouse (all of its 15 measurements) to be clustered within a single neuron. This assured that if there were no clustering structure in the data no artificial grouping of mice would be imposed (which would be the case if the number of neurons was small). Thus, for 570 measurements corresponding to 38 control mice, the chosen SOM size is 7x7. A smaller size (e.g., 6x6 or 5x5) might overly compact the data, possibly forcing clustering of measurements from more than one class of mice into a single neuron. A larger size (e.g., 8x8 or 9x9) might prevent identification of true clusters (classes when found clusters are labeled). For the 510 measurements corresponding to the 34 trisomic mice, a SOM size of 6x6 was chosen.

SOM was implemented using the neural network toolbox of Matlab (R2011b). SOM starts by assigning random weights to the neurons. Consequently, different runs produce slightly different clustering results. To choose the optimal

SOM, we use the average quantization error, ε_q, defined in Equation 4.2. It measures how well the data samples are adapted to the resulting map. x_i represents one data sample, m_{ci} represents the weight vector of the neuron in which sample x_i is clustered (its best matching unit) and n is the number of samples. The ecuation calculates the distance of each sample to the weight vector of its best matching unit. The lower the value of ε_q, the closer are data elements in average to their best matching neuron. The value of ε_q was calculated for ten runs and the SOM with the minimum value was selected.

$$\varepsilon_q = \frac{1}{n} \sum_{i=1}^{n} \parallel \mathbf{x_i} - \mathbf{m_{ci}} \parallel \tag{4.2}$$

Ideally, mice from different classes would be separated into different clusters (represented by groups of neurons of the SOM map). If SOMs from different runs produced identical minimum values of the average quantization error, then the SOM with the smallest number of neurons that grouped mice from different classes was selected. Visualizing clustering in the 2D space of the SOM facilitates biological interpretation. We further added a functionality to SOM in order to automatically label the resulting map by first, color-coding each neuron according to the majority class of its content. Second, assigning each neuron a text label with the name of that majority class (of mice clustered in it) and third, assigning a number corresponding to the number of measurements clustered in the neuron. Note that class information is not used during SOM clustering but only at the time of labeling and visualizing the results.

4.5.3. Identification of clusters and class-discriminant proteins

The distribution of the classes of mice (Table 4.1) within the final SOM map is used to evaluate the validity of the SOM approach to this specific problem. If the SOM clusters majority of mice from the same class in one or several adjacent neurons, thus forming large class-specific clusters, this suggests that the protein levels discriminate between classes. For our purpose, we define a cluster as: (i) two or more adjacent neurons that contain mice of the same class and no other classes, or (ii) a single neuron that contains >= 80 % of the measurements of one mouse (>=12 of 15) and no measurements from any other class. For identification of clusters, we do not consider neurons that group mice from different classes.

Each class of mice (represented by a group of neurons) has an associated set of weight vectors. The weight vector of a neuron is a vector of the same dimension (i.e., 77, the number of proteins measured) as the input data and represents the mice clustered in it. In order to identify proteins whose levels are significantly different between classes we use the Wilcoxon rank-sum test,

often used in gene expression analysis (Muranen et al., 2011; Rogers et al., 2012; Rosin et al., 2012; Tanaka et al., 2010; Varras et al., 2012). Here, this test was used to compare the weight vectors of the neurons defining clusters of two classes. For example, if Class A mice are found in one large cluster, C_A, composed of 15 neurons, and Class B mice are found in another large cluster, C_B, composed of 12 neurons, and P_1 is a protein of interest, then a set of values C_AP_1 contains the 15 values of the weight vectors of C_A for P_1 and a set of values, C_BP_1, contains the 12 values of the weight vectors of C_B for P_1. The Wilcoxon test was run 77 times, once with the values of each protein, for comparing Classes A and B as follows:

$$\begin{gathered} Wilcoxon\_test(C_1P_1, C_2P_1) = 0{,}001 \\ Wilcoxon\_test(C_1P_2, C_2P_2) = 0{,}356 \\ . \\ . \\ . \\ Wilcoxon\_test(C_1P_{77}, C_2P_{77}) = 0{,}003 \end{gathered}$$

Wilcoxon test returns a p-value; proteins with p $<$0.05 were considered to be significantly different between the two classes and the set of such proteins therefore discriminates between the two classes. The reason for using weight vectors is because they represent characteristics of the measurements clustered in each neuron, keeping the most characteristic values of expression levels of each class. Given the eight classes of mice, there are 28 possible pairwise comparisons but only a subset of these are biologically meaningful. Table 4.2 lists the comparisons performed and their biological relevance.

4.5.4. Validation of results

Once discriminant proteins were identified between classes, additional SOM clusterings were generated with different subsets of proteins as features, including or excluding discriminant proteins. If the reduced subsets of proteins are indeed critical to class discrimination, then including them in the dataset should improve (or remain the same) the quality of the clustering and excluding them, should deteriorate it. Also, if SOM manages to separate classes of mice using only the discriminant subsets of proteins, it will indicate that those proteins are sufficient to distinguish the classes and therefore those proteins can be consider critical to learning in the different classes of mice. To identify the proteins that are critical to learning, we further identified the intersection between different subsets of proteins that discriminate successful learning in control mice with and without memantine and failed vs. rescued learning in the trisomy mice.

Groups	**Biological Interpretation**
c-CS-s vs. c-SC-s	Effects of CFC training in saline injected controls (normal learning, NL)
c-CS-m vs. c-CS-m	Effects of CFC training in memantine injected controls (normal learning, NLm)
c-SC-m vs. c-SC-s	Effects of memantine on control baseline
c-CS-m vs. CS-s	Effects of memantine on control final conditions (normal learning +/- memantine)
t-CS-s vs. t-SC-s	Effects of CFC training in saline injected Ts65Dn (failed learning, FL)
t-CS-m vs t-SC-m	Effects of CFC training in memantine-injected Ts65Dn (rescued learning, RL)
t-SC-m vs. t-SC-s	Effects of memantine on trisomy baseline
t-CS-m vs. t-CS-s	Effects of memantine on Ts65Dn final conditions (RL vs. FL)
t-SC-s vs. c-SC-s	Initial trisomy vs. control differences

Table 4.2: Group comparisons and biological relevance

4.6. SOM based approach applied to mice data

The SOM based approach was used to cluster protein expression data from the eight classes of mice (four classes of control and four classes of trisomy mice). For both control and trisomy, two groups of mice were stimulated to learn, injected with either saline or memantine, and two groups were not stimulated to learn, also injected either with saline or memantine. The trisomy mice injected with saline fail to learn, but learn successfully when injected with memantine, while control mice learn equally well with either saline or memantine (Costa et al., 2008). Questions of particular biological interest are: (i) can SOM correctly cluster mice into classes based on patterns of expression of the 77 proteins, (ii) can the resulting clusters and class separations be improved using subsets of the 77 proteins, and (iii) can subsets of proteins that are most critical for normal, failed and rescued learning, and memantine response be identified?

We applied the SOM based approach first to the four classes of control mice to investigate its performance with protein profiles from normal learning. We then applied it to trisomic mice to investigate failed and rescued learning, and lastly, we applied it to a mix of control and trisomic mice to investigate differences most relevant to learning impairment.

Run SOM 7x7	**Average quantization error**	**# Mixed class neurons**	**Total # of measurements in mixed class neurons**
1	0.589	10	144
2	0.585	10	140
3	0.583	9	131
4	**0.579**	**8**	**110**
5	0.591	10	153
6	0.592	11	158
7	0.581	8	126
8	0.586	9	142
9	0.594	10	143
10	0.596	9	129

Table 4.3: Average quantization error, number of mixed class neurons and total number of measurements in mixed class neurons after repeating 10 times the clustering of control mice data with a SOM 6x6.

4.6.1. Control mice data

4.6.1.1. Determination of SOM size, clustering and labeling

Because the data are composed of 570 measurements that correspond to 38 control mice available (38 mice x 15 measurements per mouse), a SOM of size 7x7 that contains 49 neurons was selected, according to the explanations in section 4.5.2. The SOM was run ten times. Table 4.3 shows the average quantization error, the number of mixed-class neurons, and the total number of measurements in mixed-class neurons for all runs. Run 4 (in bold) has the lowest average quantization error, the lowest number of mixed class neurons, and minimum number of mice in those neurons; therefore this SOM was selected.

Mice are clustered in neurons in the labeled two-dimensional SOM map (Figure 4.3). We used a hexagonal layout where each neuron had six adjacent neurons, with the exception of neurons on the outside borders. This topology is the most suitable when spatial distributions are considered because the number of adjacent neurons is greater than triangular or squared topologies. Mixed topologies such as octagonal and squared are not appropriate because of the reasons expressed above. Because SOM preserves the topology of the original data, the data elements (mice) grouped in neurons that are adjacent to each other in the map correspond to mice that are close to each other in the original 77-dimensional protein expression input space. Figure 4.3 shows that the two classes of control mice that learn successfully (c-CS-s and c-CS-m, green and yellow) are clearly separated from the two SC classes

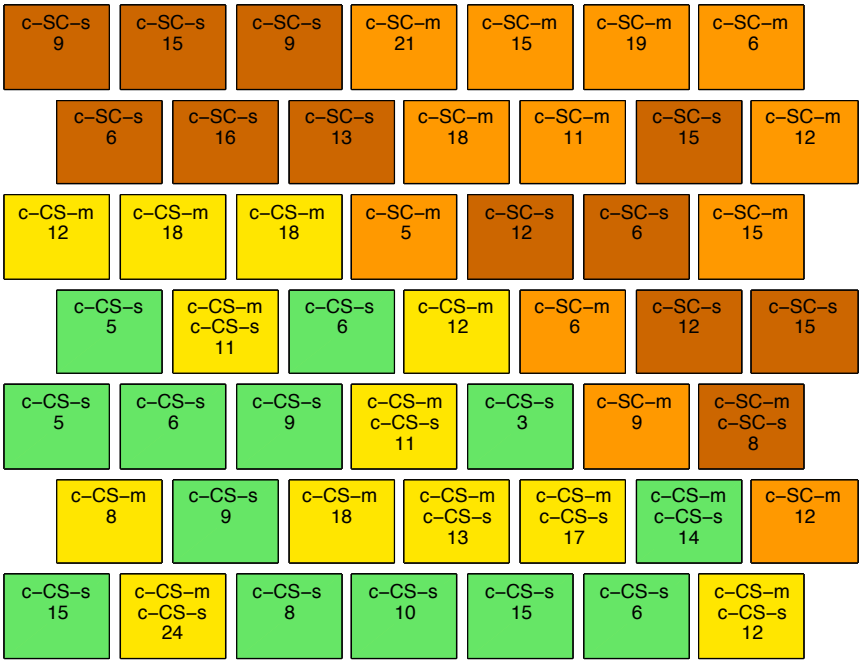

Figure 4.3: **Optimal SOM with four groups of control mice.** Neuron color indicates the majority class of the measurements clustered within it: brown, c-SC-s; orange, c-SC-m; yellow, c-CS-m, green, c-CS-s. Neurons are labeled with the name(s) of the majority and minority class(es) and the total number of measurements contained within it.

that do not learn (brown and orange). This suggests that learning in control mice is associated with distinct changes in protein expression. In relation to the number of neurons selected, it is noticeable that in many of them the total number of measurements clustered is 15, what indicates that SOM is capable of clustering the 15 measurements of the same mouse automatically. This aspect, added to the observation of clear different regions of neurons in the map that correspond to the different classes of mice are indications that the number of neurons selected is sufficiently appropriate for these data.

4.6.1.2. Identification of clusters and class-discriminant proteins

Figure 4.4 shows the same SOM as in figure 4.3 with class-specific clusters indicated, where a cluster is defined either by a group of adjacent neurons that contain measurements from the same class of mice or by a single neuron that contains a high percentage of measurements of the same type of mouse (recall that mixed class neurons are not considered here as valid cluster members, as explained in section 4.5.3). Among the two SC

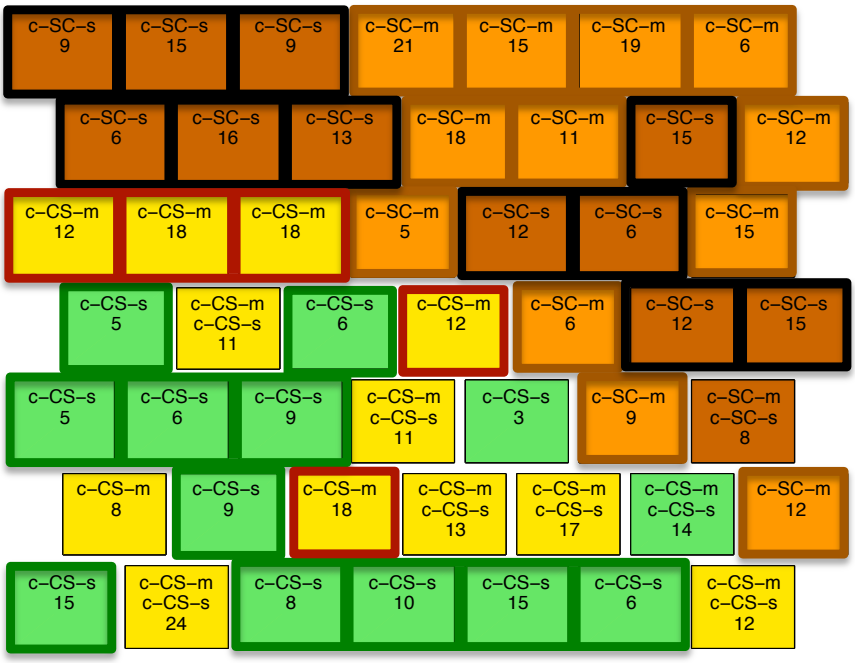

Figure 4.4: **Class-specific clusters in SOM of control mice.** Neurons forming each cluster are outlined: black, c-SC-s; brown, c-SC-m; green, c-CS-s; dark red, c-CS-m.

classes, saline treated mice (c-SC-s) form two large clusters of six and five neurons each (neurons surrounded by black line in Figure 4.4) and memantine treated mice (c-SC-m) form one large cluster of nine neurons and two additional clusters of two and a single neuron each (neurons surrounded by brown line in Figure 4.4). These clusters, plus the observation that only a single neuron mixes both c-SC-s and c-SC-m measurements, suggest that memantine alone is also associated with distinct changes in protein expression that discriminate memantine injection from saline. In contrast, the two CS classes are not so well separated into c-CS-s and c-CS-m clusters. While c-CS-s contains one large cluster of ten neurons plus a singleton (neurons surrounded in green, containing 70 % of the measurements in this class), the c-CS-m class is represented by a three neuron cluster and two single neurons that together contain only 50 % of the measurements (neurons surrounded by dark red line). Seven of the 25 CS neurons mix c-CS-s and c-CS-m. Memantine does not alter the success of learning and therefore the similarities in changes in protein levels in these CS neurons may predominantly reflect responses to learning not effects of memantine.

The Wilcoxon rank-sum test was used to compare levels of individual proteins between pairs of classes. Comparing c-CS-s vs. c-SC-s (mice stimu-

lated to learn versus mice not stimulated to learn, both treated with saline) identifies proteins that respond in normal successful learning in CFC. Thirty-one proteins were significantly changed (Table 4.4, column c1). Because these results were obtained using weight vectors, we also used the Wilcoxon test for the 31 proteins using the original values of proteins without normalization; this did not change the significance of the results (data not shown). Figure 4.5 shows boxplots of the values of c-SC-s and c-CS-s clusters for the 12 proteins with lowest p-values. Note that c-CS-s values are greater than those in c-SC-s for all of these proteins, except SOD1, IL1B and ubiquitin. The same dataset as the one used here was analyzed by Ahmed et al. (2014) using a statistical standard three-level mixed effects model plus a Bonferroni correction. In their analysis 24 of these 31 proteins were similarly identified as significantly changed, with only a single additional protein that was significantly different in (Ahmed et al., 2014) but not in our analysis.

Comparing the lists of proteins significantly changed in successful learning with saline and with memantine identified 18 proteins common to both (Table 4.4, bold, columns c1 and c2). These included BRAF, ERK and pERK, components of the MAPK signaling pathway that is well established to be critical to learning (Sweatt, 2001). Also of note are the immediate early gene protein, EGR1 and the brain-derived nerve growth factor protein, BDNF that are also relevant to learning (Liu et al., 2004; Veyrac et al., 2014). Given the interest of this study to DS, it is also notable that four Hsa21 proteins, APP, DYRK1A, ITSN1 and SOD1 are included as responding in both. Two additional comparisons reflect molecular events related to successful learning in control mice: c-CS-m vs. c-SC-s and c-CS-s vs. c-SC-m. Discriminant proteins in those two comparisons were identified (Table 4.4, c3 and c4). Taking the intersection of all four successful learning comparisons identified 11 proteins: BRAF, CaNA, CDK5, DYRK1A, GFAP, ITSN1, pERK, pGSK3B, pNUMB, S6, and SOD1. Thirteen proteins discriminate c-SC-m vs c-SC-s, indicating the effects of memantine on the initial conditions: AKT, ARC, BCL2, ELK, H3AcK18, NR1, pCAMKII, pNR1, pS6, pNUMB, pPKCG, SOD1 and ubiquitin (Table 4.4, c5). Memantine is an antagonist of the N-methyl-D-aspartate receptor (NMDAR), and it is interesting to note that among memantine effects are the NMDAR subunit NR1 and its phosphorylated form, pNR1, plus phosphorylated forms of two proteins, NUMB and CAMKII, known to interact with and modulate the activity of and signaling from the NMDAR. A second set of 12 proteins discriminates c-CS-m vs c-CS-s (Table 4.4, c6), thus identifying effects of memantine on the final protein profiles after learning: ARC, BAD, EGR1, ERBB4, H3MeK4, IL1B, nNOS, PKCA, pPKCAB, pS6, SHH, and Ubiquitin.

c1	c2	c3	c4	c5	c6
c-CS-s vs.c-SC-s	c-CS-m vs.c-SC-m	c-CS-m vs. c-SC-s	c-CS-s vs. c-SC-m	c-SC-m vs. c-SC-s	c-CS-m vs. c-CS-s
APP	ADARB1	BRAF	AMPKA	AKT	ARC
ARC	**APP**	CaNA	APP	ARC	BAD
BAD	**BDNF**	CDK5	ARC	BCL2	EGR1
BDNF	**BRAF**	DYRK1A	BAD	ELK	ERBB4
BRAF	**CaNA**	GFAP	Bcatenin	H3AcK18	H3MeK4
CaNA	**CDK5**	ITSN1	BCL2	NR1	IL1B
CDK5	**DYRK1A**	pERK	BDNF	pCAMKII	nNOS
DYRK1A	**EGR1**	pGSK3B	BRAF	pNR1	PKCA
EGR1	**ERK**	pGSK3B_ _Tyr216	CaNA	pNUMB	pPKCAB
ERBB4	**GFAP**	pNUMB	CDK5	pPKCG	pS6
ERK	**GSK3B**	pP70S6	DYRK1A	pS6	SHH
GFAP	**ITSN1**	S6	EGR1	SOD1	Ubiquitin
GSK3B	NR2A	SOD1	ELK	Ubiquitin	
H3MeK4	P3525		ERK		
IL1B	**P38**		GFAP		
ITSN1	pCAMKII		GSK3B		
nNOS	**pERK**		H3AcK18		
P38	**pGSK3B**		H3MeK4		
pERK	**pNUMB**		IL1B		
pGSK3B	**S6**		ITSN1		
PKCA	**SOD1**		nNOS		
pNUMB	TRKA		NR1		
pPKCAB	**Ubiquitin**		NR2A		
pRSK			NUMB		
pS6			P38		
PSD95			pCAMKII		
S6			pCFOS		
SHH			pERK		
SNCA			pGSK3B		
SOD1			PKCA		
Ubiquitin			pNUMB		
			pPKCAB		
			pS6		
			PSD95		
			S6		
			SNCA		
			SOD1		
			TRKA		
			Ubiquitin		

Table 4.4: Discriminant proteins found in comparisons of control mice classes. Bold indicates common proteins between c1 and c2.

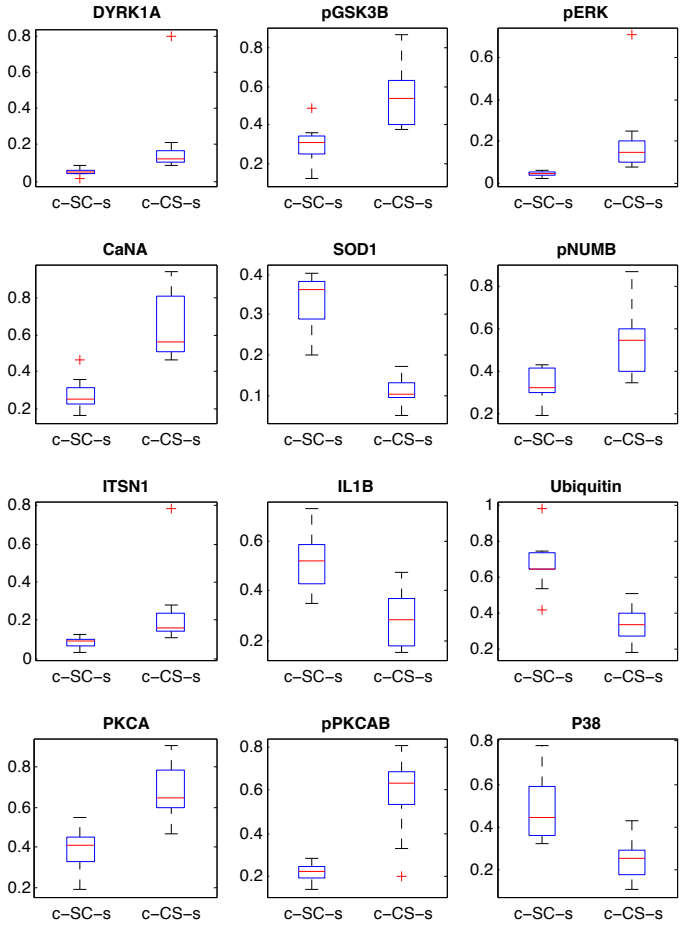

Figure 4.5: Boxplots of levels of expression of twelve proteins found discriminant between c-CS-s and c-SC-s mice. The proteins selected are the twelve with the lowest p-values obtained by Wilcoxon test.

4.6.1.3. Validation of discriminant protein subsets

The SOM described above used data from the complete set of 77 proteins. If the subsets of proteins found to discriminate class-specific clusters in this SOM are correct, then it should be possible to predict the outcome of the clustering when the input is limited to only those subsets. Figure 4.6A shows the SOM obtained using as input the 11 proteins found to be significantly different in all four comparisons related to successful learning (Table 4.4, columns c1-c4). The CS classes (yellow and green neurons) are clearly separated from the SC classes (brown and orange), indicating that proteins in this reduced subset are relevant to learning and also sufficient among the set of 77 to achieve separation. There are however differences between the

qualitative features of the SOM in Figure 4.6A and that in Figure 4.3. In Figure 4.6A, there are 12 CS mixed neurons and they contain >50 % of the measurements. This compares to only 7 CS mixed neurons containing <30 % of measurements in the original SOM in Figure 4.3. This suggests that these 11 proteins are more relevant to identifying successful learning and less relevant to discriminating memantine responses.

Figure 4.6B shows the SOM obtained using the remaining 66 proteins that were not common among the four successful learning classes. In contrast to the SOMs using the complete set of 77 and the reduced set of 11, the CS and SC groups are no longer well separated. Notably three CS (green and yellow) neurons are found in the upper left corner of the SOM, separated from the bulk of the CS neurons entirely. Two of them also contain measurements of SC mice. Similarly, there is one isolated SC neuron located within the bulk of the CS group (brown neuron on the bottom) and another neuron contains measurements from three classes, c-CS-s, c-CS-m and c-SC-s mice. These results support the critical role of the 11 common proteins: without them the discriminating clusters are lost.

To further explore the effect of memantine on learning-induced protein expression, clustering was repeated using the 11 proteins common to the four comparisons c-CS vs. c-SC (used in Figure 4.6A) plus the 12 proteins that discriminate c-CS-s and c-CS-m classes, i.e. endpoint protein expression with and without memantine. The SOM is shown in Figure 4.7. While the separation between CS and SC clusters remains clear, the separation between c-CS-s and c-CS-m has improved compared with previous SOMs. Specifically, the c-CS-s neurons form one large 9-neuron cluster plus one isolated neuron and the c-CS-m form a single 11-neuron cluster plus one isolated neuron. There are now only three mixed CS neurons and these contain only 33 measurements, a considerable reduction from the 12 mixed neurons containing 142 measurements when the 12 c-CS-s vs c-CS-m discriminating proteins were not included. These results suggest that the 12 proteins predominantly reflect memantine effects, not responses to learning. Table 4.5 summarizes the number of mixed c-CS-m and c-CS-s neurons in SOMs from Figures 4.3, 4.6A and 4.7.

Lastly, we added the 13 proteins that describe memantine effects on the initial protein profiles, c-SC-m vs c-SC-s, to the 11 that discriminate successful learning (Figure 4.8) (the total number of proteins is actually 22 not 24 because SOD1 and pNUMB are present in both groups of proteins). The SOM result revealed that the c-CS and c-SC mice remained clearly separated, but the separation between c-SC-s and c-SC-m was improved. The clusters are completely separated, with c-SC-s measurements in two clusters composed of 10 and 2 neurons and c-SC-m measurements entirely contained in a single 8 neuron cluster. In addition, there were no mixed neurons. However, there were more mixed c-CS-s and c-CS-m neurons. This reflects the

fact that some of the proteins that discriminate between c-SC-m and c-SC-s have similar values in, and do not discriminate between, c-CS-m and c-CS-s.

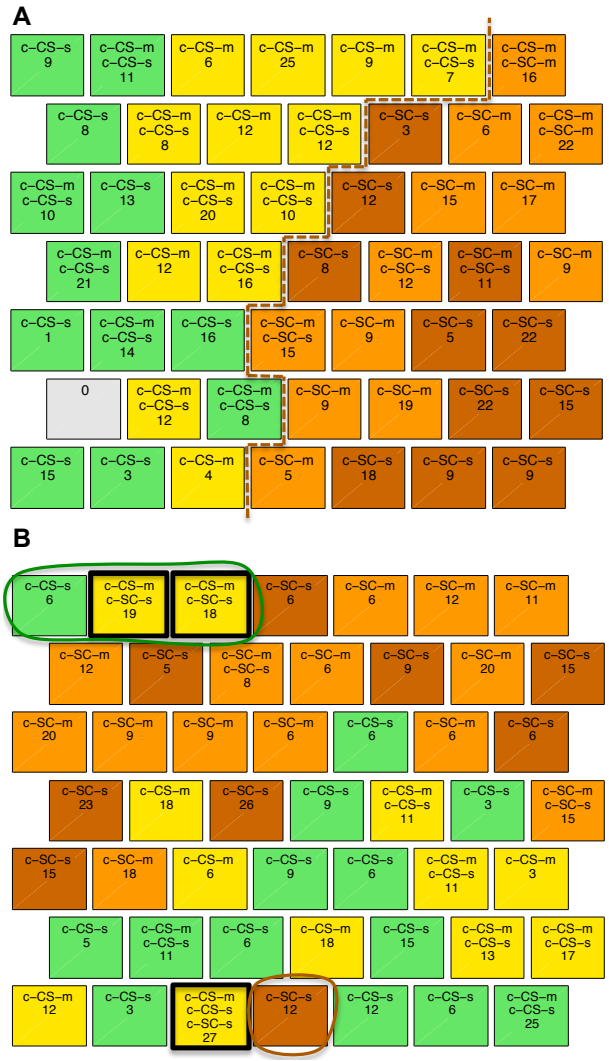

Figure 4.6: **SOM after clustering control mice with reduced subsets of proteins. A**. SOM obtained using only the 11 proteins that discriminate between the four classes of c-CS and c-SC. The dashed line indicates the border between the two main classes. **B**. SOM obtained using the remaining 66 proteins. Circled in green are neurons that contain a majority of CS mice surrounded by neurons with SC mice. Circled in brown are neurons with SC mice surrounded by neurons of CS mice.

	# Proteins used in clustering		
	77	11 (Discriminant c-CS vs. c-SC)	11 + 12 (discriminant c-CS-m vs. c-CS-s)
# neurons with mixed c-CS-m and c-CS-s	7	10	3
# measurements in mixed neurons	102	142	**33**

Table 4.5: Number of mixed c-CS-m and c-CS-s neurons and measurements.

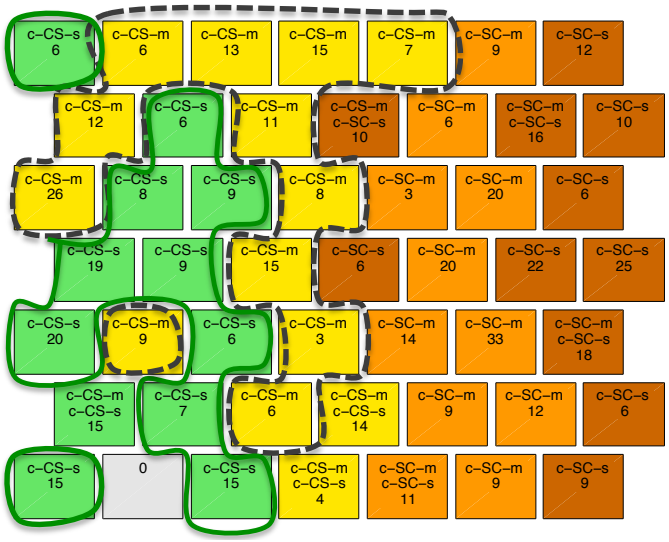

Figure 4.7: **SOM after clustering control mice with a subset of 23 protein discriminant between c-CS and c-SC and between c-CS-m and c-CS-s.** 11 discriminating proteins between context-shock and shock-context (c-CS and c-SC) plus the 12 proteins discriminating between context-shock with and without memantine (c-CS-m and c-CS-s). Dashed black line, clusters of CS-m; green, clusters of CS-s.

4.6.2. Trisomic mice data

4.6.2.1. Determination of SOM size, clustering, labeling and identification of clusters

The four classes of trisomic mice, SC and CS with saline and memantine treatment, comprised a total of 34 mice and 510 measurements. A 6X6 SOM

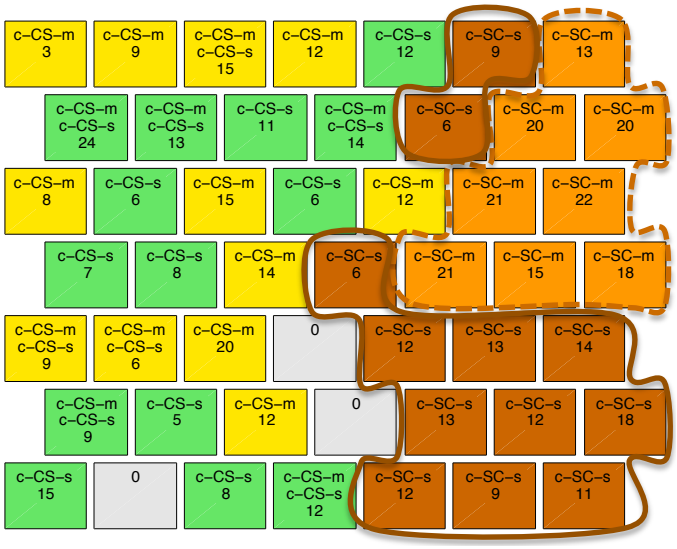

Figure 4.8: **SOM after clustering control mice with a subset of 22 protein discriminant between c-CS and c-SC and between c-SC-m and c-SC-s.** 11 learning associated proteins (c-CS vs. c-SC) plus the 13 proteins discriminating between shock-context with and without memantine (c-SC-m and c-SC-s). The total is 22 proteins instead of 24 because two proteins are found in both subsets. Brown circles, c-SC-s clusters; orange circle, c-SC-m cluster.

size was selected. Table 4.6 shows the average quantization error, the number of mixed class neurons and the total number of mice in mixed class neurons for ten runs of SOM clustering. Run 5 (bold) has the lowest values of all three measures. The optimal SOM for trisomic mice is shown in Figure 4.9. Similar to Figure 4.3 with control mice, t-SC mice are well separated from t-CS (light and dark blue vs. pink and red). In addition, mice in t-SC-s and t-SC-m classes are also completely separated, with t-SC-s mice (dark blue) in two clusters of 4 and 3 neurons, and t-SC-m (light blue) in a large compact cluster of eight neurons and a single neuron. The organization of t-CS clusters, however, is more complicated and provides interesting differences from the SOM for control mice.

In Figure 4.9, t-CS-s measurements are found in a cluster of 5 neurons (pink) and one single neuron that together contain only 64 (of 105) measurements. Unlike control mice, the t-CS-s mice fail to learn and these neurons are largely located adjacent to SC neurons, whereas, in control mice, the c-CS-s neurons (Figure 4.4) were distributed throughout the CS region. This difference suggests that protein levels in trisomy mice in failed learning are more similar to those in t-SC-s mice than they are to those in

Run SOM 6x6	**Average quantization error**	**# Mixed class neurons**	**Total # of measurements in mixed class neurons**
1	0.7	9	150
2	0.7019	7	110
3	0.7014	8	148
4	0.7053	9	150
5	**0.6981**	**5**	**84**
6	0.7015	8	115
7	0.7006	6	92
8	0.7003	7	137
9	0,7053	7	116
10	0,7181	11	191

Table 4.6: Average quantization error, number of mixed class neurons and total number of measurements in mixed class neurons after repeating 10 times the clustering of trisomic mice data with a SOM 6x6.

control mice in successful learning. This is supported by the observation of mixed SC-CS neurons. Neurons containing the t-CS-m class (red neurons), the only class in this SOM that learns successfully, form a large 6-neuron cluster plus one singleton (neurons surrounded by green line). Lastly, there are three CS-s/CS-m mixed neurons, containing 49 measurements, suggesting certain similarities between these two types of trisomy mice. This could indicate that not all responses in failed learning are incorrect and that these more closely resemble rescued learning.

4.6.2.2. Identification of class-discriminant proteins

Comparison of t-CS-s vs. t-SC-s (mice stimulated vs. not stimulated to learn both treated with saline) identifies changes in protein expression that occur when the trisomy mice fail to learn in CFC. Using the Wilcoxon rank-sum test, ten proteins were found to change significantly (Table 4.7, c1): DYRK1A, ITSN1, pERK, BRAF, SOD1, MTOR, P38, NR2B, pP70S6, and GluR4. Six of these, BRAF, DYRK1A, ITSN1, P38, pERK and SOD1, also changed in the corresponding set of control mice in normal learning, c-CS-s vs. c-SC-s (Table 4.4, c1), again indicating that some normal responses occur even in failed learning. However, changes in an additional 25 proteins that occurred in normal learning in control mice were not seen in trisomy. Such a dramatic difference, in number and identity of protein responses, is consistent with failed learning.

Comparison of t-CS-m vs. t-SC-m (mice stimulated vs. not stimulated to learn both treated with memantine) identifies the proteins that respond

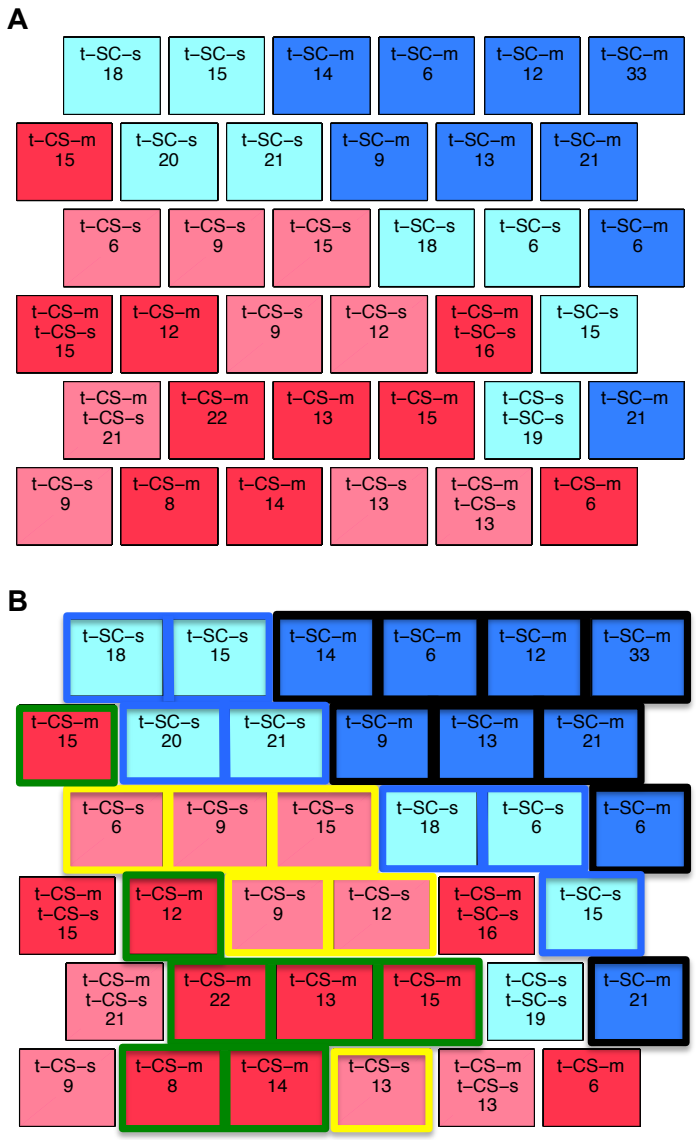

Figure 4.9: **SOM clustering with data from trisomic mice and class specific clusters** **A**. SOM after clustering trisomic mice using 77 proteins. Light blue, t-SC-s; dark blue, t-SC-m; pink, t-CS-s; red, t-CS-m. B. Class-specific clusters in trisomy mice. Neurons forming each cluster are outlined: blue, t-SC-s; black, t-SC-m; yellow, t-CS-s; green, t-CS-m.

when learning is rescued in trisomic mice with memantine. In contrast to failed learning, levels of 36 proteins are significantly different in this comparison (Table 4.7, column c2). Of these, eight were also seen in failed learning

(Table 4.7, column c1) and a total of 16 responses also occurred in control mice in normal learning (proteins in italics in Table 4.7, c2), suggesting that memantine acts to promote more, but still not all, normal responses to the stimulation of learning. A second comparison also reflecting learning, t-CS-m vs. t-SC-s, was made, identifying 20 proteins (Table 4.7, c3).

Because memantine facilitates learning in trisomy mice, it is reasonable to expect that changes produced by memantine alone, i.e. t-SC-m vs t-SC-s, include some that change the initial conditions to those more conducive to learning. Twelve proteins responded (Table 4.7, c4): pNR2A, pPKCAB, pMTOR, pP70S6, pPKCG, S6, pS6, ARC, ERBB4, P3525, SNCA, and BAD. Only three were common to memantine in control mice (Table 4.4, c4) and six also responded in normal learning (Table 4.4, c1).

The final comparison examines the differences in the protein profiles of t-CS-m vs. t-CS-s, i.e. compares the protein expression outcomes of rescued and failed learning. Only nine proteins were found (Table 4.7, c4): DYRK1A, pERK, BRAF, CDK5, RRP1, GFAP, GLUR3, P3525 and Ubiquitin.

4.6.2.3. Validation of discriminant protein subsets

To discover subsets of the 77 proteins that effectively discriminate classes of trisomic mice requires biological considerations different from those used for control mice. The trisomic mice fail to learn when injected only with saline and therefore the comparisons of t-CS-s vs. t-SC-s and t-CS-s vs. t-SC-m do not reflect changes associated with successful learning. Instead of the four comparisons used with control mice, we considered first the protein changes that were common to the two comparisons that reflected rescued learning: t-CS-m vs. t-SC-m and t-CS-m vs. t-SC-s (15 proteins; Table 4.7, c3*), plus the initial effects of memantine: t-SC-m vs. t-SC-s (12 proteins; Table4.7, c4). The SOM generated with these proteins (a total of 26) is shown in Figure 4.10A. No neurons that mix CS and SC mice occur and all the measurements of the t-SC-s and t-SC-m classes are completely separated in two different clusters with no mixed class neurons. This distribution of the data is even better than that obtained with the 77 proteins. A second SOM, shown in Figure 4.10B, was generated using the 15 proteins that discriminate between rescued learning and the lack of stimulation to learn and the 9 proteins that discriminate between rescued learning and failed learning (Table 4.7, c5) a total of 22 unique proteins. In this second clustering, there is a large cluster of t-CS-m neurons, containing 96 of 135 measurements of this type of mouse. There are also only five neurons containing only t-CS-s mice, (containing 44 of 105 measurements). Instead, most appear mixed with measurements of t-CS-m mice.

c1 (FL)	**c2 (RL)**	**c3**	**c4**	**c5**
t-CS-s vs. t-SC-s	t-CS-m vs. t-SC-m	t-CS-m vs- t-SC-s	t-SC-m vs. t-SC-s	t-CS-m vs. t-CS-s
BRAF	AKT	AcetylH3K9	ARC	BRAF
DYRK1A	AMPKA	AKT*	BAD	CDK5
GluR4	*ARC*	BAD	ERBB4	DYRK1A
ITSN1	*BRAF*	BRAF*	P3525	GFAP
MTOR	CAMKII	CaNA*	pMTOR	GluR3
NR2B	*CaNA*	DYRK1A*	pNR2A	P3525
P38	DSCR1	EGR1*	pP70S6	pERK
pERK	*DYRK1A*	H3AcK18*	pPKCAB	RRP1
pP70S6	*EGR1*	H3MeK4*	pPKCG	Ubiquitin
SOD1	*ERBB4*	ITSN1*	pS6	
	GSK3B	MTOR*	S6	
	H3AcK18	P38*	SNCA	
	H3MeK4	pERK*		
	ITSN1	pMEK*		
	MTOR	pP70S6		
	NR2B	RRP1		
	P38	SNCA*		
	pAKT	SOD1*		
	pBRAF	Tau		
	pCREB	Ubiquitin*		
	pERK			
	pGSK3B			
	pGSK3B_ _Tyr216			
	pJNK			
	pMEK			
	pMTOR			
	pNR2A			
	pNR2B			
	pS6			
	RAPTOR			
	RSK			
	S6			
	SNCA			
	SOD1			
	TIAM1			
	Ubiquitin			

Table 4.7: Discriminant proteins in four comparisons of trisomic mice. FL, failed learning; RL, rescued learning. Italics, proteins that also changed in control mice in normal learning (c-CS-s vs. c-SC-s; Table 4.4, c1) and proteins with the symbol * are proteins common to comparisons c2 and c3.

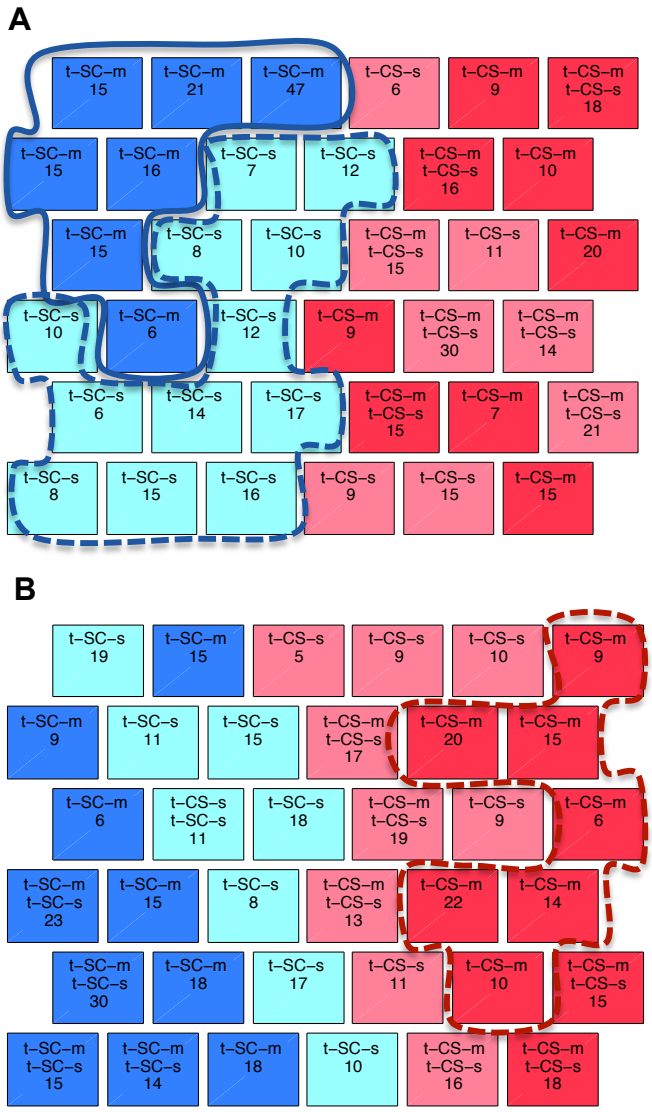

Figure 4.10: **SOM after clustering trisomic mice with reduced subsets of proteins. A.** Clustering of trisomic mice with: proteins common to the two comparisons that reflected rescued learning: t-CS-m vs. t-SC-m and t-CS-m vs. t-SC-s (15 proteins; Table 4.7, c3*), plus the proteins that reflect initial effects of memantine: t-SC-m vs. t-SC-s (12 proteins; Table4.7, c4). A total of 26 unique proteins. Cluster of neurons of t-SC-m mice are surrounded in blue and the cluster of t-SC-s mice is surrounded in dashed blue line. **B.** Clustering with the former 15 proteins (4.7, c3*) plus the 9 discriminant between t-CS-m and t-CS-s (4.7, c5). A total of 22 unique proteins.

4.6.3. Control and trisomy mice

We applied the same approach for the identification of proteins that best discriminate failed learning in trisomic mice from normal successful learning in control mice. Clustering using all eight groups of control and trisomic mice failed to produce a SOM with any clear clusters (data not shown). To search for an informative SOM, we then reduced the number of input classes to five: the three classes of successful learning (c-CS-s, c-CS-m, and t-CS-m), the one class of failed learning (t-CS-s) and one control class not stimulated to learn (c-SC-s). Together, these classes comprise 44 mice and 660 measurements. Figure 4.11 shows the resulting 7x7 SOM. The control baseline mice, c-SC-s, are found in a single large cluster of ten adjacent neurons at the top of the SOM. None of the other four classes forms a large cluster; the largest clusters include three neurons and fewer than 20 % of the measurements of a single class, and individual neurons of each class are interspersed throughout the map. Most interesting is the presence of 16 CS mixed neurons (of a total of 39), each containing measurements from 2, 3 or all 4 CS classes. While this indicates poor separation of the four classes, it is also biologically reasonable. Three classes learn successfully, and therefore similarities among subsets of their proteins are logical. Furthermore, from the trisomy SOMs described above, we also know that responses in the failed learning class do include a subset of the protein changes seen in successful learning. This latter is reflected in the presence of 10 of 16 mixed neurons that include measurements from t-CS-s mice plus one or more of the successful learning classes. Excluding the t-CS-s class from the clustering did not improve the clustering of successful learning classes (data not shown).

In an attempt to identify proteins that best discriminate failed from successful learning, we next clustered using t-CS-s, c-CS-s and c-CS-m. These classes comprise 26 mice and 390 measurements. Using a 6x6 matrix produced poor separation of clusters (data not shown). We, therefore, chose to cluster with a SOM of 8x8 to allow clustering within a wider area. As shown in Figure 4.12, t-CS-s mice are found in three clear clusters of 4-5 neurons containing 93 of the total of 105 measurements. Mice from the c-CS-s and c-CS-m are also found in significant clusters, the largest containing 9 and 14 neurons, respectively. The Wilcoxon rank-sum test identified 14 proteins significantly different between t-CS-s and c-CS-s (BDNF, pCAMKII, PKCA, pNR1, APP, MTOR, P38, AMPKA, NR2B, RAPTOR, S6, Tau, GluR3, Ubiquitin and EGR1) and 16 proteins differing in levels between t-CS-s with c-CS-m (NR2A, pNR1, APP, MTOR, P38, NR2B, RAPTOR, pGSK3B, S6, RRP1, BAX, nNOS, Tau, GFAP, GluR3, IL1B). To validate these results, we performed clustering using the ten proteins from the intersection of these two sets (APP, pNR1, NR2B, GLUR3; P38; MTOR, S6, RAPTOR; EGR1 and Tau). In the resulting SOM (Figure 4.13), the t-CS-s mice are found in a single large cluster completely separated from c-CS-s and

Figure 4.11: **SOM clustering with the 77 proteins of the four classes of control and trisomic mice stimulated to learn and one class of control not stimulated to learn.** Clustering of classes c-CS-s (green neurons), c-CS-m (yellow), t-CS-s (pink), t-CS-m (red) and c-SC-s (brown).

c-CS-m. This suggests that the levels of these ten proteins critically discriminate between failed learning in these Down syndrome mice and successful learning in controls in this task. As a further test, we used the same ten proteins to cluster rescued learning (t-CS-m) with the two classes of successful learning in control mice. As shown in Figure 4.14, the t-CS-m mice are dispersed throughout the SOM. There are four small clusters of 2-4 neurons, each containing only 15-27 measurements. Indeed, 40 % of the t-CS-m measurements are found in neurons mixed with c-CS-s or c-CS-m, and in one neuron with both. Together the SOMs in Figures 4.12, 4.13 and 4.14 suggest that abnormal responses of these ten proteins are critical to failed learning, and that memantine treatment induces changes in these responses that not only result in successful (rescued) learning, but also in a protein profile that is not distinguished from those of normal successful learning.

The last clustering with control and trisomy mice used SC classes: t-SC-s, t-SC-m, c-SC-s and c-SC-m, and the 77 proteins. Figure 4.15 shows that t-SC-s clusters are almost completely separated from the other three classes in one large cluster at the bottom and one isolated neuron of 15 measurements at the top of the map. Of 135 measurements of t-SC-s, 120

are clustered in class specific neurons and only 15 are clustered with t-SC-m (11 t-SC-s measurements) or c-SC-m (4 t-SC-s measurements). We then calculated the discriminant proteins between t-SC-s and the other classes, identifying 21 discriminant proteins: pNR2A, pPKCAB, pRSK, APP, P38, pMTOR, pP70S6, pGSK3B, pPKCG, CDK5, S6, AcetylH3K9, ARC, Tau, IL1B, P3525, SNCA, Ubiquitin, pGSK3B_Tyr216, pS6 and CaNA. Figure 4.15B shows clustering with these 21 proteins. All t-SC-s mice are present in a single cluster of adjacent neurons separated from the rest of the classes. Also the appearance of an empty neuron at the border of the cluster reinforces the dissimilarity between the three classes and t-SC-s. These 21 proteins differentiate between trisomic mice that are incapable of learning successfully and the three classes of mice that are capable of learning with stimulation.

Figure 4.12: **SOM clustering of context-shock classes of control and trisomic mice using as input the levels of all 77 proteins.** In pink neurons: class t-CS-s (failed learning), in green and yellow neurons classes c-CS-s and c-CS-m respectively (normal learning). Circled in pink, clusters of t-CS-s.

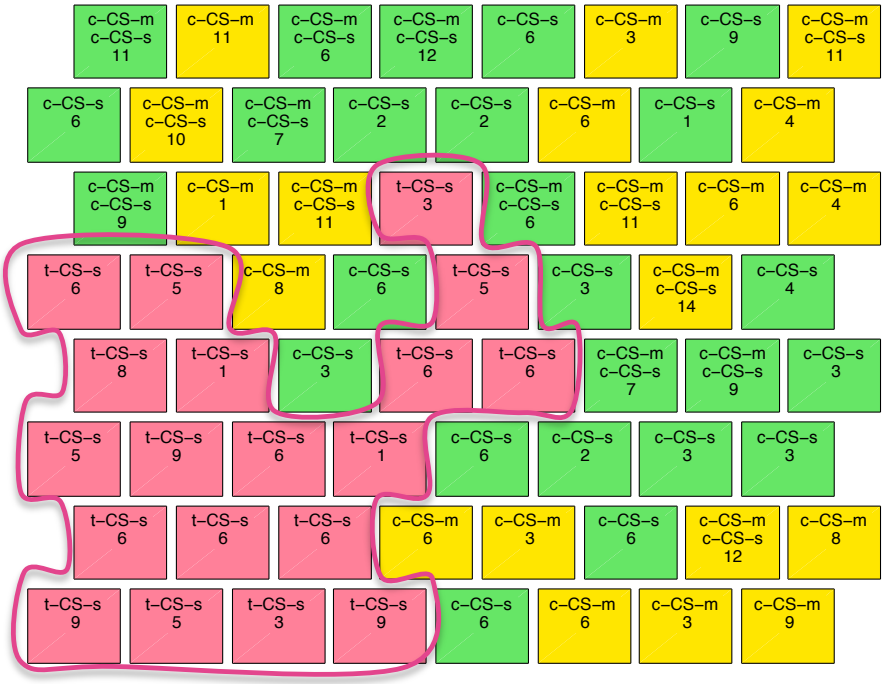

Figure 4.13: **SOM clustering of context-shock classes of control and trisomic mice using as input only the set of 10 proteins that discriminate t-CS-s from both c-CS-s and c-CS-m.** Circled in pink cluster of t-CS-s mice.

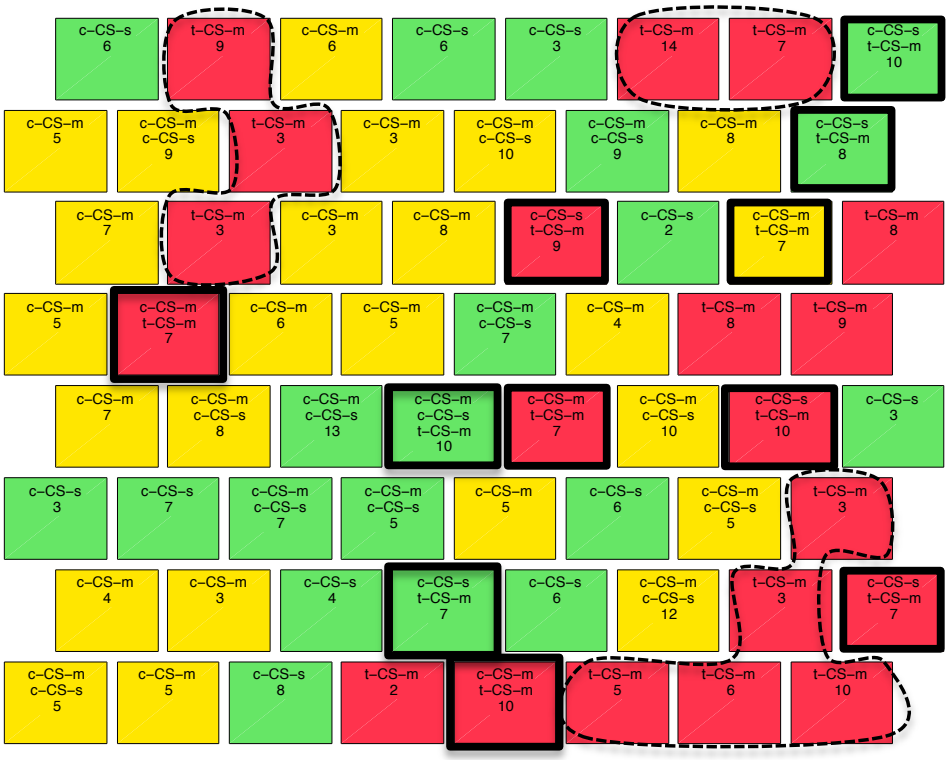

Figure 4.14: **Clustering of classes t-CS-m (red, rescued learning), c-CS-s and c-CS-m (yellow and green, normal learning) using as input the 10 proteins that discriminate t-CS-s from both c-CS-s and c-CS-m.** Dashed black line: clusters of t-CS-m mice. Black squares: neurons with mixed classes of t-CS-m and controls.

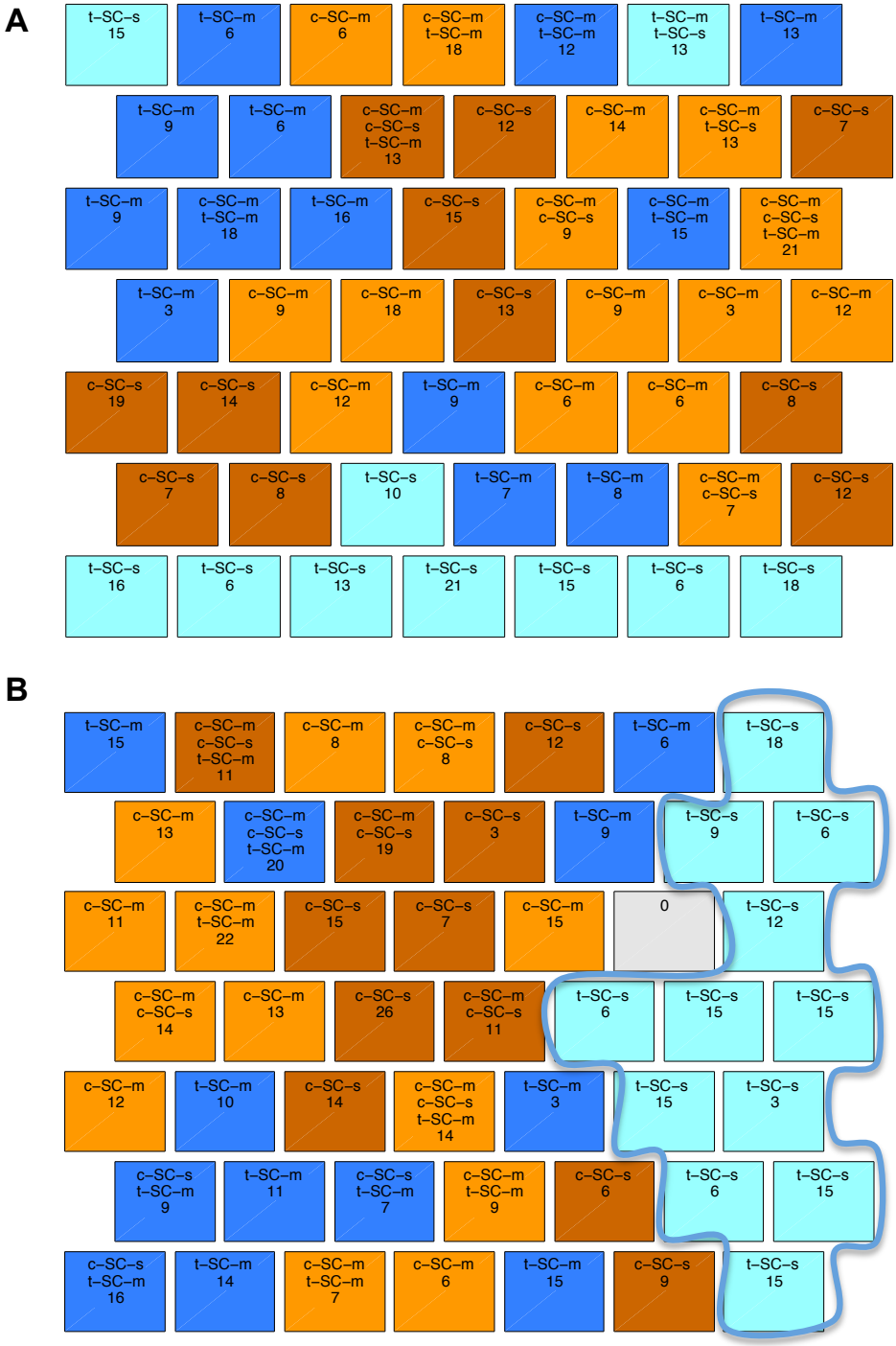

Figure 4.15: SOM after clustering shock-context classes of control and trisomic mice. **A**. Clustering of classes: t-SC-s (light blue), t-SC-m (dark blue), c-SC-m (orange), c-SC-s (brown) with 77 proteins. **B.** Clustering of the three classes with the 21 discriminant proteins found between t-SC-s and the two other classes.

4.7. Relevance of the proposed approach

Regarding the relevance of the proposed approach, it can be summarized as follows:

(a) *Complex data management*
The molecular processes underlying learning and memory are among the most complex in mammalian systems. Mutation detection in humans has shown that several hundred genes can cause intellectual disability (ID) (Ahmed et al., 2012; Gilissen et al., 2014). Genetic and pharmacological manipulations in mice have shown that several hundred additional genes can cause learning, memory and synaptic plasticity deficits. How these many genes normally work together to produce successful learning is largely unknown, yet it is an important problem to address. Intellectual disability with a genetic cause affects approximately 1 % of the world population and pharmacotherapies are largely unknown (Leonard and Wen, 2002). DS represents a significant subpopulation of ID because it has an incidence of approximately one in 1000 live births worldwide (Irving et al., 2008). DS is also particularly challenging because it is not due to a single gene mutation but to an extra copy of a normal chromosome that encodes approximately 500 genes. An unknown subset of these genes are overexpressed due to copy number and it is hypothesized that this overexpression is sufficient to perturb normal pathways and normal responses to stimulation. Understanding the perturbations and the abnormal responses to stimulation is a logical first step to understanding how drugs may rescue or be designed to rescue the abnormalities. Because of the large number of genes involved, generation of large and complex datasets is required. This in turn requires computational methods for analysis that have not previously been directed at such a problem.

(b) *Unsupervised SOM approach for data arrangement and knowledge discovery*

We have designed a data mining approach, combining SOM and the Wilcoxon statistical test, to identify biologically important changes in protein levels in mice that have been exposed to context fear conditioning. We used protein expression data generated from normal genotype control mice that learn successfully and from their trisomic littermates (Ts65Dn) that are a partial model of Down syndrome that fail learning in CFC. Memantine treatment changes the protein profiles in both controls and Ts65Dn, does not alter learning in controls, and rescues the learning impairment in Ts65Dn mice. The CFC protocol requires four groups each of control and trisomic mice, two groups (CS) stimulated

to learn in CFC with and without memantine, and two groups (SC) not stimulated to learn, with and without memantine injection. Expression levels of 77 proteins were quantitated in 7-10 mice per group and approximately one third of proteins changed in each group. Measurement of this number of proteins provides a glimpse of the true complexity of responses to the stimulation to learn in this task. A challenge is to identify subsets of these protein responses that make the most critical contributions to normal learning, to failed learning and to rescued learning. The longer term goal is to identify drugs to more specifically target abnormalities in failed learning for correction in DS. Standard statistical tests can identify proteins that differ significantly between classes of mice, but by themselves, they do not discriminate those differences that are most important to learning outcome.

The application of the strategy based on SOM to complex protein profiles has provided a novel look at the molecular consequences of normal and failed learning. Firstly, clustering different types of mice by their protein expression profile, we observed that the different classes of mice were distributed in different regions of the SOM. This allowed to corroborate that the levels of expression of the 77 proteins measured effectively discriminate between the different types of learning in control and Down syndrome mice. Secondly, reduced subsets of critical discriminant proteins between classes of mice were identified. This later result constitutes novel knowledge discovery in the present biological problem because it sheds light into the detection of which proteins are more relevant in each type of learning. Thirdly, the value of SOM was also illustrated as a tool for results validation. The later was performed by repeating the clustering using data from different subsets of the eight classes of mice and the different class-discriminant subsets of the 77 proteins.

(c) *Data validation and interpretation*
When using the discriminant subsets of proteins as SOM features, informative patterns in the resulting Self Organizing Map arise. This includes the separation of different classes of mice into non-overlapping clusters of neurons in the 2-D map. These cases indicate that the levels of the selected proteins are effectively sufficient to discriminate the different classes. It is however also informative to observe patterns where the neurons from different classes of mice are not so well separated and measurements from different classes may even appear within the same, mixed-class, neurons. These latter cases indicate that the levels of the selected proteins reflect common features shared by the different classes of mice, which, in turn, may reflect novel, common, underlying biological responses.

The designed approach was applied to three sets of data, control mice alone, trisomic mice alone and combinations of control and trisomic mice. The first SOM, using the complete dataset of 77 proteins from control mice, clearly separated c-CS mice from c-SC mice, thus showing that changes in these proteins discriminate successful learning from the lack of stimulation to learn. There are four comparisons associated with learning: c-CS-s/m vs. c-SC-s/m. The subsets of proteins that changed significantly in each of the four comparisons were identified and changes in 11 proteins were found to be common to all four. These proteins therefore formed the candidate set most critical to successful learning in control mice. Clustering with only these 11 generated a SOM that maintained the separation of CS from SC mice but also increased the number of mixed neurons. In this case, the mixed neurons suggest that these 11 proteins are more significant for discriminating CS vs. SC than for saline vs. memantine. In contrast, clustering with the remaining 66 proteins produced a SOM where even the separation of CS from SC was no longer clear. Together these results strongly support the relative biological importance to learning of the 11 proteins.

Many of these proteins have well established critical roles in learning and memory or synaptic plasticity. Mutations in three, BRAF, GFAP and DYRK1A, cause both intellectual disability in people and learning/memory/synaptic plasticity deficits in mice (Brenner et al., 2001; Courcet et al., 2012; Sarkozy et al., 2009). Mutations in five others, CDK5, ERK, GSK3B, ITSN1 and SOD1 impair learning and memory in mouse. Pharmacological studies have demonstrated that even acute inhibition of phosphorylation ERK impairs learning specifically in CFC (Fanselow, 1990). Proteins S6 and GSK3B are involved in the MTOR pathway, that is also well established with roles in learning and memory. The CANA protein is a catalytic subunit of the protein phosphatase, calcineurin. Genetic and pharmacological inhibition of calcineurin activity enhances learning and memory in mice, and overexpression of calcineurin is proposed to play a role in cognitive impairment in aging and in Alzheimer's Disease. Lastly, NUMB has not been previously associated with learning and memory, thus suggesting a novel discovery for this protein. With the exception of SOD1, expression levels of each of these proteins increased with learning, such as pERK, DYRK1A,„BRAF, ITSN1, SOD1, GSK3B,„S6 and CANA. Each of these proteins functions as one component of a multiple component pathway or a multisubunit complex. How these disparate proteins in particular may work synergistically to facilitate learning in CFC remains to be assessed with further experiments and analyses.

A similar application of SOM to the full 77 protein dataset from the trisomic mice revealed interesting differences from controls. Trisomic

mice, t-CS-s, fail to learn in CFC and this is reflected in the SOM by the location of t-CS-s neurons in closer proximity to the t-SC neurons and the presence of two neurons that mix CS and SC measurements; this is in contrast to control mice, where the c-CS-s neurons were more dispersed and were never mixed with SC. Together these SOM characteristics indicate that protein responses when trisomic mice fail to learn in CFC more closely resemble responses in mice that were not stimulated to learn than do control mice after successful learning. Consistent with this, of the ten proteins that changed significantly in failed learning, only five were common to the subset of 11 critical proteins in control mice successful learning. These are logical reflections of the failure to learn.

The analysis with the SOM based approach using data from control and trisomic mice as input features illustrated this concept further. A common set of ten proteins was identified that discriminated between t-CS-s and both c-CS-s and c-CS-m, i.e. between failed learning in trisomic mice and successful learning in control mice. These proteins included subunits of the glutamate receptors, NR1, NR2B and GluR3, and an NMDAR interacting protein APP. Also included were P38 that signals downstream of the NMDAR; and three components of the MTOR pathway. Changes in the levels of these proteins, or the lack thereof, were therefore inadequate for, or inhibitory to, learning.

(d) *SOM advantages*

The proposed approach exploits the many functionalities and characteristics of SOM to analyze this kind of data. In comparison to other clustering methods SOM presents several advantages. On one hand, it clusters the data and visualizes the result in a 2-dimensional representation of high-dimensional input space (in our case, 77-D corresponding to the 77 proteins measured), what results extremely helpful for the discovery of the unknown structure of the data and the validation of hypothesis. Also, for this special problem an extra visualization functionality has been added to SOM. A labeling function, that labels the neurons of the map using the class information, here with colors and text to identify classes of mice, and with the number of measurements grouped in each neuron. The labeling process aids biologists in seeing patterns that may underlie the biological responses, here normal, failed or rescued learning. On the second hand, because of the unsupervised nature of SOM it is not necessary to establish the number of classes a priori, like in other supervised methods such as K-means for example. Data elements are distributed along a map of neurons preserving the topology of the data. Close neurons in the map group close data elements in the input space. This is of special importance in this spe-

cific problem because it allows the observation of informative patterns such as the location of clusters of neurons of one class of mice close to other cluster of adjacent neurons of other class. This advantage allows the biologist to easily interpret similarities and dissimilarities of the different classes in terms of their proximity in the map. This fact also provides valuable information about the responses to learning in the different classes. It should be mentioned that this type of information cannot be retrieved with other clustering methods where the result is a fix set of individual clusters of elements, selected a priori.

(e) *Future improvements and challenges*

The approach designed has been adapted to the specific data of the experiment considered and its results are novel and relevant. However these data are far from complete. Only a single time point after training was assessed. Following the time course of responses will provide better understanding of exactly what fails to occur in the trisomic mice: are there failures to reach adequate levels of required proteins, or failures to reach these levels in the properly orchestrated time frame? Alternatively, are there incorrect dynamic responses that inhibit learning? Another major consideration with these datasets and use is the proteins that were selected for analysis. While each protein was selected for its known role in brain development, or learning, memory or synaptic plasticity, there are many other proteins of interest that were not measured. Reverse Phase Protein Arrays that were used for protein measurement have great advantages in scaling up throughput and requiring very small sample amounts, both of which make it possible to measure far more proteins than traditional methods of Western blots. RPPA is also sensitive and accurate. The drawback, however, is that it requires highly specific antibodies and these are not available for some proteins of interest. Antibodies that cross react with non-target proteins can still be used on Western blots where analysis can be restricted to the bands of interest and non-specific bands ignored. Because RPPA uses an array format, spurious signals cannot be similarly subtracted or ignored. The SOMs generated here therefore are affected by the specific sets of proteins used as input and measurement of different proteins could change the discriminating sets.

Successful learning in CFC requires a functional hippocampus and it is hippocampus that has been studied most often at the molecular level. Protein levels in other brain regions are clearly dynamic, however, and undoubtedly important (Zelikowsky et al., 2014). For this initial analysis, we chose to analyze the data from the nuclear fraction of cortex because the dataset was the most complete. It remains important to

extend this type of analysis to the datasets from the cytosol and membrane fractions of cortex, and to all three fractions of the hippocampus. Succesful learning in CFC requires a functional hippocampus and it is hippocampus that has been studied most often at the molecular level. The complexity of such an analysis will be significantly increased because relationships among responses in different brain regions and fractions will also need to be considered. Future work requires additional experimental datasets. Further validation of the observations regarding normal learning in control mice will require repeating CFC studies with mice from different genetic background strains (Stiedl et al., 1999). It is well known that the same mutation present on different inbred backgrounds can give rise to very different phenotypic features, in some cases eliminating any observable consequence of the mutation. It will be of interest to apply the strategy proposed to data from several genetic backgrounds where learning is normal in CFC. This might serve to identify new classes of critical proteins. For trisomic mice, important data would be protein responses after treatment with other drugs that also rescue learning in CFC. Currently there are more than 15 such drugs and they are diverse in their known targets and mechanisms of action, and thus are expected to also be diverse in at least some of their protein responses. The approach presented could be used to identify common critical protein responses, which would help to define potentially more effective targets. Additional experiments with other mouse models of DS, models that are trisomic for different sets of orthologs of Hsa21 genes, will also provide useful information to understand the set of perturbations that might arise in people with DS who are trisomic for all the genes encoded by Hsa21.

Chapter 5

Conclusions and future work

5.1. General Conclusions

In this thesis, novel computational methods and approaches based on machine learning have been proposed with the goal of providing new tools for the study of metabolism. The approaches have been applied to solve three different problems and the results, described in chapters 2, 3 and 4, represent on one hand, an advance in each of the biological areas of study, and on the other hand, advances in computer science through the novel application and development of machine learning based methods. The three strategies proposed are applied to solve biological problems, however all of them could be adapted to other problems or datasets from other fields. They can serve as inspiration to solve state of the art problems in data mining in general and also in the field of optimization.

Firstly, we showed that the application of multi-criteria optimization (less used in life sciences than in other fields like physics or engineering) can be a powerful tool in the resolution of problems that have been faced long ago with mono-objective approaches. In particular, we demonstrated that this kind of optimization provides important advantages at the time of studying the regulation of metabolic cycles and finding universal schemes of regulation. The regulation of metabolic networks in general and metabolic cycles in particular is a problem of capital importance in molecular biology. The results of the work presented in chapter 2 provide novel indications on one hand about the existence of universal patterns of regulation in this kind of systems and on the other hand that multi-criteria optimization can result in a better tool to simulate the natural process of evolution that metabolic pathways and networks may have suffered.

Secondly, two data mining strategies based on Self-organizing maps have been designed with the goal of discovering underlying knowledge from two different types of datasets. Since the first publication of the unsupervised classification method by Teuvo Kohonen 1982 there have been many

applications of the method, however as new problems arise new versions of the method, or its combination with other strategies have shown to obtain successful results. The works described in chapters 3 and 4 of this thesis are examples of this idea. In chapter 3 a novel approach was designed using SOM combined with clustering validity indices to cluster a set of prokaryotic species by their similarity in a set of metabolic features, gaining advantage over other classical and existing approaches. The system is able to clustering a complex dataset following a hierarchical strategy and help in the extraction of underlying information that may relate metabolism with phenotypic, environmental or evolutionary characteristics in prokaryotic species. In chapter 4, also a SOM based approach, this time combined with a statistical test, was designed for the analysis of protein expression data from control and Down syndrome mice exposed to context fear conditioning. The novel approach exploits the functionalities and properties of SOM to provide a novel look at the molecular consequences of normal, failed and rescued learning by identifying reduced subsets of proteins more critical in the different types of learning. Also, the approach proposed allows the identification of the differences in protein expression levels that are more important to learning outcome and that standard statistical studies employed in similar studied can not.

The thesis presented is therefore, the product of a multidisciplinary project categorized within the wide fields of bioinformatics and computational biology. It proposes artificial intelligence techniques to solve molecular biology problems where metabolism is involved. Because metabolism is present at different biological levels and can be studied from different perspectives, several specific subfields, like regulation in metabolic networks, microbial communities or protein expression, have been explored by means of machine learning approaches, conceiving an innovative and original research work.

In previous chapters the approaches and methods proposed, as well as the results and conclusions of the three problems addressed, have been exhaustively discussed. In this chapter the main contributions and conclusions of each of them are summarized according to the proposed objectives of the thesis. Furthermore new perspectives and future work in each of the three areas of research are described.

5.2. Finding an enzymatic regulation pattern of a metabolic network by means of optimization methods.

Metabolic cycles are important because they occur frecuently in metabolism and in some occasions they are involved in crucial points of metabolic control (Gilman1995). The main goal of the work described in chapter 2 con-

sisted of the search of a universal regulation pattern in a metabolic cycle. The starting hypothesis was the fact that metabolic networks and metabolic pathways have undergone a natural process of optimization through time to be how they currently are. We studied the model of a metabolic cycle previously studied by Gilman and Ross in the 90s (Gilman and Ross, 1995). The authors tried to find a unique and universal scheme of enzymatic regulation for this model by means of the estimation of a set of parameters, responsible of the dynamic behavior of the cycle. Their main goal was to find the specific values of these parameters that allow the cycle to behave optimally under varying concentration profiles of its metabolites in a certain period of time.

In this thesis a multi-criteria approach has been proposed, which consists of the simultaneous optimization of two objectives. On one hand, the correct direction of the metabolic flux (f_1) and on the other hand (f_2) a minimum energetic cost. Gilman and Ross combined these two objectives into one objective function and optimized the parameters using a genetic algorithm. According to their results they only managed to find solutions that worked optimal in some courses of concentrations but badly in the others. They called them "specialist" solutions against "generalist" solutions, that were what they were searching. In our work we simultaneously optimized both objectives obtaining a family of solutions called Pareto front (set of trade-off solutions with respect to two objective f_1 and f_2) for each course of concentrations.

In each Pareto front corresponding to each course of concentrations we observed the appearance of what is called a knee-point or knee-solution. These solutions were identified as preferred optimal solutions between the two objectives. We discovered that interchanging the different knee-points in the different courses we obtained optimal values of f1 and f2 as well. This result was a possible indication of the existence of an underlying universal regulation scheme. Repeating several times the multi-objective optimization in each course we observed a frequent pattern of regulation that was optimal in terms of optimal behavior and energetic cost in all the courses. Also it was seen that the regulation scheme found is frequent in real metabolic cycles regulated by substrate, like the conversion of fructose 6-phosphate (F6p) in fructose 2,6-biphosphate (F2,6BP) described in text books (Berg et al., 2006) or the regulation of glicolisis in liver ((Berg et al., 2006, figure 16.28).

It is important to mention that it is not possible to find these solutions systematically using a classical mono-objective approach, what was verified in the first part of the study where Gilman and Ross results were reproduced with three different mono-objective optimization methods. In this work we proved that by switching from mono- to multi-objective optimization an optimal and universal scheme of regulation for this metabolic cycle could be found. We have established the basis to other scientific experts in the field, that applying a multi-criteria approach to similar problems as the one

presented here can provide novel results and a new insight into the regulation of metabolic networks.

The multi-criteria approach used can be easily scalable to bigger metabolic networks formed by more than one regulation unit; such scalability poses no major problems than increased computational requirements. This possibility can provide a more systemic way of optimizing metabolic systems. Naturally, this strategy can be used in other contexts such as synthetic biology in order to optimally design biological circuits. Optimization methods have already been used with that purpose as reviewed in (Marchisio and Stelling, 2009). We suggest increasing the robustness and feasibility of synthetic biology designs by adopting a multi-criteria framework similar to the one presented in this thesis.

5.3. Development of an expert system based on unsupervised classification for clustering bacterial species by metabolic features

The study of microbial communities has a great importance because many natural and artificial processes are mediated by groups of bacteria rather than by isolated entities. One way of studying these communities is the search of common metabolic features among different species. This is important because not only can it be very useful for their classification, but also because it allows the identification of common functional properties that traditional methods such as the 16S rRNA are not capable to find. The study of these properties can be of great help at the time of describing the way of life of organisms or entire species. In the second work described in chapter 3 of this thesis and centered in a structural study of metabolism, an ES was designed capable of clustering a set of 365 prokaryotic species by structural metabolic features at the level of metabolic pathways. For each of the species we had information about the percentage of annotated enzymes in 114 metabolic pathways, obtained from the KEGG database. Because enzymes are a key part of the structure of metabolic pathways, the percentage of enzymes is an indication of how complete certain pathway in a species is and therefore it is possible to cluster them according to their similarities in the presence or absence of a set of metabolic pathways.

The dataset was complex and classical clustering methods like fuzzy clustering did not achieve good results. Besides, there existed the possibility of incomplete information from some species. For this reason, an ES adapted to the nature of the data and inspired by human reasoning was designed. The system is based on the unsupervised SOM clustering method and the Davies Bouldin validity index (DB) and it estimates first the optimal number of neurons to cluster the data samples (species) and afterwards it initiates

an iterative process of clustering in stages. The iterative process follows a hierarchical strategy during which relevant groups of species with similar metabolic structure are identified. This way, the system is able to successfully find groups of bacteria that are easily identifiable in the first place and discriminate others more problematic to be clustered in later steps. To perform this strategy an adapted version of DB (DB') was calculated to identify the relevant groups, which were the neurons most distant to others and with more compact elements. In order to monitor the process and the performance of the ES the partition obtained in each stage was validated with the DB validity index, making another important contribution.

The results show that the method manages to automatically cluster the different species in biologically coherent groups. Besides, when analyzing the resulting groups of species we found that those species generally shared other functional characteristics such as pathogenicity, environmental preferences or ability to form spores to resist external threads. Most of the species grouped in the same neuron belonged to the same taxonomic category, what indicates that in general the metabolic structure and therefore metabolism is related to phylogenetic proximity. However, neurons that grouped taxonomically different species were also found and those species also shared environmental preferences and other behaviors or phenotypic features. These facts make the method more valuable because it suggests that metabolism may not only be related to phylogeny but also to external conditions as can be the adaptation to the habitat where species live. This adaptation may have consisted of acquiring new metabolic pathways that their ancestors did not have or had lost. The common metabolic pathways that were critical to cluster those species together could be considered environment specific and their study can help to better understand bacterial communities and their relation to their habitat.

It is also worth mentioning that the strategy proposed could be applied to other complex datasets made up of heterogeneous or incomplete data, what gives a general validity to the method to cluster complex datasets.

Following the same line of research in the future it could be of great relevance to combine in a similar system, as the one presented here, metabolic and environmental information. This could be addressed firstly by generating a training data set composed of metabolic information, as used in our ES, plus the probability of each species to be found in certain environments. Afterwards, a system capable of predicting the probability of an unknown species to survive in certain environments only by knowing its metabolic features could be designed. This aspect can have great significance in fields such as microbial ecology, that studies the relationship between microorganisms and their environment or metagenomics that studies the genetic material extracted from environmental samples.

5.4. Development of a data mining approach based on unsupervised classification for the analysis of experimental protein expression data

Down syndrome mice is considered a genetic disorder. Patients of this syndrome suffer deficits in the learning ability and memory, high level functional activities. Due to the lack of pharmacotherapies for these deficits and the high incidence of the syndrome (one of 1000 live births) there exists much interest in the preclinical evaluation of potential drugs in mouse models.

In the third work presented in this thesis and described in chapter 4, a data mining approach to extract underlying information from protein expression data has been designed. The dataset consisted on the levels of expression of 77 proteins measured in the brains of control and Down syndrome (DS) mice after having been exposed to context fear conditioning, an experiment to measure the ability of associative learning. The protein expression levels were measured in eight different types of mice, some of them pretreated with memantine, a drug used in patients with Alzheimer to rescue the ability of learning. Control mice learnt successfully whereas DS mice failed unless they had been treated with the drug. Standard statistical analysis performed previously with this dataset did not allow to extract as much information as was desirable to give an answer to important questions such as: what changes in the level of expression are necessary for normal learning in control mice? Or what anomalies contribute to failed learning in DS mice? Or what changes induced by memantine are critical in mice that have rescued their learning ability? In order to answer all these questions we designed a machine learning approach based on Self-organizing maps (SOM) and the Wilcoxon rank-sum statistical test with three objectives: 1) determine if it was possible to automatically cluster mice based on their expression profiles of the 77 proteins into genetic (control vs. DS) or treatment (memantine vs. not memantine) clusters and 2) Identify subsets of proteins that best discriminate between classes of mice and define changes in the level of expression due to genetic or treatment causes. The strategy had to contemplate and face the problem of missing data. Because the dataset was the outcome of an experimental process some of the measurements corresponding to certain proteins could not be registered, what constituted an additional challenge.

The results obtained show that with the strategy proposed, the objectives considered have been achieved. Firstly, we observed that clustering the data with SOM we managed to separate with high accuracy the different types of mice, what indicates that the expression levels of the proteins measured effectively discriminate between the different types of learning. Secondly, using the weight vectors associated to the neurons of SOM and the Wilcoxon test we obtained subsets of discriminant proteins between classes. This was validated by repeating the clustering with the original data but only using the

subsets of discriminant proteins as features. We confirmed that in most of the cases the results were the same or better than using the 77 proteins. This is a proof that the subsets found are sufficient to cluster mice into the different classes. Lastly, we corroborated in the existing literature that the majority of the proteins found in the different subsets played an important role in learning and memory. One example was the case of proteins BRAF, DYRK1A and GFAP that were found to discriminate between classes stimulated and not stimulated to learn in control mice. It is known that mutations in these proteins cause intellectual inability in humans and learning, memory and synaptic plasticity deficits in mice

The use of SOM in this case was advantageous over other clustering methods because it allowed the visual representation of the clustering. Moreover after adding an extra functionality to SOM, it automatically labeled the neurons by colors according to the majority class of mice in each neuron, what facilitated enormously the interpretation of results. Another advantage of SOM over other methods in this particular problem was the preservation of topology of the input data. The result of the clustering was a 2D map of neurons where the close ones clustered data samples also similar in the input space. Thank to this property of SOM we could observed overlapping between classes, in some cases the same neuron grouped two or more mice of different classes. In other cases it was useful because it could be possible to observe which classes of mice were closer to others.. These latter cases indicated that the levels of the selected proteins reflected common features shared by the different classes of mice which, in turn, may reflect novel, common, underlying biological responses. In general, this approach represents a novel analysis of protein expression data capable of providing new valuable information, which was not possible to obtain by standard statistical methods previously used in this type of problem. The results suggest that this approach, applied to additional datasets, can help to identify protein abnormalities in DS mice, and those proteins that need to be altered by drug treatments to facilitate the rescue of learning deficits.

The strategy proposed was used only with part of the data measured in the whole experiment, particularly the nuclear fraction of the cortex from brains of mice. Firstly, it would be of great interest for the research in Down syndrome to perform the same analysis with the rest of the data from the membrane and cytoplasm fraction and also from the hippocampus tissue, a region of the brain more studied in relation with learning and memory. The results obtained in this work on the data used suggests that the application of this strategy can be of great help for the identification of anomalies in expression level of proteins from DS mice, which would be necessary to alter by pharmacological treatment to reduce learning and memory deficits. Secondly, it would be important to apply this strategy with mice data of different genetic origin where learning is normal in context fear conditioning.

This can be useful to identify new classes of critical proteins. In the case of Down syndrome, relevant data would be also protein expression levels of mice after being exposed to context fear conditioning previously treated with other types of drugs. There exist more than 15 drugs with different targets and mechanisms of action thus it is expected that they have different impact on the protein response. SOM could be used to identify critical proteins with similar expression profiles, what would help to define new more potentially effective targets. Lastly, new experiments with other models of DS (here Ts65Dn was used) can help in the extraction of useful information, such as patterns of protein responses, in order to better understand the series of perturbations that can arise in Down syndrome patients.

Chapter 6

Summary in Spanish

6.1. Introducción

6.1.1. Antecedentes

Desde hace décadas, la informática ha tenido una gran repercusión en los avances en biología molecular. A principios de los años 60 los computadores comenzaron a ser recursos disponibles para los investigadores del mundo académico (Hagen, 2000). Tras la aparición del lenguaje de programación FORTRAN, especialmente apropiado para aplicaciones científicas y fácil de aprender, expertos en biología molecular comenzaron a crear sus propios programas que les asistían en sus investigaciones. Entre los científicos pioneros en la aplicación de la informática a sus problemas en biología se encuentran Margaret Oakley Dayhoff (1962; 1965; 1969; 1966) o Walter Fitch 1967; 1966 que sentaron las bases de lo que más adelante se llamaría biología computacional, inspirando y animando a muchos científicos de la época a aplicar la informática como herramienta de gran utilidad en sus investigaciones. En 1970, ya en el pasado siglo, bastantes biólogos computacionales habían desarrollado ya aplicado diversas técnicas informáticas para diferentes estudios en determinados campos de la biología destacando el análisis de la estructura molecular o aspectos relacionados con el estudio de la evolución.

En los primeros años de la década de los 70, los investigadores Paulin Hogeweg y Ben Hesper, expertos en biología teórica comenzaron a utilizar el término bioinformática (Hesper and Hogeweg, 1970; Hogeweg and Hesper, 1978; Hogeweg, 1978) entendido como el estudio de los procesos informáticos y en concreto la acumulación, transmisión y procesamiento de información en sistemas biológicos. Tanto Hogeweg y Hesper como otros autores e investigadores de la época comenzaron a utilizar y desarrollar tanto técnicas de reconocimiento de patrones (Lance and Williams, 1966; Macnaughton-Smith et al., 1964) como métodos de agrupamiento o "clustering" (Hogeweg, 1976) en sus investigaciones. También comenzaron a introducir los autómatas ce-

lulares como un formalismo de modelización en ecología (Hogeweg, 1988) y evolución (Boerlijst and Hogeweg, 1991). Simultáneamente, se desarrollaron métodos de simulación basados en eventos (event-based) u orientados a individuos (más conocidos por su nombre en inglés como individual-oriented), son los que actualmente se denominan métodos de simulación basados en agentes (agent-based). Todas estas técnicas y muchas otras utilizadas en estos primeros años de inicio de la bioinformática pertenecen al paradigma de la Inteligencia Artificial (IA), lo cual nos permite establecer las implicaciones que este campo ha podido tener y tiene en el campo de la biología molecular. Sin embargo, también la IA comenzó a beneficiarse de la biología en los años 60, inspirándose en ella para diseñar métodos novedosos en el campo bioinformático, así como nuevas representaciones de sistemas de procesamiento de información, modelos basados en redes neuronales, reconocimiento de patrones para el aprendizaje (Rosenblatt, 1962), o los algoritmos genéticos y evolutivos como métodos de optimización (Goldberg, 1989; Holland, 1992; Rechenberg, 1973; Schwefel, 1977), que nacieron en un principio por la necesidad de simular la evolución y la selección natural (Crosby et al., 1973; Fraser and Burnell, 1970; Fraser, 1957).

Durante la década de los 80 y 90 siguieron aplicándose y desarrollándose métodos novedosos basados en IA para resolver distintos problemas biológicos. Sin embargo, es a partir del año 2000, tras la publicación del primer borrador completo del genoma humano, hito científico sin precedentes, cuando estos métodos adquieren especial importancia. A partir de ese momento las técnicas de secuenciación genómica comenzaron a evolucionar de manera asombrosa. A estas técnicas se le suman otras más de proteómica, transcriptómica, fluxómica o metabolómica, denominadas ómicas, que permiten tener una visión completa de una determinada célula u organismo. El problema principal radica en el manejo, análisis e integración de la gran cantidad de información que generan estas técnicas. En este momento la biología molecular ha ingresado de pleno derecho en el grupo de los Big Data y precisamente la bioinformática se sitúa como la llave para poder conducir con éxito una investigación en biología molecular. Para procesar e interpretar la información compleja y variada que se deriva de los laboratorios es necesario el desarrollo de nuevas técnicas computacionales que sean capaces de procesar y analizar toda esta información. En los últimos años se han desarrollado y aplicado un gran número de métodos computacionales basados en IA, y más concretamente en aprendizaje automático con este fin; donde el objetivo ha sido construir herramientas, que por un lado ayuden a interpretar y extraer información valiosa de dichos datos y por otro, sean capaces de simular procesos biológicos y construir modelos que ayuden a entender dichos procesos. Larrañaga y colaboradores (Larranaga et al., 2006) realizan una revisión exhaustiva de métodos de aprendizaje automático utilizados en bioinformática y dividen los principales problemas biológicos en

dos categorías: problemas de modelado y problemas de optimización. En los primeros el aprendizaje consiste en ejecutar un programa que induzca a la construcción de un modelo basado en unos datos de entrenamiento para poder posteriormente inferir información a partir de él. A este tipo pertenecen principalmente problemas de clasificación. Estos problemas se resuelven generalmente mediante métodos de clasificación supervisada (cuando para una parte del conjunto de datos, llamado conjunto de entrenamiento, se conoce la clase a la que pertenecen, en este caso se dice que los datos están etiquetados) o clasificación no supervisada (si no se dispone de dicho conjunto de entrenamiento). En el primer caso, se construye un modelo generalizador que permite posteriormente clasificar nuevos datos similares y en el segundo su principal objetivo es identificar automáticamente patrones interesantes en los datos no perceptibles a simple vista. Ejemplos de métodos característicos de los dos tipos de clasificación pueden encontrarse en (Pajares and de la Cruz, 2010), si bien desde la perspectiva de la visión por computador. Un esquema representativo relativo a clasificación se presenta en la Figura 1.3.

Ejemplos de la utilización de métodos de clasificación supervisada tales como redes neuronales, máquinas vector soporte (support vector machines en inglés) o la estrategia del vecino más próximo en bioinformática pueden encontrarse en (Bao and Cui, 2005; Carter et al., 2001; Cypess et al., 2013; Jagga and Gupta, 2014; Kim, 2004; López-Bigas and Ouzounis, 2004; Salamov and Solovyev, 1995; Sø rlie et al., 2001; Yi and Lander, 1993). Por otro lado, ejemplos de métodos de clasificación no supervisada o clustering, también en el entorno de la bioinformática, son (Bohlin et al., 2009; Brohée and van Helden, 2006; Herrero et al., 2001; Lorenzo-Redondo et al., 2014; Sheng et al., 2003; Spellman et al., 1998; Tamayo et al., 1999). Estos últimos se han utilizado en numerosas ocasiones, principalmente para el análisis de datos de expresión de genes, con el fin de encontrar la relación entre genes con un perfil de expresión similar y algún tipo de similitud funcional o de regulación.

En el segundo tipo de problemas, cuya estrategia de solución se basa en optimización, se encuentran aquellos en los que se requiere encontrar una solución óptima dentro de un espacio de múltiples soluciones posibles. Existen varios tipos de métodos de optimización, si bien en biología los más utilizados y que mejores resultados ofrecen son los métodos de optimización global. Esto es así porque muchos de los problemas de optimización en biología son no lineales, con muchos óptimos locales y habitualmente NP-completos. Los métodos de optimización global son capaces de explorar un espacio de soluciones más amplio que otros métodos, escapando de óptimos locales. Éstos se dividen a su vez en deterministas y estocásticos. Los métodos estocásticos se basan en estrategias probabilísticas y tienen por tanto cierta componente aleatoria. Por esta razón no pueden ofrecer la garantía absoluta de haber encontrado el óptimo global. Aunque los métodos deter-

ministas pueden garantizar la óptimalidad global para algunos problemas, ninguno de ellos puede resolver de manera general problemas de optimización global en tiempo finito. El coste computacional derivado de aplicar estos métodos aumenta muy rápidamente con el tamaño de datos asociados al problema. Los métodos estocásticos sí son capaces de localizar la vecindad de soluciones globales con relativa eficiencia pero con el coste de no poder garantizar la optimalidad global de la solución (Moles et al., 2003). Ejemplos de métodos estocásticos de optimización global son por ejemplo los algoritmos genéticos o la programación evolutiva, comentados previamente, el método de enfriamiento simulado (simulated annealing en inglés) (Kirkpatrick et al., 1983; van Laarhoven and Aarts, 1987) o los métodos de optimización basados en las colonias de hormigas (Colorni et al., 1991). Diversos trabajos sobre su aplicación en biología pueden encontrarse en (Falkenauer and Marchand, 2002; Fogel et al., 2002; Keith et al., 2002; Kikuchi et al., 2003; Krasnogor et al., 2002; Nguyen et al., 2002; Park et al., 1997; Rodriguez-Fernandez et al., 2006; Smith, 2004; Wren et al., 2004). En (Moles et al., 2003) se hace también una interesante revisión y comparación de métodos de optimización global aplicados al caso particular de la estimación de parámetros en rutas metabólicas.

Por último, otro tipo de optimización que se está empezando a aplicar a la resolución de problemas biológicos y que en otros ámbitos ya se utiliza desde hace más tiempo es la optimización multi-objetivo, que trata de optimizar simultáneamente dos o más funciones objetivo. Este tipo de optimización, en lugar de obtener una única solución, encuentra una familia de soluciones llamada frente Pareto de soluciones óptimas en el que no es posible mejorar ninguna de ellas en un objetivo concreto sin empeorar en otro. Cada una de ellas es una solución consenso entre los distintos objetivos y es necesario disponer de un proceso de decisión por parte del experto para decantarse por una u otra solución como la óptima. El método más ampliamente utilizado es el Non-dominated Sorting Genetic Algorithm-II (NSGA-II) (Deb et al., 2002) basado en un algoritmo evolutivo, sin embargo otros autores han implementado nuevas estrategias con resultados exitosos (Sendín et al., 2009) principalmente orientadas a la estimación de parámetros en problemas de biología de sistemas. (Handl and Knowles, 2007) hacen una revisión muy completa de diferentes problemas biológicos en los que la utilización de optimización multi-objetivo ha sido relevante.

De lo expuesto anteriormente se deduce la gran importancia que tienen los métodos de aprendizaje automático, considerando clasificación y optimización, en los avances en biología molecular. Sin embargo, la biología molecular es un campo amplísimo y dentro de ella existen una gran cantidad de áreas de investigación como son la genómica, la proteómica, la metabolómica o la biología de sistemas, entre otras. De todas ellas se obtiene información para la determinación de lo que se llama metabolismo. En esencia, el metabolismo

es la expresión fenotípica del genotipo (genómica, proteómica, metabolómica) y al mismo tiempo, su concreción en organismos concretos resulta ser la consecuencia de un proceso evolutivo sometido a optimización. Justamente el metabolismo constituye el tema central de investigación presentado en esta tesis mediante la aplicación, diseño y desarrollo de métodos computacionales basados en aprendizaje automático aplicados a tres problemas biológicos distintos relacionados con el metabolismo.

6.1.2. Identificación del problema de investigación

El metabolismo consiste en una intrincada red de reacciones químicas que ocurren dentro de las células mediante el cual los organismos se mantienen vivos. En términos generales, el metabolismo se encarga de llevar a cabo transformaciones a partir de un producto inicial, como pueden ser los alimentos, en un producto final, tal como la energía, construcción de nuevas moléculas para formar estructuras (lípidos de membranas, proteínas, material genético como DNA, etc.), así como de degradar biomoléculas (catabolismo). A cada una de estas transformaciones particulares se le suele denominar "ruta metabólica". Pero las rutas metabólicas están a su vez conectadas entre sí por reacciones que transforman un metabolito [1] de una de ellas en otro metabolito de una segunda ruta metabólica. De este modo, el metabolismo se puede concebir, más que como un conjunto de rutas metabólicas, como una red de reacciones a la que, de aquí en adelante, denominaremos "red metabólica". Alternativamente, y de una forma más rigurosa, una red metabólica se puede representar como un grafo bipartito donde los nodos alternos de la misma son los metabolitos y los intermedios, que aparecen entre dos externos, las reacciones que transforman un metabolito en otro. Estas reacciones son mediadas o catalizadas por unas moléculas llamadas enzimas, que se activan o se inhiben en función de las necesidades de la célula. La Figura 1.4 mostrada en el Capítulo 1 representa una red metabólica donde los círculos A, B, C, D se corresponden con los metabolitos (sustratos y productos de las reacciones) y los cuadrados con las reacciones catalizadas por las correspondientes enzimas. El resultado de la expresión del genoma consiste, entre otras cosas, en una serie de proteínas (indicadas en la figura 1.4 dentro de rectángulos) que poseen una actividad catalítica (enzimas), las proteínas en la red tienen una numeración estándar de acuerdo con la reacción que son capaces de catalizar. Estas actividades relacionan determinados metabolitos: por ejemplo, la enzima 2.3.2.1 es capaz de catalizar la conversión del metabolito A en el B. Como resultado aparece una red metabólica que se puede representar como un grafo bipartito en el cual los nodos alternativos son los metabolitos (A, B, C y D) y los intermedios las reacciones (2.3.2.1 y 2.4.1). Sobre esta red metabólica, representada en el interior del cuadrado azul, se sobrepone una

[1] Un metabolito es cualquier molécula que intervenga bien como sustrato o como producto en una reacción metabólica.

red de regulación. Así, un metabolito (por ejemplo, el D) puede ejercer una regulación sobre la reacción 2.3.2.1. Dicha regulación puede tener lugar a dos niveles: a) a nivel de la actividad enzimática, actuando directamente sobre una enzima (línea marrón) haciendo que aumente o disminuya la producción de un determinado metabolito o b) una regulación a nivel de la expresión inhibiendo (-) o activando (+) la formación de la proteína que posee dicha actividad enzimática (línea verde).

A la hora de estudiar el metabolismo y cómo éste influye en el comportamiento de una célula u organismo, y atendiendo al esquema anteriormente comentado, existen varias vertientes dignas de consideración:

1. Análisis del comportamiento dinámico de las redes metabólicas.
2. Estudio de la estructura de la red y su relación con características fenotípicas o con preferencias ambientales.
3. Descubrimiento de patrones de expresión específicos vinculados a determinadas actividades metabólicas, funcionales o comportamientos

El trabajo de investigación que se recoge en la presente tesis pretende dar respuesta a tres tipos de problemas diferentes dentro de las tres vertientes indicadas anteriormente mediante la aplicación y desarrollo de métodos computacionales novedosos basados en aprendizaje automático. El primer trabajo se corresponde con un análisis dinámico del metabolismo, en el que se estudia la regulación a nivel enzimático de un ciclo metabólico mediante la estimación de unos parámetros cinéticos, cuyo valor determina el funcionamiento dinámico de las redes metabólicas. El estudio se propone desde una perspectiva de la optimización multi-objetivo empleándose de manera novedosa en este tipo de estudios.

El segundo trabajo se corresponde con un estudio estructural del metabolismo en el que el objetivo es agrupar un conjunto de especies bacterianas por similitud en las características estructurales de sus metabolismos. Para ello, se ha diseñado un Sistema Experto basado en la combinación de técnicas de clasificación no supervisada e índices de validación adaptado a la naturaleza de los datos. Un segundo objetivo del método consiste en identificar características fenotípicas o de comportamiento comunes dentro de los grupos resultantes que ayuden a extraer información de los datos, inapreciable a simple vista, como puede es la relación entre metabolismo y ambiente.

El tercer trabajo se corresponde con un estudio funcional del metabolismo en el que se desarrolla una nueva estrategia de minería de datos basada también en técnicas de clasificación no supervisada. Esta vez, sin embargo, estas técnicas se combinan con un test estadístico para identificar proteínas involucradas en actividades funcionales de alto nivel, tales como el aprendizaje y la memoria en ratones control y síndrome de Down (SD). La estrategia se presenta como una manera novedosa de tratar datos de expresión

de proteínas, que son resultado del segundo tipo de regulación de las redes metabólicas, concretamente la regulación a nivel de expresión. La estrategia propuesta es capaz de descubrir patrones de comportamiento a partir de datos experimentales imposibles de identificar a través de análisis estadísticos estándar, utilizados hasta ahora en este tipo de estudios.

En definitiva, el trabajo de investigación resuelve problemas en el ámbito biológico mediante técnicas computacionales novedosas. Se trata de un proyecto netamente multidisciplinar con implicaciones en los campos de la biología molecular y la bioquímica mediante técnicas computacionales en el ámbito de la IA.

6.2. Objetivos

El objetivo general que se plantea es el estudio, aplicación y desarrollo de métodos computacionales basados en aprendizaje automático aplicados al campo de la biología para el estudio del metabolismo y su efecto en el comportamiento. Los objetivos específicos se concretan en los tres siguientes:

(a) Estudio y aplicación de métodos de optimización para la identificación de un patrón de regulación óptimo de una red metabólica.

(b) Desarrollo de un sistema experto para la clasificación automática no supervisada de especies bacterianas según características metabólicas.

(c) Desarrollo de una estrategia de minería de datos basada en clasificación no supervisada para el análisis de datos experimentales de expresión de proteínas en ratones control y síndrome de Down.

6.3. Principales resultados

Los objetivos anteriores responden a tres problemas biológicos diferentes y para cada uno de ellos se han propuesto soluciones computacionales distintas. Por ello, los principales resultados y conclusiones relativos a los tres problemas planteados se concretan por separado en esta misma sección, mientras que las conclusiones generales resultantes de la investigación como conjunto se describen en la Sección 4 de este resumen.

6.3.1. Aproximación a la regulación de redes metabólicas mediante optimización multi-objetivo

El trabajo descrito en el Capítulo 2 de la presente tesis tiene como objetivo el estudio dinámico de redes metabólicas mediante métodos de optimización. En concreto, dicho estudio se afronta a través de la estimación de un

conjunto de parámetros, llamados cinéticos, encargados de regular el funcionamiento dinámico de un ciclo metabólico. El modelo utilizado representado en la Figura 2.1 del Capítulo 2 fue estudiado en los años 90 por Gilman y Ross (1995) y consiste en un modelo de un ciclo substrato, que idealiza una célula animal en la que se metaboliza glucosa en sangre en energía para dicha célula mientras la concentración de glucosa en sangre es adecuada, pero que por otra parte sintetiza glucosa para exportar cuando la concentración de glucosa en sangre disminuye en exceso. El sistema se describe mediante las ecuaciones 2.1,2.2, 2.3 (Capítulo 1)donde los valores de ocho parámetros: $K_{\alpha,F}$, $K_{\alpha,T}$, $K_{\beta,F}$, $K_{\beta,T}$, $r_{\alpha,F}$, $r_{\alpha,T}$, $r_{\beta,F}$, $r_{\beta,T}$ determinan la dirección del flujo en un sentido u otro. El valor de estos parámetros se puede plasmar en un esquema de regulación específico, un ejemplo de este esquema puede observarse en la Figura 2.1B. El objetivo de Gilman y Ross consistía en encontrar un único conjunto de valores para los ocho parámetros que permitiera al sistema comportarse de manera óptima en condiciones externas cambiantes. Estas condiciones vienen representadas por las concentraciones de F (glucosa en el exterior de la célula) y T (glucosa en el interior).

Se sabe que el metabolismo y las rutas metabólicas han sufrido un proceso natural de optimización a lo largo del tiempo para convertirse en lo que actualmente son. Gilman y Ross utilizaron un enfoque mono-objetivo para resolver el problema, simulando en cierto modo el proceso de evolución natural. Diseñaron un algoritmo genético para estimar el valor de esos parámetros en distintos perfiles de concentración variantes de F y T (véase ejemplo en 2.3) con la esperanza de encontrar algún esquema de regulación universal para este sistema que se comportara de manera óptima en todos los perfiles. Sin embargo, sólo encontraron soluciones que llamaron "especialistas", soluciones que funcionaban bien en un determinado perfil pero mal en otros.

En esta tesis se afronta el mismo problema de encontrar un conjunto de valores para estos parámetros, y por tanto un esquema de regulación universal para este sistema, sin embargo esta vez desde un enfoque de optimización multi-objetivo, cuyos resultados se concretaron en (Higuera et al., 2012). Gilman y Ross combinaron dos funciones objetivo (ecuaciones: 2.5,2.6) en una única función, en nuestro caso optimizamos las dos funciones simultáneamente. Para ello aplicamos el método NBIWT *weighted Tchebycheff* descrito en (Sendín et al., 2010) con seis perfiles distintos de concentraciones (Figura 2.3). Como resultado se obtuvo un conjunto de soluciones para cada perfil de concentraciones, en vez de una única solución, que representaban un consenso entre los dos objetivos. Este conjunto de soluciones se llama frente Pareto de soluciones óptimas. Todas ellas son matemáticamente óptimas y en general en el caso de desear una solución única es necesario un proceso de decisión en el que basándose en información adicional se sacrifique un objetivo con respecto a otro. Sin embargo, en este problema concreto se observó

que en el frente Pareto, resultado de la optimización en cada perfil, aparecían soluciones 'rodilla" en cada uno de ellos. Estas soluciones se consideran soluciones preferidas dentro del frente Pareto porque una leve mejora en uno de los objetivos supone un empeoramiento en el segundo objetivo. Los frentes Pareto resultantes en cada perfil se pueden observar en la Figura 2.5 del Capítulo 2, denotándose en rojo la mencionada solución rodilla.

Considerando las soluciones rodilla de cada perfil como óptimas se obtuvieron los esquemas de regulación correspondientes. Estos esquemas resultaron ser ligeramente diferentes en cada rodilla, sin embargo, se observó que intercambiando los parámetros de las rodillas de un perfil en el resto de perfiles se obtenía el valor óptimo obtenido en cada perfil. Este resultado indica que las seis rodillas se comportaban de manera óptima en los seis perfiles, lo que induce a pensar en la existencia de un esquema universal de regulación para este sistema. Se ejecutó el método multi-objetivo diez veces en cada uno de los perfiles y se encontró un esquema de regulación que aparecía con más frecuencia como rodilla en todos los perfiles. A continuación se obtuvo un conjunto de parámetros consenso calculando la media de los valores de los parámetros de cada rodilla que respondía a ese esquema de regulación. El conjunto de parámetros calculado también resultó óptimo en cada uno de los perfiles, como se puede observar en la gráfica de la Figura 2.10. Este resultado puede verse como una indicación de la existencia de mecanismos universales de regulación en ciclos substrato, los cuales son muy frecuentes en el metabolismo. Se pueden encontrar diversos ejemplos en la literatura (Berg et al., 2006; Morán and Goldbeter, 1984). En este sentido tiene especial importancia el ciclo PFK2-FBPase2 (Berg et al., 2006), que presenta exactamente el mismo patrón de regulación que el encontrado por medio del procedimiento de optimización multi-objetivo. Es necesario remarcar que los resultados obtenidos mediante optimización multi-objetivo no se pueden derivar mediante un enfoque mono-objetivo, lo que se comprobó en una primera parte del trabajo reproduciendo los resultados de Gilman y Ross con tres métodos distintos de optimización mono-objetivo.

Gracias a la aplicación de la optimización multi-objetivo se consiguió encontrar un patrón universal de regulación para el ciclo metabólico estudiado. El estudio de este tipo de ciclos es de especial importancia porque ocurren con mucha frecuencia en el metabolismo y en muchos casos están implicados en puntos muy sensibles del control metabólico. Este trabajo constituye una aportación en el complejo campo de la regulación metabólica, pudiéndose en un futuro aplicar a redes metabólicas más grandes y complejas. Sin embargo, las implicaciones de este trabajo van más allá del análisis de la regulación metabólica. Por ejemplo, podría aplicarse una aproximación multi-objetivo en el campo de la biología sintética que se encarga principalmente del diseño de circuitos biológicos, lo cual podría ayudar a aumentar su robustez y viabilidad.

6.3.2. Diseño de un sistema experto para el agrupamiento de especies procariotas según sus características metabólicas

En el Capítulo 3 de esta tesis se describe el diseño de un sistema experto (SE) cuya base es el trabajo de (Higuera et al., 2013), para agrupar un conjunto de 365 especies procariotas de acuerdo a su similitud según un conjunto de características metabólicas. Estas características se basan en elementos estructurales del metabolismo de cada especie, concretándose en la ausencia o presencia de determinadas rutas metabólicas en las especies en cuestión y de forma más precisa en cómo de completa está una determinada ruta metabólica en cada especie (porcentaje de enzimas anotadas en la base de datos KEGG (Kanehisa, 2002)). Un ejemplo del conjunto de datos puede observarse en la Tabla 3.1.

El SE se inspira en el razonamiento humano a la hora de asignar categorías a determinados elementos, primero agrupando los elementos más sencillos para más adelante agrupar los que resultan más difíciles. El SE se basa en la combinación de la técnica de clasificación no supervisada: Mapas auto-organizativos (más conocido por sus siglas en inglés: Self organizing Maps, SOM) así como en determinados índices de validación siguiendo una estrategia jerárquica. El sistema agrupa las especies en fases como se describe en el esquema de la Figura 3.2. En primer lugar determina el número óptimo de clases mediante la utilización del índice de validación de Davies Bouldin (DB) (Davies and Bouldin, 1979) y a continuación comienza un proceso iterativo en el que se identifican clases relevantes mediante un índice inspirado en DB.

El resultado de aplicar el SE a los datos en cuestión permite, en primer lugar, determinar el número óptimo de clases que mejor se adapta a la topología de los datos. El agrupamiento se llevó a cabo con once SOMs de diferentes tamaños desde 7x7 hasta 17x17 y se calculó el índice DB para cada partición resultante. Este proceso se realizó diez veces para cada tamaño. El índice DB es una medida de calidad de una determinada partición de los datos. Se basa en la distancia intra- e inter- clase, de esta forma cuanto menor es el valor de DB más separados estarán las clases y más cercanos serán los elementos dentro de cada clase. Por lo tanto, a medida que disminuye DB aumentará la calidad del agrupamiento o partición. En este caso se obtuvo un valor de DB mínimo para una SOM de dimensiones 10x10. Estos resultados se pueden observar tanto en la tabla 3.2 como en la Figura 3.5 del Capítulo 3. A continuación se inició el proceso iterativo repitiendo las fases L2 - L4 del esquema mostrado en la Figura 3.2. Durante la fase L2 se identifican clases relevantes, para ello se utiliza un índice de validación inspirado en DB que llamaremos DB', el cual permite ordenar las clases según su distancia con respecto a otros clusters y la compacidad de los elementos dentro de ellos. Las especies agrupadas en estas clases se separan del con-

junto de datos. Se estimó como adecuada una tasa de eliminación inicial del 30 %, esto es, para una SOM de dimensión 10x10 se separarían del conjunto de datos las especies agrupadas en los 30 clases con mejor valor de DB'. Para las siguientes iteraciones se estableció un decrecimiento del 33.3 %. Así para tres iteraciones del SE las tasas se corresponderían con 30 %, 20 % y 10 %. También, como se explica en la Sección 3.5.3, la dimensión de la SOM disminuye en cada iteración con la intención de compactar los datos tras haber eliminado varias de las muestras del conjunto precedente.

Para evitar malinterpretar los resultados se ejecutó el SE 100 veces (fases L2-L4) y se observó que comenzando con una SOM 10x10 en el 31 % de los casos DB decrecía en las siguientes fases: SOM 9x9, SOM 8x8 y continuaba decreciendo al menos hasta SOM 7x7. En el 45 % de los casos DB disminuye hasta la SOM 8x8 y después aumenta. El 24 % restante presenta otro tipo de comportamiento. Estos resultados indican que en el 76 % de los casos el índice DB disminuye al menos hasta 8x8. Los dos tipos de comportamiento principales pueden observarse en la Figura 3.6. Por lo tanto se muestra que una excesiva reducción de SOM no mejora la partición. La causa estriba en el hecho de que el reducido conjunto de datos usado en la cuarta iteración contiene elementos residuales o elementos muy difíciles de clasificar. Esto explicaría la variabilidad del comportamiento del SE en la cuarta iteración.

Tras este análisis se llegó a la conclusión de que para garantizar un buen funcionamiento del SE, para este tipo de datos disponibles, el sistema debe detener su ejecución cuando DB deja de decrecer, lo que ocurre de manera general cuando se alcanza una SOM de dimensión 8x8. El resultado final del agrupamiento está formado por las clases eliminadas en las primeras iteraciones de ejecución del SE y la última partición de SOM.

Mediante este sistema se garantiza que los datos se agrupan con suficiente nivel de confianza, gracias al uso de índices cuantitativos de validación. Por último, se analizó el contenido de las clases obtenidas iterativamente en uno de los casos en que DB decrece hasta SOM 8x8. En primer lugar se observó que las clases eran biológicamente coherentes. En segundo lugar, con el objetivo de ahondar en la relevancia biológica de las clases obtenidas, se identificaron características fenotípicas comunes dentro de los elementos de la misma clase. Muchas de las especies agrupadas en la misma clase no solo compartían similitudes metabólicas sino también preferencias ambientales o de comportamiento, como puede ser la patogenicidad, la habilidad de crecer en ambientes hostiles o la formación de esporas para protegerse de amenazas externas. El hecho de encontrar estas características comunes resulta de gran ayuda a la hora de explorar hasta qué punto el metabolismo puede estar relacionado con esas características fenotípicas y más concretamente si determinadas rutas metabólicas pueden estar involucradas en el desarrollo de esas características y comportamientos.

Las figuras 3.8, 3.9 y 3.10 muestran el contenido de las clases obtenidas

en distintas fases del SE y las características fenotípicas comunes. Además, se identificaron clases que agrupaban especies metabólicamente similares y de diferente orden taxonómico, lo que reflejaba su lejanía filogenética. Además, estas especies compartían preferencias ambientales, lo cual indica que existen muchas probabilidades de que su metabolismo estuviera involucrado en la adaptación a dicho ambiente. El estudio de las rutas metabólicas concretas más relevantes, a la hora de agrupar estas especies taxonómicamente diferentes en la misma clase, puede conducir a la identificación de rutas específicas de determinados ambientes. Estos resultados tienen especial relevancia a la hora de entender mejor cómo funcionan las comunidades bacterianas, responsables de números procesos naturales y artificiales.

En conclusión, el SE diseñado nos permite agrupar de manera no supervisada un conjunto complejo de datos biológicos del que no se conoce el número de clases. Además, gracias a la utilización de índices de validación, que permiten monitorizar el proceso, se obtiene cierto nivel de garantía de que el agrupamiento se ha llevado a cabo correctamente, algo difícil de conseguir especialmente cuando se trata de datos biológicos. El análisis de los resultados muestra su coherencia biológica y permite extraer información adicional sobre los datos, al igual que la extracción de hipótesis como puede ser la relación entre metabolismo y ambiente.

6.3.3. Nueva estrategia de minería de datos basada en clasificación no supervisada para el análisis de datos de expresión de proteínas

En el Capítulo 4 de esta memoria de tesis se describe una nueva estrategia novedosa propia de minería de datos basada en clasificación no supervisada para analizar de forma eficiente un conjunto de datos experimentales. En primer lugar, la estrategia se basa en agrupar automáticamente y sin supervisión ocho clases diferentes de ratones control y síndrome de Down (SD), con diferentes respuestas de aprendizaje y tratamiento, según su similitud en el valor de expresión de 77 proteínas. En segundo lugar, la estrategia se sirve de un índice estadístico para identificar subconjuntos de proteínas discriminantes entre las distintas clases y con un papel relevante en aprendizaje y memoria. En su conjunto, la estrategia propuesta ayuda también a encontrar patrones biológicos de comportamiento en las distintas clases de ratones y diferentes tipos de aprendizaje. Patrones que no se pueden identificar con métodos estadísticos estándar utilizados hasta la fecha en este tipo de estudios.

Tanto las distintas clases de ratones como la base del experimento y el conjunto de datos pueden observarse en la Figura 4.1 y la Tabla 4.1. El método de clasificación utilizado fue el mismo que en el trabajo anterior, es decir SOM, con la diferencia de que en este caso se explotaron otras funcionalidades de SOM no usadas anteriormente, como son el agrupamiento de

datos multidimensionales en un mapa visual de dos dimensiones, la conservación de la topología de los datos de entrada y la posibilidad de etiquetar el mapa con información adicional no utilizada en el proceso de agrupamiento.

El esquema de funcionamiento de la estrategia propuesta se muestra en la Figura 4.2. Las consideraciones adicionales con respecto al tamaño de la SOM y los detalles específicos sobre el diseño de la estrategia se describen en la Sección 4.5. La estrategia se aplicó en primer lugar a los datos de las cuatro clases de ratones control, dos de ellas estimuladas al aprendizaje y las otras dos sin estimulación. Dentro de estas dos clases un cierto número de ratones había sido tratado con mementina, fármaco que restaura la capacidad de aprendizaje en enfermos de Alzheimer y otros no. Se llevó a cabo el agrupamiento con SOM considerando los valores de expresión de las 77 proteínas como características.

Uno de los resultados más relevantes obtenidos en esta fase fue que en el mapa de neuronas resultante (4.3) se observó que las dos clases de ratones no estimulados al aprendizaje se encontraban en una región del mapa mientras que las otras dos, sí estimuladas, aparecían en una región opuesta, separadas por una clara frontera. Dado que SOM conserva la topología de los datos, se deduce fácilmente que elementos similares en el espacio de entrada se agrupan en zonas próximas entre sí dentro del mapa, este resultado fue una indicación de que las proteínas usadas para el agrupamiento efectivamente servían para discriminar al menos aprendizaje de no aprendizaje. A continuación se identificaron grupos de neuronas adyacentes que agrupaban ratones de la misma clase, considerando estas neuronas representantes de clase. Esto se llevó a cabo de acuerdo con los criterios establecidos en la Sección 4.6.1.2. A continuación se compararon las distintas clases de ratones aplicando el test de Wilcoxon entre los vectores de pesos de las neuronas representantes de cada clase. Con esto se consiguió determinar qué proteínas tenían valores significativamente diferentes en las distintas clases y por lo tanto discriminantes entre distintos tipos de aprendizaje. Los distintos subconjuntos de proteínas discriminantes aparecen en la Tabla 4.4. Posteriormente se comprobó, según se acredita en diversos trabajos existentes en la bibliografía, que la mayoría de las proteínas de estos subconjuntos jugaban un papel importante en el aprendizaje o memoria. Un ejemplo de proteínas discriminantes encontradas entre aprendizaje y no aprendizaje fueron BRAF, ERK y pERK, conocidas por su papel en la ruta de señalización MAPK crítica en el aprendizaje.

Por último, se utilizó SOM como herramienta de validación de los resultados. Para ello, se repitió el agrupamiento de las 4 clases de ratones utilizando únicamente los subconjuntos de proteínas discriminantes. Se observó por ejemplo que utilizando únicamente 11 de las 77 proteínas (Figura 4.6A) se conseguía una separación similar entre aprendizaje y no aprendizaje a la que se obtenía con 77 (Figura 4.4), lo cual indica que únicamente con esas 11 proteínas era posible separar las dos clases principales, revelando la

importancia biológica de ese subconjunto. También se comprobó la validez de los subconjuntos de proteínas que diferenciaban entre no estimulación con y sin mementina y estimulación con y sin mementina (Figura 4.7 y 4.8). Una vez comprobada la eficacia de la estrategia con datos de ratones control se aplicó el mismo procedimiento a los datos de las cuatro clases de ratones con síndrome de Down.

En este caso, la interpretación de los resultados resultaba un poco más complicada, debido a que de las dos clases estimuladas al aprendizaje únicamente conseguía aprender satisfactoriamente aquella que había sido tratada con mementina. En el mapa de SOM resultante del agrupamiento (Figura 4.9) se puede observar la separación entre las dos clases estimuladas al aprendizaje y las dos no estimuladas. El hecho de poder visualizar el agrupamiento y la ventaja de la conservación de la topología de SOM permite encontrar patrones inherentes a la estructura de los datos. Por ejemplo, los ratones estimulados al aprendizaje sin mementina aparecen en el mapa más próximos entre sí que los no estimulados al aprendizaje, cosa que no pasaba en control. Esto puede indicar que los ratones que fracasan en el aprendizaje se asemejan más, en cuanto a su nivel de expresión se refiere, a los no estimulados al aprendizaje que aquellos estimulados al aprendizaje tratados con mementina y que por lo tanto aprenden con éxito.

Se obtuvieron a continuación las proteínas discriminantes entre las distintas clases que resultaban interesantes desde el punto de vista biológico, tal y como se puede observar en la Tabla 4.7. Por último, se procedió a la validación de los subconjuntos de proteínas encontrados repitiendo el agrupamiento de los ratones con SD únicamente con 15 proteínas, que discriminaban entre las clases de no estimulación con y sin mementina. El resultado es una separación completa en dos regiones del mapa de las dos clases de ratones (Figura 4.10A). A la hora de validar el subconjunto de 12 proteínas discriminantes entre las clases estimuladas al aprendizaje con y sin mementina se observó, en el mapa resultante (Figura 4.10B), que existía un grupo de neuronas adyacentes que agrupaban únicamente ratones que aprendían con éxito (tratados con mementina); por tanto, estas proteínas eran capaces de separar la gran mayoría de los ratones con SD que aprendían con éxito de los estimulados a aprender que fracasaban. Sin embargo, en algunas de las neuronas restantes se mezclaban las dos clases. Esto puede ser un indicativo de que o bien no se pudieron identificar todas las proteínas relevantes para la discriminación o que estas dos clases compartían similitudes difíciles de diferenciar.

Por último, se aplicó la estrategia a un conjunto de datos de ratones control y síndrome de Down juntos para intentar dilucidar las diferencias relativas al nivel de proteínas de ratones sanos con respecto a ratones enfermos. Primero se obtuvo un subconjunto de 10 de las 77 que discriminaban aquellos ratones SD que presentaban aprendizaje fallido (estimulados al aprendizaje

sin mementina) de ratones control que aprendían con normalidad tanto con mementina como sin ella. Repitiendo el agrupamiento únicamente con las 10 proteínas se observa que los ratones con SD se separan completamente de los control (Figura 4.12B). Esto indica que el nivel de expresión de esas 10 proteínas son determinantes del fracaso en el aprendizaje en DS. Se hizo una segunda prueba agrupando con SOM los mismos ratones control y los SD tratados con mementina utilizando las 10 proteínas anteriores, observándose que los ratones con SD ya no se separan de los de control tan claramente como los de aprendizaje fallido. Esto muestra que ambos tipos de ratones comparten similitudes en esas 10 proteínas que no permite diferenciarlos en el mapa auto-organizativo. La explicación lógica a esta situación es que la mementina ayuda a alcanzar unos niveles de expresión que se acercan más a los de control y que son más adecuados para el aprendizaje. Se llevaron a cabo un conjunto de pruebas similares con ratones control y síndrome de Down que se representan en las Figuras 4.14A y B.

La contribución principal de este trabajo ha sido el diseño de una estrategia basada en clasificación no supervisada para extraer gran cantidad de información de un complejo conjunto de datos experimentales. Por un lado, la identificación de subconjuntos de proteínas que describen los distintos tipos de aprendizaje en ratones control y SD, así como la influencia de la mementina en la recuperación del aprendizaje. Por otro lado, la visualización y las propiedades de SOM que le diferencian de otros métodos de clasificación no supervisada, especialmente la conservación de la topología, permiten conocer la estructura de los datos y ayudar a los biólogos a identificar patrones informativos biológicos e interpretar las similitudes entre clases de ratones por su proximidad topológica en el mapa. Algo que no permiten otros métodos de esta naturaleza. Los resultados obtenidos sugieren que esta estrategia, aplicada a nuevos conjuntos de datos, puede ayudar en la identificación por un lado de anomalías a nivel de expresión de proteínas y por otro de aquellas proteínas que necesitan ser alteradas mediante tratamientos farmacológicos para facilitar la recuperación de déficits en aprendizaje y memoria en pacientes con SD.

6.4. Conclusiones generales

En la sección anterior, tras la descripción de los resultados se comentaron algunas de las conclusiones principales de cada trabajo, en esta sección comentamos las conclusiones generales de la tesis como conjunto.

Tal y como se planteó en la introducción, en la presente tesis se han afrontado y proporcionado soluciones a tres problemas diferentes relacionados con el estudio del metabolismo en el campo de la bioquímica mediante diferentes técnicas computacionales. Concretamente técnicas basadas en aprendizaje automático incluyendo métodos de optimización y clasificación. Los resulta-

dos de la aplicación de dicha metodología a estos problemas junto con las conclusiones específicas más relevantes, descritos brevemente en la sección anterior, suponen por un lado, un avance en las áreas biológicas de investigación concretas de cada problema y por otro, avances en investigación en informática a través de la aplicación y desarrollo de nuevos métodos basados en aprendizaje automático. Las tres estrategias propuestas, que conforman en su conjunto la tesis, se han diseñado y aplicado para resolver problemas biológicos, sin embargo las tres podrían ser utilizadas para otros problemas y conjuntos de datos de diferentes campos. Pueden servir como inspiración para resolver nuevos problemas actuales tanto en el campo de la minería de datos como en problemas de optimización. En primer lugar, se mostró que la aplicación de optimización multi-objetivo, aplicada con menos frecuencia en las ciencias de la vida que en otros campos como la física o la ingeniería, puede ofrecer nuevas y mejores soluciones a problemas que llevan tiempo afrontándose desde perspectivas mono-objetivo. También, se han propuesto dos estrategias novedosas de minería de datos basadas en el método SOM de clasificación no supervisada con el objetivo de descubrir información novedosa subyacente a dos conjuntos de datos de distinto tipo. Desde su primera publicación en 1982 por el científico Teuvo Kohonen 1982 SOM se ha aplicado en numerosas ocasiones, sin embargo según van surgiendo nuevos problemas, nuevas versiones del método o incluso la combinación del mismo con otras estrategias han mostrado resultados exitosos. En la investigación desarrollada se han mostrado dos ejemplos, en uno de ellos se combina SOM con índices de validación de agrupamiento para el diseño de un SE capaz de agrupar especies microbianas por similitudes en su metabolismo y en otro se combinó con un test estadístico para identificar proteínas relevantes en aprendizaje y memoria.

Por tanto, esta tesis es en su conjunto el producto resultante de un proyecto multidisciplinario categorizado dentro de los campos de bioinformática y biología computacional. Propone técnicas de inteligencia artificial para resolver tres problemas de compleja solución dentro de la biología molecular y en los que el metabolismo está siempre involucrado. Debido a que el metabolismo está presente en diferentes niveles biológicos y puede ser estudiado desde distintas perspectivas, a lo largo de este trabajo se han explorado diferentes áreas específicas de investigación mediante técnicas de aprendizaje automático como son: la regulación a nivel enzimático de redes metabólicas, la determinación de características fenotípicas en especies microbianas a partir de sus estructuras metabólicas y el análisis de expresión de proteínas, consecuencia de la regulación génica.

Bibliography

Ahmed, M. M., Dhanasekaran, A. R., Block, A., Tong, S., Costa, A. C. S. and Gardiner, K. J. Protein profiles associated with context fear conditioning and their modulation by memantine. *Molecular & cellular proteomics : MCP*, vol. 13(4), pages 919–37, 2014. ISSN 1535-9484.

Ahmed, M. M., Sturgeon, X., Ellison, M., Davisson, M. T. and Gardiner, K. J. Loss of correlations among proteins in brains of the Ts65Dn mouse model of down syndrome. *Journal of proteome research*, vol. 11(2), pages 1251–63, 2012. ISSN 1535-3907.

Andrés-Toro, B., Girón-Sierra, J. M., Fernández-Blanco, P., López-Orozco, J. A. and Besada-Portas, E. Multiobjective optimization and multivariable control of the beer fermentation process with the use of evolutionary algorithms. *Journal of Zhejiang University. Science*, vol. 5(4), pages 378–389, 2004. ISSN 1009-3095.

Banga, J. R. Optimization in computational systems biology. *BMC systems biology*, vol. 2(1), page 47, 2008. ISSN 1752-0509.

Bao, L. and Cui, Y. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics*, vol. 21(10), pages 2185–2190, 2005.

Berg, J., Tymoczko, J. L. and Stryer, L. *Biochemistry*. W. H. Freedman and Company., New York, USA, sixth edition, 2006. ISBN 978-0716787242.

Bezdek, J. C., Ehrlich, R. and Full, W. Fcm: The fuzzy c-means clustering algorithm. *Computers and Geosciences*, vol. 10(2-3), pages 191–203, 1984.

Boada, R., Hutaff-Lee, C., Schrader, A., Weitzenkamp, D., Benke, T. A., Goldson, E. J. and Costa, A. C. S. Antagonism of NMDA

receptors as a potential treatment for Down syndrome: a pilot randomized controlled trial. *Translational psychiatry*, vol. 2, page e141, 2012. ISSN 2158-3188.

BOERLIJST, M. and HOGEWEG, P. Spiral wave structure in pre-biotic evolution: Hypercycles stable against parasites. 1991.

BOHLIN, J., SKJERVE, E. and USSERY, D. W. Analysis of genomic signatures in prokaryotes using multinomial regression and hierarchical clustering. *BMC genomics*, vol. 10(1), page 487, 2009. ISSN 1471-2164.

BROHÉE, S. and VAN HELDEN, J. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC bioinformatics*, vol. 7(1), page 488, 2006. ISSN 1471-2105.

CARTER, R. J., DUBCHAK, I. and HOLBROOK, S. R. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic acids research*, vol. 29(19), pages 3928–38, 2001. ISSN 1362-4962.

CASASNOVAS, J., CLEMENTE, J. C., MIRÓ-JULIA, J., ROSSELLÓ, F., SATOU, K. and VALIENTE, G. Fuzzy clustering improves phylogenetic relationships reconstruction from metabolic pathways. In *Proc. of the 11th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. 2006.

CHUBUKOV, V., ZULETA, I. A. and LI, H. Regulatory architecture determines optimal regulation of gene expression in metabolic pathways. 2012.

CLEMENTE, J. C., SATOU, K. and VALIENTE, G. Reconstruction of phylogenetic relationships from metabolic pathways based on the enzyme hierarchy and the gene ontology. *Genome informatics. International Conference on Genome Informatics*, vol. 16(2), pages 45–55, 2005. ISSN 0919-9454.

COLORNI, A., DORIGO, M. and VITTORIO, M. Distributed Optimization by Ant Colonies. In *ECAL91-EUROPEAN CONFERENCE ON ARTIFICIAL LIFE*, pages 134–142. Elsevier Publishing, Paris, France, 1991.

COSTA, A. C. S., SCOTT-MCKEAN, J. J. and STASKO, M. R. Acute injections of the NMDA receptor antagonist memantine rescue performance deficits of the Ts65Dn mouse model of Down syndrome on a fear conditioning test. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, vol. 33(7), pages 1624–32, 2008. ISSN 0893-133X.

CROSBY, J. L. ET AL. *Computer simulation in genetics.*. John Wiley and Sons., London, New York, Sidney, Toronto, 1973. ISBN 0-471-18880-8.

CUTELLO, V., NARZISI, G. and NICOSIA, G. A multi-objective evolutionary approach to the protein structure prediction problem. *Journal of the Royal Society, Interface / the Royal Society*, vol. 3(6), pages 139–151, 2006. ISSN 1742-5689.

CYPESS, A. M., WHITE, A. P., VERNOCHET, C., SCHULZ, T. J., XUE, R., SASS, C. A., HUANG, T. L., ROBERTS-TOLER, C., WEINER, L. S., SZE, C., CHACKO, A. T., DESCHAMPS, L. N., HERDER, L. M., TRUCHAN, N., GLASGOW, A. L., HOLMAN, A. R., GAVRILA, A., HASSELGREN, P.-O., MORI, M. A., MOLLA, M. and TSENG, Y.-H. Anatomical localization, gene expression profiling and functional characterization of adult human neck brown fat. *Nature medicine*, vol. 19(5), pages 635–9, 2013. ISSN 1546-170X.

DAS, I. and DENNIS, J. E. Normal-Boundary Intersection: A New Method for Generating the Pareto Surface in Nonlinear Multicriteria Optimization Problems. *SIAM Journal on Optimization*, vol. 8(3), pages 631–657, 1998. ISSN 1052-6234.

DAVIES, D. L. and BOULDIN, D. W. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1(2), pages 224–227, 1979. ISSN 0162-8828.

DAY, R., ZYDALLIS, J., LAMONT, G. and PACHTER, R. Solving the Protein Structure Prediction Problem Through a Multiobjective Genetic Algorithm. 2002.

DAYHOFF, M. O. Computer aids to protein sequence determination. *Journal of Theoretical Biology*, 1965.

DAYHOFF, M. O. Computer analysis of protein evolution. *Scientific American*, vol. 221(1), pages 86–95, 1969. ISSN 0036-8733.

DAYHOFF, M. O., ECK, R. V., CHANG, M. A. and SOCHARD, M. R. *Atlas of Protein Sequence and Structure*. Silver Spring, Md. National Biomedical Research Foundation, 1965.

DAYHOFF, M. O. and LEDLEY, R. S. Comprotein. In *Proceedings of the December 4-6, 1962, fall joint computer conference on - AFIPS '62 (Fall)*, pages 262–274. ACM Press, New York, New York, USA, 1962.

DEB, K. *Multi-Objective Optimization using Evolutionary Algorithms*, vol. 16. John Wiley & Sons, 2001. ISBN 978-0-471-87339-6.

DEB, K. and GUPTA, S. Understanding knee points in bicriteria problems and their implications as preferred solution principles. *Engineering Optimization*, vol. 43(11), pages 1175–1204, 2011. ISSN 0305-215X.

DEB, K., PRATAP, A., AGARWAL, S. and MEYARIVAN, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, vol. 6(2), 2002. ISSN 1089-778X.

EBENHÖH, O. and HEINRICH, R. Evolutionary optimization of metabolic pathways. Theoretical reconstruction of the stoichiometry of ATP and NADH producing systems. *Bulletin of mathematical biology*, vol. 63(1), pages 21–55, 2001. ISSN 00928240.

ECK, R. and DAYHOFF, M. O. *Atlas of protein sequence and structure, 1966*. National Biomedical Research Foundation, Silver Spring, Maryland, 1966.

EGEA, J. A., MARTÍ, R. and BANGA, J. R. An evolutionary method for complex-process optimization. *Computers & Operations Research*, vol. 37(2), pages 315–324, 2010. ISSN 03050548.

ESMAEILI, A. and JACOB, C. A multi-objective differential evolutionary approach toward more stable gene regulatory networks. *Biosystems*, vol. 98(3), pages 127–36, 2009. ISSN 1872-8324.

FALKENAUER, E. and MARCHAND, A. Clustering microarray data with evolutionary algorithms. In: Gary B. Fogel, David W. Corne (eds). *Evolutionary Computation in Bioinformatics. Morgan Kaufmann*, pages 219–30, 2002.

FANSELOW, M. S. Factors governing one-trial contextual conditioning. *Animal Learning & Behavior*, vol. 18(3), pages 264–270, 1990. ISSN 0090-4996.

FITCH, W. M. An improved method of testing for evolutionary homology. *Journal of molecular biology*, vol. 16(1), pages 9–16, 1966.

FITCH, W. M. and MARGOLIASH, E. Construction of phylogenetic trees. *Science (New York, N.Y.)*, vol. 155(760), pages 279–284, 1967. ISSN 0036-8075.

FOGEL, G. B., PORTO, V. W., WEEKES, D. G., FOGEL, D. B., GRIFFEY, R. H., MCNEIL, J. A., LESNIK, E., ECKER, D. J. and SAMPATH, R. Discovery of RNA structural elements using evolutionary computation. 2002.

FRASER, A. S. Simulation of genetic systems by automatic digital computers. I. Introduction. *Australian Journal of Biological Science*, vol. 10, pages 484 – 491, 1957.

FRASER, A. S. and BURNELL, D. *Computer Models in Genetics*. McGraw-Hill, New York, New York, USA, 1970. ISBN 0-07-021904-4.

GEVERS, D., COHAN, F. M., LAWRENCE, J. G., SPRATT, B. G., COENYE, T., FEIL, E. J., STACKEBRANDT, E., VAN DE PEER, Y., VANDAMME, P., THOMPSON, F. L. and SWINGS, J. Re-evaluating prokaryotic species. *Nature Reviews Microbiology*, vol. 3(9), pages 733–9, 2005. ISSN 1740-1526.

GILISSEN, C., HEHIR-KWA, J. Y., THUNG, D. T., VAN DE VORST, M., VAN BON, B. W. M., WILLEMSEN, M. H., KWINT, M., JANSSEN, I. M., HOISCHEN, A., SCHENCK, A., LEACH, R., KLEIN, R., TEARLE, R., BO, T., PFUNDT, R., YNTEMA, H. G., DE VRIES, B. B. A., KLEEFSTRA, T., BRUNNER, H. G., VISSERS, L. E. L. M. and VELTMAN, J. A. Genome sequencing identifies major causes of severe intellectual disability. *Nature*, vol. 511(7509), pages 344–7, 2014. ISSN 1476-4687.

GILMAN, A. and ROSS, J. Genetic-algorithm selection of a regulatory structure that directs flux in a simple metabolic model. *Biophysical journal*, vol. 69(4), pages 1321–1333, 1995.

GOLDBERG, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989. ISBN 0201157675.

GOODACRE, R. Making sense of the metabolome using evolutionary computation: seeing the wood with the trees. *Journal of experimental botany*, vol. 56(410), pages 245–254, 2005. ISSN 0022-0957.

GUO, H., MENG, Y. and JIN, Y. A cellular mechanism for multi-robot construction via evolutionary multi-objective optimization of a gene regulatory network. *Bio Systems*, vol. 98(3), pages 193–203, 2009. ISSN 1872-8324.

HAGEN, J. B. The origins of bioinformatics. *Nature reviews. Genetics*, vol. 1(3), pages 231–6, 2000. ISSN 1471-0056.

HALKIDI, M., BATISTAKIS, Y. and VAZIRGIANNIS, M. On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, vol. 17(2-3), pages 107–145, 2001. ISSN 0925-9902.

HALSALL-WHITNEY, H., TAYLOR, D. and THIBAULT, J. Multicriteria optimization of gluconic acid production using net flow. *Bioprocess and biosystems engineering*, vol. 25(5), pages 299–307, 2003. ISSN 1615-7591.

HANDL, J. and KNOWLES, J. An Evolutionary Approach to Multiobjective Clustering. *IEEE Transactions on Evolutionary Computation*, vol. 11(1), pages 56–76, 2007. ISSN 1089-778X.

HANDL, J., KNOWLES, J. and KELL, D. B. Computational cluster validation in post-genomic data analysis. *Bioinformatics (Oxford, England)*, vol. 21(15), pages 3201–12, 2005. ISSN 1367-4803.

HERRERO, J., VALENCIA, A. and DOPAZO, J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics (Oxford, England)*, vol. 17(2), pages 126–136, 2001. ISSN 1460-2059.

HESPER, B. and HOGEWEG, P. Bioinformatica: een werkconcept. *Kameleon 1.6*, pages 28–29, 1970.

HIGUERA, C., PAJARES, G., TAMAMES, J. and MORÁN, F. Expert system for clustering prokaryotic species by their metabolic features. *Expert Systems with Applications*, vol. 40(15), pages 6185–6194, 2013. ISSN 09574174.

HIGUERA, C., VILLAVERDE, A. F., BANGA, J. R., ROSS, J. and MORÁN, F. Multi-criteria optimization of regulation in metabolic networks. *PloS one*, vol. 7(7), page e41122, 2012. ISSN 1932-6203.

HOGEWEG, P. Iterative character weighing in numerical taxonomy. *Computers in biology and medicine*, vol. 6(3), pages 199–211, 1976. ISSN 0010-4825.

HOGEWEG, P. Simulating the growth of cellular forms. *SIMULATION*, vol. 31(3), pages 90–96, 1978. ISSN 0037-5497.

HOGEWEG, P. Cellular automata as a paradigm for ecological modeling. *Applied Mathematics and Computation*, vol. 27(1), pages 81–100, 1988. ISSN 00963003.

HOGEWEG, P. The roots of bioinformatics in theoretical biology. *PLoS computational biology*, vol. 7(3), page e1002021, 2011. ISSN 1553-7358.

HOGEWEG, P. and HESPER, B. Interactive instruction on population interactions. *Computers in biology and medicine*, vol. 8(4), pages 319–27, 1978. ISSN 0010-4825.

HOLLAND, J. H. *Adaptation in natural and artificial systems*. MIT Press, 1992. ISBN 0-262-58111-6.

HONG, S. H., KIM, T. Y. and LEE, S. Y. Phylogenetic analysis based on genome-scale metabolic pathway reaction content. *Applied microbiology and biotechnology*, vol. 65(2), pages 203–10, 2004. ISSN 0175-7598.

HUNT, L. T. Margaret O. Dayhoff 1925-1983. *DNA (Mary Ann Liebert, Inc.)*, vol. 2(2), pages 97–8, 1983. ISSN 0198-0238.

IERAPETRITOU, M., SHARMA, N. and YARMUSH, M. L. Novel quantitative tools for engineering analysis of hepatocyte cultures used in bioartificial liver systems. *Computer Aided Chemical Engineering*, vol. 18(C), pages 1057–1062, 2004.

IRVING, C., BASU, A., RICHMOND, S., BURN, J. and WREN, C. Twenty-year trends in prevalence and survival of Down syndrome. *European journal of human genetics : EJHG*, vol. 16(11), pages 1336–40, 2008. ISSN 1018-4813.

JAGGA, Z. and GUPTA, D. Supervised learning classification models for prediction of plant virus encoded RNA silencing suppressors. *PloS one*, vol. 9(5), page e97446, 2014. ISSN 1932-6203.

JAIN, G., WANG, H., LIAO, L. and BOYD, E. F. Genomic Comparison of Bacterial Species Based on Metabolic Characteristics. In *2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*, pages 77–83. IEEE, 2009. ISBN 978-0-7695-3739-9.

KANEHISA, M. The KEGG database. *Novartis Foundation Symposium*, vol. 247, pages 91–101; discussion 101–103, 119–128, 244–252, 2002.

KEITH, J. M., ADAMS, P., BRYANT, D., KROESE, D. P., MITCHELSON, K. R., COCHRAN, D. A. E. and LALA, G. H. A simulated annealing algorithm for finding consensus sequences. *Bioinformatics (Oxford, England)*, vol. 18(11), pages 1494–1499, 2002. ISSN 1367-4803.

KIKUCHI, S., TOMINAGA, D., ARITA, M., TAKAHASHI, K. and TOMITA, M. Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics*, vol. 19(5), pages 643–650, 2003. ISSN 1367-4803.

KIM, S. Protein beta-turn prediction using nearest-neighbor method. *Bioinformatics (Oxford, England)*, vol. 20(1), pages 40–44, 2004. ISSN 1460-2059.

KIRKPATRICK, S., GELATT, C. D. and VECCHI, M. P. Optimization by simulated annealing. *Science (New York, N.Y.)*, vol. 220(4598), pages 671–80, 1983. ISSN 0036-8075.

KOHONEN, T. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, vol. 43(1), pages 59–69, 1982.

KOHONEN, T. Learning vector quantization. *Neural Networks*, vol. 1(1), pages 303–309, 1988.

KOSKI, J. and SILVENNOINEN, R. Norm methods and partial weighting in multicriterion optimization of structures. *International Journal for Numerical Methods in Engineering*, vol. 24(6), pages 1101–1121, 1987. ISSN 0029-5981.

KRASNOGOR, N., BLACKBURNE, B. P., BURKE, E. K. and HIRST, J. D. Multimeme Algorithms for Protein Structure Prediction. *Proceedings of the Parallel Problem Solving from Nature VII. Lecture Notes in Computer Science*, vol. 2439, pages 769–778, 2002.

VAN LAARHOVEN, P. and AARTS, E. *Simulated Annealing: Theory and Applications*. Springer, Reidel, Dordrecht, The Nederlands, 1987.

LANCE, G. N. and WILLIAMS, W. T. A Generalized Sorting Strategy for Computer Classifications. *Nature*, vol. 212(5058), pages 218–218, 1966. ISSN 0028-0836.

LANNING, O. J., HABERSHON, S., HARRIS, K. D., JOHNSTON, R. L., KARIUKI, B. M., TEDESCO, E. and TURNER, G. W. Definition of a guiding function in global optimization: a hybrid approach combining energy and R-factor in structure solution from powder diffraction data. *Chemical Physics Letters*, vol. 317(3-5), pages 296–303, 2000. ISSN 00092614.

LARRANAGA, P., CALVO, B., SANTANA, R., BIELZA, C., GALDIANO, J., INZA, I., LOZANO, J. A., ARMANANZAS, R., SANTAFE, G., PEREZ, A. and ROBLES, V. Machine learning in bioinformatics. *Briefings in Bioinformatics*, vol. 7(1), pages 86–112, 2006. ISSN 1467-5463.

LARSEN, P. E., FIELD, D. and GILBERT, J. A. Predicting bacterial community assemblages using an artificial neural network approach. *Nature methods*, vol. 9(6), pages 621–5, 2012. ISSN 1548-7105.

LEE, C. C., LO, W. C., LAI, S. M., CHEN, P., YI PING, TANG, C. Y. and LYU, P. C. Metabolic classification of microbial genomes using functional probes. *BMC genomics*, vol. 13(1), page 157, 2012. ISSN 1471-2164.

LEE, S. Y., LEE, D.-Y. and KIM, T. Y. Systems biotechnology for strain improvement. *Trends in biotechnology*, vol. 23(7), pages 349–58, 2005. ISSN 0167-7799.

LEE, Z.-J., SU, S.-F., CHUANG, C.-C. and LIU, K.-H. Genetic algorithm with ant colony optimization (GA-ACO) for multiple sequence alignment. *Applied Soft Computing*, vol. 8(1), pages 55–78, 2008. ISSN 15684946.

LEONARD, H. and WEN, X. The epidemiology of mental retardation: challenges and opportunities in the new millennium. *Mental retardation and developmental disabilities research reviews*, vol. 8(3), pages 117–34, 2002. ISSN 1080-4013.

LIU, I. Y. C., LYONS, W. E., MAMOUNAS, L. A. and THOMPSON, R. F. Brain-derived neurotrophic factor plays a critical role in contextual fear conditioning. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 24(36), pages 7958–63, 2004. ISSN 1529-2401.

LÓPEZ-BIGAS, N. and OUZOUNIS, C. A. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Research*, vol. 32(10), pages 3108–3114, 2004.

LORENZO-REDONDO, R., DELGADO, S., MORÁN, F. and LOPEZ-GALINDEZ, C. Realistic three dimensional fitness landscapes generated by self organizing maps for the analysis of experimental HIV-1 evolution. *PloS one*, vol. 9(2), page e88579, 2014. ISSN 1932-6203.

MACNAUGHTON-SMITH, P., WILLIAMS, W. T., DALE, M. B. and MOCKETT, L. G. Dissimilarity Analysis: a new Technique of Hierarchical Subdivision. *Nature*, vol. 202(4936), pages 1034–1035, 1964. ISSN 0028-0836.

MARCHISIO, M. A. and STELLING, J. Computational design tools for synthetic biology. *Current opinion in biotechnology*, vol. 20(4), pages 479–485, 2009. ISSN 09581669.

MELÉNDEZ-HEVIA, E., WADDELL, T. G. and MONTERO, F. Optimization of metabolism: the evolution of metabolic pathways towards simplicity through the game of the pentose phosphate cycle. *Journal of theoretical Biology*, vol. 166, pages 201–220, 1994. ISSN 00225193.

MENDES, P. and KELL, D. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics (Oxford, England)*, vol. 14(10), pages 869–883, 1998. ISSN 1367-4803.

MIETTINEN, K. *Nonlinear multiobjective optimization*, vol. 12. Kluwer Academic Publishers, 1999. ISBN 978-0-7923-8278-2.

MOLES, C. G., MENDES, P. and BANGA, J. R. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome research*, vol. 13(11), pages 2467–74, 2003. ISSN 1088-9051.

MORÁN, F. and GOLDBETER, A. Onset of birhythmicity in a regulated biochemical system. *Biophysical chemistry*, vol. 20(1-2), pages 149–156, 1984. ISSN 03014622.

MURANEN, T. A., GRECO, D., FAGERHOLM, R., KILPIVAARA, O., KÄMPJÄRVI, K., AITTOMÄKI, K., BLOMQVIST, C., HEIKKILÄ, P., BORG, A. and NEVANLINNA, H. Breast tumors from CHEK2 1100delC-mutation carriers: genomic landscape and clinical implications. *Breast cancer research : BCR*, vol. 13(5), page R90, 2011. ISSN 1465-542X.

NEEDLEMAN, S. B. and WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, vol. 48(3), pages 443–453, 1970. ISSN 00222836.

NGUYEN, H. D., YOSHIHARA, I., YAMAMORI, K. and YASUNAGA, M. Aligning multiple protein sequences by parallel hybrid genetic algorithm. *Genome informatics. International Conference on Genome Informatics*, vol. 13, pages 123–132, 2002. ISSN 0919-9454.

NIELSEN, J. Principles of optimal metabolic network operation. *Molecular systems biology*, vol. 3, page 126, 2007. ISSN 1744-4292.

NISHIZUKA, S., CHARBONEAU, L., YOUNG, L., MAJOR, S., REINHOLD, W. C., WALTHAM, M., KOUROS-MEHR, H., BUSSEY, K. J., LEE, J. K., ESPINA, V., MUNSON, P. J., PETRICOIN, E., LIOTTA, L. A. and WEINSTEIN, J. N. Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100(24), pages 14229–34, 2003. ISSN 0027-8424.

OLIVARES, D., DESHPANDE, V. K., SHI, Y., LAHIRI, D. K., GREIG, N. H., ROGERS, J. T. and HUANG, X. N-methyl D-aspartate (NMDA) receptor antagonists and memantine treatment for Alzheimer's disease, vascular dementia and Parkinson's disease. *Current Alzheimer research*, vol. 9(6), pages 746–58, 2012. ISSN 1875-5828.

PAJARES, G. and DE LA CRUZ, J. *Aprendizaje automático. Un enfoque práctico*. RA-MA Editorial, 2010. ISBN 8476534604.

PARK, L. J., PARK, C. H., PARK, C. and LEE, T. Application of genetic algorithms to parameter estimation of bioprocesses. *Medical & biological engineering & computing*, vol. 35(1), pages 47–49, 1997. ISSN 0140-0118.

PARK, Y.-S., CÉRÉGHINO, R., COMPIN, A. and LEK, S. Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. *Ecological Modelling*, vol. 160(3), pages 265–280, 2003. ISSN 03043800.

RABOW, A. A., SHOEMAKER, R. H., SAUSVILLE, E. A. and COVELL, D. G. Mining the National Cancer Institute's Tumor-Screening Database: Identification of Compounds with Similar Cellular Activities. *Journal of Medicinal Chemistry*, vol. 45(4), pages 818–840, 2002. ISSN 0022-2623.

RADULOVIC, J., KAMMERMEIER, J. and SPIESS, J. Generalization of fear responses in C57BL/6N mice subjected to one-trial foreground contextual fear conditioning. *Behavioural brain research*, vol. 95(2), pages 179–89, 1998. ISSN 0166-4328.

RECHENBERG, I. *Evolutionsstrategie; Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog Verlag, Stuttgart, 2nd editio edition, 1973. ISBN 978-3772803734.

RODRIGUEZ-FERNANDEZ, M., MENDES, P. and BANGA, J. R. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Bio Systems*, vol. 83(2-3), pages 248–65, 2006. ISSN 0303-2647.

ROGERS, A., SMITH, M. J., DOOLAN, P., CLARKE, C., CLYNES, M., MURPHY, J. F., MCDERMOTT, A., SWAN, N., CROTTY, P., RIDGWAY, P. F. and CONLON, K. C. Invasive markers identified by gene expression profiling in pancreatic cancer. *Pancreatology : official journal of the International Association of Pancreatology (IAP) ... [et al.]*, vol. 12(2), pages 130–40, 2012. ISSN 1424-3911.

ROSENBLATT, F. *Principles of neurodynamics; perceptrons and the theory of brain mechanisms*. Washington, Spartan Books, 1962.

ROSIN, G., HANNELIUS, U., LINDSTRÖM, L., HALL, P., BERGH, J., HARTMAN, J. and KERE, J. The dyslexia candidate gene DYX1C1 is a potential marker of poor survival in breast cancer. *BMC cancer*, vol. 12, page 79, 2012. ISSN 1471-2407.

RUEDA, N., FLÓREZ, J. and MARTÍNEZ-CUÉ, C. Mouse models of Down syndrome as a tool to unravel the causes of mental disabilities. *Neural plasticity*, vol. 2012, page 584071, 2012. ISSN 1687-5443.

SALAMOV, A. A. and SOLOVYEV, V. V. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *Journal of molecular biology*, vol. 247(1), pages 11–15, 1995. ISSN 0022-2836.

SANDER, J., QIN, X., LU, Z., NIU, N. and KOVARSKY, A. Automatic Extraction of Clusters from Hierarchical Clustering Representations. *Advances in knowledge discovery and data mining*, vol. 2637, pages 75–87, 2003.

SCHUETZ, R., KUEPFER, L. and SAUER, U. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Molecular systems biology*, vol. 3, page 119, 2007.

SCHWEFEL, H. P. *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie Systems Research*. Birkhäuser Verlag, Basel, german edition, 1977. ISBN 978-3764308766.

SENDÍN, J. O. H., ALONSO, A. A. and BANGA, J. R. Multi-objective optimization of biological networks for prediction of intracellular fluxes. *Advances in Soft Computing*, vol. 49, pages 197–205, 2009.

SENDÍN, J. O. H., EXLER, O. and BANGA, J. R. Multi-objective mixed integer strategy for the optimisation of biological networks. *IET systems biology*, vol. 4(3), pages 236–248, 2010. ISSN 17518849.

SENDÍN, J. O. H., VERA, J., TORRES, N. V. and BANGA, J. R. Model based optimization of biochemical systems using multiple objectives: a

comparison of several solution strategies. *Mathematical and Computer Modelling of Dynamical Systems*, vol. 12(5), pages 469–487, 2006.

SHENG, Q., MOREAU, Y. and DE MOOR, B. Biclustering microarray data by Gibbs sampling. *Bioinformatics*, vol. 19(SUPPL. 2), pages 196–205, 2003.

SHOVAL, O., SHEFTEL, H., SHINAR, G., HART, Y., RAMOTE, O., MAYO, A., DEKEL, E., KAVANAGH, K. and ALON, U. Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science (New York, N.Y.)*, vol. 336(6085), pages 1157–60, 2012. ISSN 1095-9203.

SMITH, J. The Co-Evolution of Memetic Algorithms for Protein Structure Prediction. In: William WH, Krasnogor N, Smith JE (eds). *Recent Advances in Memetic Algorithms, Studies in Fuzziness and Soft Computing.*, pages 105–128, 2004.

SØ RLIE, T., PEROU, C. M., TIBSHIRANI, R., AAS, T., GEISLER, S., JOHNSEN, H., HASTIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., THORSEN, T., QUIST, H., MATESE, J. C., BROWN, P. O., BOTSTEIN, D., LØ NNING, P. E. and BØ RRESEN DALE, A. L. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98(19), pages 10869–74, 2001. ISSN 0027-8424.

SPELLMAN, P. T., SHERLOCK, G., ZHANG, M. Q., IYER, V. R., ANDERS, K., EISEN, M. B., BROWN, P. O., BOTSTEIN, D. and FUTCHER, B. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, vol. 9(12), pages 3273–3297, 1998. ISSN 1059-1524.

STEGMAYER, G., GERARD, M. and MILONE, D. Data Mining Over Biological Datasets: An Integrated Approach Based on Computational Intelligence. *IEEE Computational Intelligence Magazine*, vol. 7(4), pages 22–34, 2012. ISSN 1556-603X.

STIEDL, O., RADULOVIC, J., LOHMANN, R., BIRKENFELD, K., PALVE, M., KAMMERMEIER, J., SANANBENESI, F. and SPIESS, J. Strain and substrain differences in context- and tone-dependent fear conditioning of inbred mice. *Behavioural brain research*, vol. 104(1-2), pages 1–12, 1999. ISSN 0166-4328.

STRASSER, B. J. Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954-1965. *Journal of the History of Biology*, vol. 43(4), pages 623–660, 2010.

STURGEON, X. and GARDINER, K. J. Transcript catalogs of human chromosome 21 and orthologous chimpanzee and mouse regions. *Mammalian genome : official journal of the International Mammalian Genome Society*, vol. 22(5-6), pages 261–71, 2011. ISSN 1432-1777.

SUEN, G., GOLDMAN, B. S. and WELCH, R. D. Predicting prokaryotic ecological niches using genome sequence analysis. *PloS one*, vol. 2(8), page e743, 2007. ISSN 1932-6203.

SUN, J., GARIBALDI, J. M. and HODGMAN, C. Parameter estimation using meta-heuristics in systems biology: a comprehensive review. *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 9(1), pages 185–202, 2012. ISSN 1557-9964.

SWEATT, J. D. The neuronal MAP kinase cascade: a biochemical signal integration system subserving synaptic plasticity and memory. *Journal of neurochemistry*, vol. 76(1), pages 1–10, 2001. ISSN 0022-3042.

SZALENIEC, M. Prediction of enzyme activity with neural network models based on electronic and geometrical features of substrates. *Pharmacological reports : PR*, vol. 64(4), pages 761–81, 2012. ISSN 1734-1140.

TAMAYO, P., SLONIM, D., MESIROV, J., ZHU, Q., KITAREEWAN, S., DMITROVSKY, E., LANDER, E. S. and GOLUB, T. R. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96(6), pages 2907–2912, 1999.

TANAKA, S., MOGUSHI, K., YASEN, M., NOGUCHI, N., KUDO, A., NAKAMURA, N., ITO, K., MIKI, Y., INAZAWA, J., TANAKA, H. and ARII, S. Gene-expression phenotypes for vascular invasiveness of hepatocellular carcinomas. *Surgery*, vol. 147(3), pages 405–14, 2010. ISSN 1532-7361.

TKACIK, G., WALCZAK, A. M. and BIALEK, W. Optimizing information flow in small genetic networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 80(3 Pt 1), page 031920, 2009.

VAN SOMEREN, E., WESSELS, L., BACKER, E. and REINDERS, M. Multi-criterion optimization for genetic network modeling. *Signal Processing*, vol. 83(4), pages 763–775, 2003. ISSN 01651684.

VARRAS, M., GRIVA, T., KALLES, V., AKRIVIS, C. and PAPARISTEIDIS, N. Markers of stem cells in human ovarian granulosa cells: is there a clinical significance in ART? *Journal of ovarian research*, vol. 5(1), page 36, 2012. ISSN 1757-2215.

VEYRAC, A., BESNARD, A., CABOCHE, J., DAVIS, S. and LAROCHE, S. The transcription factor Zif268/Egr1, brain plasticity, and memory. *Progress in molecular biology and translational science*, vol. 122, pages 89–129, 2014. ISSN 1878-0814.

WALCZAK, A. M., TKACIK, G. and BIALEK, W. Optimizing information flow in small genetic networks. II. Feed-forward interactions. *Physical review. E*, vol. 81(4), page 041905, 2010. ISSN 1539-3755.

WISEMAN, F. K., ALFORD, K. A., TYBULEWICZ, V. L. J. and FISHER, E. M. C. Down syndrome–recent progress and future prospects. *Human molecular genetics*, vol. 18(R1), pages R75–83, 2009. ISSN 1460-2083.

WREN, J. D., YAO, M., LANGER, M. and CONWAY, T. Simulated annealing of microarray data reduces noise and enables cross-experimental comparisons. *DNA and cell biology*, vol. 23(10), pages 695–700, 2004. ISSN 1044-5498.

YI, T. M. and LANDER, E. S. Protein secondary structure prediction using nearest-neighbor methods. *Journal of molecular biology*, vol. 232(4), pages 1117–1129, 1993. ISSN 0022-2836.

ZELIKOWSKY, M., HERSMAN, S., CHAWLA, M. K., BARNES, C. A. and FANSELOW, M. S. Neuronal ensembles in amygdala, hippocampus, and prefrontal cortex track differential components of contextual fear. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 34(25), pages 8462–6, 2014. ISSN 1529-2401.