UNIVERSIDAD DE GRANADA

NUEVOS MÉTODOS DE PREDICCIÓN DE INTERACCIÓN DE PROTEÍNA-PROTEÍNA UTILIZANDO SISTEMAS INTELIGENTES EN BASES DE DATOS DE PROTEÓMICA

TESIS DOCTORAL

José Miguel Urquiza Ortiz

September 12, 2011

Departamento de Arquitectura y Tecnología de Computadores

Editor: Editorial de la Universidad de Granada
Autor: José Miguel Urquiza Ortiz
D.L.: GR 1067-2012
ISBN: 978-84-695-1074-2

UNIVERSIDAD DE GRANADA

NUEVOS MÉTODOS DE PREDICCIÓN DE INTERACCIÓN DE PROTEÍNA-PROTEÍNA UTILIZANDO SISTEMAS INTELIGENTES EN BASES DE DATOS DE PROTEÓMICA

Memoria presentada por

José Miguel Urquiza Ortiz

Para optar al grado de

DOCTOR EUROPEO EN INGENIERÍA INFORMÁTICA

Fdo. José Miguel Urquiza Ortiz

D. Ignacio Rojas Ruiz, Catedrático de Universidad, **D. Héctor Emilio Pomares Cintas**, Profesor Titular de Universidad y **D. Luis Javier Herrera Maldonado**, Profesor Contratado-Doctor del Departamento de Arquitectura y Tecnología de Computadores

CERTIFICAN

Que la memoria titulada: **"NUEVOS MÉTODOS DE PREDICCIÓN DE INTERACCIÓN DE PROTEÍNA-PROTEÍNA UTILIZANDO SISTEMAS INTELIGENTES EN BASES DE DATOS DE PROTEÓMICA**

" ha sido realizada por **D. José Miguel Urquiza Ortiz** bajo nuestra dirección en el Departamento de Arquitectura y Tecnología de Computadores de la Universidad de Granada para optar al grado de Doctor Europeo en Ingeniería Informática.

Granada, 12 de septiembre de 2011

Fdo. Ignacio Rojas Ruiz	Héctor Emilio Pomares Cintas	Luis Javier Herrera Maldonado
Director de la Tesis	Director de la Tesis	Director de la Tesis

RESUMEN

Es evidente que el siglo XX fue el siglo de la Física, en este comienzo de siglo XXI todo parece apuntar que será el siglo de la Biología. Queda patente este hecho si observamos la inversión realizada en la investigación de los campos que conforman la Biología; este esfuerzo es desempeñado tanto por instituciones públicas como por privadas.

Uno de los puntos de partida fundamentales ha sido la secuenciación del genoma completo de difererentes organismos, lo que nos invita a mejorar nuestro conocimiento sobre la biología de estos organismos. En los últimos años, hemos oido hablar de la genómica y de la emergente disciplina de la proteómica. Por tanto, daremos una definición de ambos términos en este trabajo. Realizando un paseo y explicando términos básicos y fundamentales para que finalmente nos empapamos del tremendo campo de la bioinformática aplicada a la proteómica. En concreto, esta tesis se centra en el estudio de las interacciones proteína-proteína, ya que estas desempeñan un papel fundamental en el control de los principales procesos celulares y su análisis permitiría un mejor entendimiento de los mecanismos celulares. Todo ello permitiería el desarrollo de mejores terapias y tratamientos en determinadas enfermedades de actualidad.

Por ello, en este trabajo de investigación se proponen métodos de predicción de interacción proteína-proteína mediante información genética y proteómica, utilizando métodos de aprendizaje supervisado y de reconocimiento de patrones. Así tendremos como resultado la obtención de las mejores características que definan este problema y la consecución de un modelo predictor de interacciones que permita la validación y ayuda en la experimentación. Además, un enfonque más específico fué aplicado para la caracterización de asociaciones moleculares complejas como es el caso del sistema ubiquitina proteasoma que tan íntimamente está relacionado con ciertas enfermedades neurodegenerativas.

Aunque en principio, impacte al lector, poco a poco, a lo largo de este trabajo, comprenderá la complejidad de la rama, hasta entender completamente el objetivo fundamental de este estudio.

ABSTRACT

It is a well-known fact that the twentieth century was the Century of Physics, however, the twenty-first century points to be the Century of Biology. A huge effort in investments from public and private organizations has already applied in order to research in the fields of Biology. One of the crucial starting point has been obtaining the completely sequenced genomes of different model organism. In recent years, genomics and the emerging discipline of proteomics have risen strongly. Here, in this thesis, a definition of both terms are described, explaining from the basic and fundamental terms until finishing within bioinformatics applied to proteomics. Specifically, this thesis focuses on the analysis of protein-protein interactions, which plays a fundamental role in the control of the main cellular processes. The analysis of such interactions allows a better knowledge of cellular mechanisms and open the possibility to develop improved therapies and treatments for important diseases.

Therefore, in this research work, methods for protein-protein interactions are proposed through proteomic and genomic information, using supervised learning and pattern recognition methods. Hence, this approach results a relevant set of features to define this problem and an protein-protein interaction predictor model to help in the validation and guide for experimentalists. In addition, with the purpose to face a specific problem an approach is proposed in an effort to characterize complex molecular associations such as in the ubiquitin-proteasome system which plays a fundamental role in the case of neurodegenerative diseases.

AGRADECIMIENTOS

A mis directores de tesis, Ignacio Rojas, Héctor Pomares y Luis Javier Herrera por su apoyo y dirección en la realización de esta tesis.

A mis padres (José y Francisca), hermanos (Pedro y Gema) y familia, ya estén en Granada, en Moratalla (mi tito Gabi) o en cualquier otro lugar.

Es una labor ardua e ímproba el hacer mención a todos quienes en cierta medida me han ayudado ya sea moralmente como desde el punto de vista técnico durante la realización de esta tesis, debería llenar varias hojas nombrándolos y si no aparecen en estos agradecimientos que no se sientan cohibidos ni olvidados, simplemente me he tomado la licencia de resumir esta sección sin la más mínima intención de infundir algún mal sentimiento a nadie. Realizada dicha aclaración, quisiera agradecer el apoyo de todos los compañeros del Departamento de Arquitectura y Tecnología de Computadores de la Universidad de Granada, desde los becarios Niceto, Juanma, Fran, Matteo, Ginés Rubio, Mario, Pablo, Gregorio, Ana Belén, Curro, Javi, Juanlu, Antonio, Karl, José Luis, David, Richard, Miguel Ángel, Raquel, Leo... hasta profesores como Alberto Guillén, Jesús González Peñalver y Manuel Rodríguez Álvarez por siempre estar dispuestos a escucharme en los momentos más difíciles y por su constante apoyo. No sólo he recibido el apoyo de este departamento sino de otros compañeros de la universidad como Diego Salas, Ignacio Álvarez, Míriam López, Juanma Górriz,... o de la Asociación de Jóvenes Investigadores Precarios como Pablo en otros.

A mis amigos de aquí y allá como Isidro y Alberto González, Alberto Ariza, José Miguel Sánchez, Roberto Vega y su hermano Gustavo que en paz descanse, Roberto y Germán Arroyo, Miguel Méndez, Víctor Fortis, María del Mar Padial, José María Ríos Pavón... a los que trabajaron conmigo en Intecna S.L. como Pedro Salido, Pedro Jimenez, Carlos, Mati, Fernando, Emilio, Ismael, Antonia...a Esperanza, Inma, Clara, Luís y Míriam, Jesús el químico, Sebas, ...a Leiron y Silvia, Francesco y Valentina, Pablo Porras y Elena, Juan Miguel y Pablo del Amo, ... y sin embargo me quedo corto nombrando.

Quisiera también hacer mención que sin la beca FPU AP2006-01748 no podría haberse llevado a cabo el desarrollo de esta tesis, haciendo también mención a la financiación de los proyectos del MEC CICYT SAF2010-20558 y Junta de Andalucía P07-TIC-02768 y P09-TIC-17547.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 SUMMARY OF THE PRESENT WORK

1.1 Introduction

It is a well-known fact that the twentieth century was the Century of Physics, however, the twenty-first century points to be the Century of Biology. One important objective of modern biology is the extraction of functional modules [240]. In the case of functional genomics, one of the main goal is to determine the function of genes predicted from the completely sequenced genomes [162].

In the last decades, it has been produced a massive quantity of biological data obtained from genome sequencing, transcriptomes, proteomes and interactomes. It represents a major challenge to integrate such relevant data sources to describe and give expression to the comprehensive knowledge of genes within and between genomes, which supply specialized information to explain the biological roles of the products of genes [162]. Therefore, genomics has already provided a huge amount of molecular interaction data, through which was constructed maps of specific cellular networks [96]. Such maps provide a valuable framework [202] to a better understanding of the mechanisms of protein function and cellular processes, and the design of new effective therapeutic approaches [240].

In the contemporary proteome research, one of most important targets is the elucidation of the structure, interactions, and functions of proteins that constitute cells and organisms [96]. Proteins participate in virtually every aspect of cellular function within an organism, however they act seldom alone, i.e, in isolation. Most proteins reach a specific function by interacting with other proteins [240]. Generally it is an accepted fact that most of the essential biological processes involve protein-protein interactions (PPIs), and exhaustive identifying them is crucial to systematically defining their cellular role [106].

Therefore protein-protein interactions (PPIs) play an essential role in almost

any biological function carried out within the cell and studying biological protein networks is of vital importance with the purpose of understanding the main cell mechanisms [181] [96] [80]. This reason motivates the development and improvement of technologies in the experimental techniques for detecting PPIs, such as mass spectrometry studies, yeast two-hybrid (Y2H) or co-imnoprecipitation (CoIP) [202] [216] [30] [234]. Nevertheless, computational approaches have been proposed in PPI prediction helping to experimentalist, used as guides in essays, saving cost and time generally associated with the experimental techniques [80].

However, it remains still difficult to precisely measure the quality and coverage of these interactome maps [162], what proportion of currently available interactome maps represents true biophysical interactions and what proportion represents artifacts [216]. In addition, supervised learning methods require gold standard positive (GSP) and negative (GSN) interaction sets, i.e, set of positive interacting and non-interacting pair of proteins. A more reliable set of PPI examples should enhance results from such methods [185]. There exist several methodologies with different degrees of accuracy and bias to create the GSP. The simplest way is to collect interactions that have been appeared in more than one experimental study; however the generated set is only useful for yeast [185]. Other methods select only those interactions exclusively obtained from small-scale experiments and interactions extracted from curated databases or interactions are present at protein complexes [145, 184].

Computational methodologies also generated reliable interaction sets using protein function and cellular localization information [230]. Nevertheless the application of these methodologies is required by annotation knowledge are required and so is limited to those species for which this information is available. There is the possibility to construct GSP sets using high-scoring interactions extracted from interaction confidence assignment schemes but there is circularity in these sets because they depend on other GSPs to yield confidence scores [185]. For example, Patil and Nakamura [162] proposed the creation of a GSP set combination of several genomic features, one of them is homology information. However, the verification of this set was based in the degree of overlapping with another GSP set.

Negative interactions sets (GSN) are also crucial for the prediction and confidence calculation of PPI sets. However negative interactions are unfortunately not reported by experimentalists in contrast with positive interactions [185]. It quite usual that computationally pairing proteins up as non-interactors selecting each pair of protein uniformly [173, 238, 232] or by protein pairs that do not share similar cellular compartments [106]. Other work suggests the creation of negative interactions using exclusively cellular compartment information can conduct to biased estimates of prediction accuracy and it could affect to several predictive methods [16]. For example Wu et al. [230] predicted negative interactions using a semantic similarity measure [222] based on the three ontologies of GO annotations. Nevertheless they did not distinguish which of the three ontologies contributed more to the determination of

negative interactions. Saeed and Deane [185] proposed an integrated method for generating GSN sets using functional, localization, expression and homology-based data. It was considered that an interaction between two proteins is tagged as a negative interaction only if the two interacting proteins have no overlap in any of these considered features. Yu et al. [233] presented an approach to picking an unbiased negative using a simply scoring pairs based on the frequency of the proteins in the considered dataset (positive set) can already achieve better than random performance.

1.2 Objectives

The main aim of the present research work is to provide a reliable, accurate and general methodology for predicting protein-protein interactions. This methodology was developed through a classification technique known as support vector machine (SVM) which has been already demonstrated its reliability and generalisation capability [46, 88], including bioinformatics approaches [238, 129, 16]. In model organism selected for this work was yeast (S. cerevisiae) because is the most widely analysed organism thus far, though despite this its interactome is still far from complete [26, 185]. However, with the purpose to face a specific problem, in the last part of this thesis, an approach is proposed in a effort to characterize complex molecular associations. Therefore, an analysis about E3-machinery – protein substrate relationship in the ubiquitin-proteasome system (UPS) for human (H. sapiens) was developed. UPS identifies and degrade the marked proteins to be eliminated (because they are misfolded, degraded or simply there were marked for regulation). Consequently UPS plays a fundamental role in the case of neurodegenerative diseases.

In the case of yeast, the methodology proposed in this thesis presents several different approaches along the chapters showing the evolution of work developed during 4 years. Hence, the common ideas are described in the following lines. This methodology is based in a initial step of feature extraction to considered datasets (training and testing) from well-known databases and datasets in proteomics, providing a number of features (25 or 26 according on the chapter) enough to define protein-protein interactions problem. However, this set of features may include irrelevant or redundant features. In this way, by the means of a pattern recognition process called feature selection, these irrelevant or redundant features are filtered out and greatly improves the learning performance of classifiers [23] [146]. Once this feature selection is applied to training set, SVM predictors are constructed using the selected set of most relevant features. The training set composed by a high-quality GSP set and GSN set. In order to verify the prediction power of the proposed SVM approaches, a comparative analysis of SVM models were carried out under different dataset or models depending on the chapter. In summary, the specific goals purposed in this thesis are:

1. To implement a new approach to PPI dataset processing based on the extraction of genomic and proteomic information from well-known databases and the application

of data mining techniques which provide very accurate models with high levels of sensitivity and specificity in the classification of PPIs.

2. To design a reliable features extraction supplying a significant number of features to define protein-protein interactions problem.

3. The predictor model has to be validated using different nature datasets, i.e, experimental, computational and literature-collected datasets.

4. To define an confidence score for each predicted pair of proteins. This score may help to validate PPI, thanks to slight modification of support vector machine model implementation. A more universal confidence scoring approach is more preferable because it is free of biases due to the influence of particular experimental setup context [30].

5. To present reliable feature selection methods in order to filter out the irrelevant or redundant features and so obtaining a reliable and less complex model in PPI prediction. This methods can be designed using filter-wrapper approach or an "ensemble" approach combining several methods.

6. To implement a parallel feature selection method motivated for the computational requirements such as memory and calculation and so saving time to receive results. This is specially useful in the case of filter-wrapper approach using SVM models and huge training and testing sets.

7. Incorporating similarity measures for each pair of proteins using information from known datasets in order to yield more information and more discriminative features about the considered interaction and so improving the accuracy of the final classifier.

8. To propose a reliable method for selecting negative sets, such a clustering approach to select most representative negative pairs instead of selecting randomly as traditionally. This approach is useful using a huge high-quality negative set as the set of more 4 million of negative samples was provided by Saeed and Deane [185].

9. To provide an Bioinformatics approach that aims to characterize complex molecular associations such as the E3-machinery – protein substrate relationship, with the purpose of sorting out potential ubiquitylation substrates and helping to define the structural constituent of the E3-substrate complexes.

1.3 Structure of the present work

Subsequently these sections of introduction and objectives, the structure of this thesis is given as follows:

- **Chapter 1. State of the art: Fundamentals of machine learning and pattern recognition methods in bioinformatics**. In this chapter is described the machine learning and pattern recognition methods utilised in the development of this work. A clustering description and k-means algorithm are explained. Feature selection section is included deals the filter and wrapper approaches, the concept of mutual information and the proposed algorithms based on the methods: Simba, Relief, G-flip and the minimal-redundancy-maximal-relevance criterion (mRMR) criterion method. A summary of the paradigm support vector machines with a description of cross validation and grid search are drawn in this chapter.

- **Chapter 2. Métodos de Predicción de Interacción Proteína-Proteína**. This chapter was written in Spanish. This chapter summarises and explain classic methods for protein-protein interactions such associative method (AM), maximum-likelihood estimation (MLE). Classic measures used in the prediction analysis are detailed. A description of selected works such PPI prediction method using domain-domain interactions by Huang et al [96] or filtering high-throughput PPI data using a combination of genomic features by Patil and al. [162] are dealt.

- **Chapter 3. Bases de datos en Proteómica: Interacciones proteína-proteína**. This chapter was written in Spanish. In this chapter, the most common databases in proteomics are described, specially for the case of protein-protein interactions and utilised in this research work. Explanations of data formats, identifiers, file headers associated to these databases are also included. Databases as Gene Ontology, Pfam, 3did, PDB, MIPS, UniProt and others are explained.

- **Chapter 4. Method for Prediction of Protein-Protein Interactions in Yeast using Genomics/Proteomics Information and Feature Selection**. In this chapter is presented a method for prediction of protein-protein interactions through a support vector machines. The dataset considered is a set of positive and randomly selected negative examples were extracted from Saeed et al. [185]. A description of extracted 26 proteomics/genomics features using well-known databases and datasets is drawn. Feature selection method is a emsemble method based on combining the G-Flip, Simba and Relief algorithms. Using the described way to extract and calculate the features, a comparison analysis between selected features and features based on the proposal by Patil et al. [162] is developed.

- **Chapter 5. Chapter Using Machine Learning Techniques and Genomic/Proteomic Information from Known Databases for Defining Relevant Features for PPI Classification**. In this chapter, and based upon the previous chapter, a approach is proposed to PPI data classification based on the extraction of genomic and proteomic information from well-known databases and the incorporation of semantic measures. The same 26 features are extracted. However,

by applying a filter-wrapper algorithm for feature selection, obtaining a final set composed of the 8 most relevant features for predicting PPIs. An analysis with the whole set of features and this sub-optimum set trained models was carried out, and which was validated by a ROC analysis. A filtering process between training and testing sets are detailed. This chapter concludes with a comparative analysis in order to evaluate the prediction capability of the support vector machine model using these 8 features testing set of external experimental, computational and literature-collected datasets extracted from Yu et al. [234]. This final model was trained using a set of "balanced" negative samples using the approach proposed by Yu et al. [233].

- **Chapter 6. Selecting negative samples for PPI prediction using hierarchical clustering methodology** In this chapter, a new approach is proposed to construct a PPI predictor training a support vector machine model through a mutual information filter-wrapper parallel feature selection algorithm and a iterative and hierarchical k-means clustering to select a relevance negative training set. The parallel was implemented using the library MPI-MEX. A comparative analysis between SVM models trained using the set of negative samples selected by the proposed hierarchical clustering, a set of randomly selected negative samples and a "balanced" negative set built using the approach proposed by Yu et al. [233].

- **Chapter 7. Characterization of the protein interaction neighbourhood of E3 enzymes in the Ubiquitin-Proteasome System**. In this chapter is outlined a bioinformatics approach that aims to characterise complex molecular associations such as the E3-machinery – protein substrate relationship. Studying the network neighbourhood between an E3 and a potential substrate, a connectivity profile of each E3 domain family and the domains of their interacting partners is drawn in order to sort out potential ubiquitylation substrates and help to define the structural constituents of the E3-substrate complexes. An E3-family connectivity profile analysis was applied with the purpose of finding interacting partners between families. Extending this approach through a biologically enrichment process and clustergram analysis for detecting similar patterns and functions for those E3 domain families grouped under the obtained connectivity groups.

- **Chapter 8. Summary, Conclusions and Contributions** is the last chapter and it offers a summary of the conclusions brought in this thesis along with the main contributions made and the publications that have been produced. It also includes the future work that opens from this thesis.

- **Bibliography**. This section gathers a updated list of references of works, books, urls and other additional resources used in this thesis.

- **Apéndice 1. Introducción Biológica: De la célula al ADN**. This appendix was written in Spanish. This appendix summarises a brief introduction to biology

from cell to DNA making a description of cell structure for eukaryote, prokaryote and archaea organisms.

- **Apéndice 2. Biología Molecular y Bioquímica**. This appendix was written in Spanish. This appendix gives a minimum description of biology paradigm, description of nucleic acids, DNA, RNA, transcription and translation.

- **Apéndice 3. Proteínas**. This appendix was written in Spanish. Properties and structure (primary, secondary, tertiary and quaternary) of proteins are explained in this appendix.

- **Apéndice 4. Proteómica**. This appendix was written in Spanish. Mainly, this appendix deals with experimental methods to detect protein-protein interactions such as yeast-two hybrid assays or. Although Blotting methods are used to detect biomolecules, they were also described briefly because of complements PPI methods.

- **Apéndice 5. Enfermedades Neurodegenarativas**. This appendix was written in Spanish. This appendix is also devoted to explain the relation between

In order to facilitate the reader to navigate through this thesis, some lists that reference and gather useful different information were included. They are the following ones:

- **List of Abbreviations**, available after this chapter, where the most frequent abbreviations used in this thesis with a description can be found.

- **Glossary**. This section refers to a glossary that contains the new or difficult terms that are distributed along this dissertation. It is written in Spanish.

- **List of figures**. A List of Figure which compiles all the figures that are displayed in this thesis.

- **List of tables** is found after the List of Figures. This list give a easy way to find any table written in the thesis.

LIST OF ABBREVIATIONS (ACRÓNIMOS)

- CV: Cross-validation (see 2.2.2)
- SVM: Support Vector Machine (see 2.2.1)
- RBF: Radial Basis Function (see 2.2.1)
- LOO: Leave-one-out cross-validation (see 2.2.2)
- UPS: Ubiquitin-Proteasome System (see 8.2)
- TAP: Tandem affinity purification (D)
- Y2H: Yeast to hybrid (doble híbrido de la levadura ver D)
- CO-IP: Co-Immunoprecipitation Assays (Co-inmunoprecipitación D)
- MS: Mass-spectrometry (espectrometría de masas D)
- GSP: Gold Standard Positive (see 1)
- GSN: Gold Standard Negative (see 1)
- ROC: Receiver operating characteristic (curve) (curva ROC 3.2.2)
- HT: High-throughput
- DDI: Domain-Domain Interactions
- PPI: Protein-Protein interactions

2

STATE OF THE ART: FUNDAMENTALS OF MACHINE LEARNING AND PATTERN RECOGNITION METHODS IN BIOINFORMATICS

2.1 Introduction

This chapter collects the fundamentals of machine learning and pattern recognition methods used along this thesis. In the case of machine learning, support vector machine (SVM) for classification is used and a description of cross-validation and grid-search methods is also included in this chapter. Those methods are used to optimize the parameters of SVM model. A briefly description of clustering and the classical k-means are also explained, its usage is motivated for huge amount of protein data processed in this thesis. Finally, this chapter concludes with a explanation about feature selection and the used algorithms. The algorithms Simba, Relief and G-flip are based on margin criterion instead of mRMR method (minimal-redundancy-maximal-relevance criterion) based on mutual information. In this thesis, the PPI prediction problem is considered as a classification problem, so each sample point represents a pair of proteins which must be classified into one out of two possible classes: non-interacting or interacting pair.

2.2 Machine Learning

2.2.1 Support vector machine

Support vector machines (SVM) are a classification and regression paradigm first developed by Vladimir Vapnik [46, 88]. The SVM approach is quite popular in the literature related to classification and regression problems because of its good generalisation capability and its superiority in comparison with other machine learning

paradigms [200]. SVMs were originally designed for binary-class classification; hence it is straightforward to use this paradigm in the present problem for classification between interacting and non-interacting pairs of proteins.

Given a training set of instance-label pairs $\{\mathbf{x}_i, y_i\}$, $i = 1, ..., N$ with input data $\mathbf{x}_i \in \mathbb{R}^n$ and labelled output data $y_i \in \{-1, +1\}$, a support vector machine [46] solves the following optimization problem:

$$\begin{array}{rc} min_{\mathbf{w},b,\epsilon} & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i \\ subject\ to & y_i(\mathbf{w}^T\phi(x_i) + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \end{array} \tag{2.1}$$

The training data vectors $\mathbf{x}_i$ are mapped into a higher dimensional space by means of the ϕ function. The hyperparameter C is the penalty parameter of the error term, or in other words, it is a real positive constant that controls the amount of misclassification allowed in the model.

From the primal problem given in equation 2.1, the dual form of a SVM can be obtained:

$$\begin{array}{rc} min_{\alpha} & \frac{1}{2}\alpha^T Q\alpha - \mathbf{e}^T\alpha \\ subject\ to & y^T\alpha = 0 \\ 0 \leq \alpha_i \leq C, & i = 1, ..., N \end{array} \tag{2.2}$$

where $\mathbf{e}$ is the all-ones vector. Q is an N by N positive semidefinite matrix given by $Q_{ij} \equiv y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$. $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ is called the *kernel function*, and allows the algorithm to fit a maximum-margin hyperplane in a transformed feature space; i.e., the classifier is a hyperplane in the high-dimensional feature space that may be non-linear in the original input space.

Therefore, for the general nonlinear SVM classifier, the decision function can be written as:

$$y(x) = sign[\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b] \tag{2.3}$$

The parameters α_i correspond to the solution of the Quadratic Programming problem that solves the maximum margin optimization problem. The training data points corresponding to non-zero α_i values are called *support vectors* [204] because they are the ones that are really required to define the separating hyperplane.

The two most typical kernels in the bioinformatics literature are the Linear kernel and the Radial Basis Function (RBF) kernel [178]. The linear kernel is represented by:

$$K(x, x_i) = x_i^T x \tag{2.4}$$

that corresponds to the linear SVM. The Radial Basis Function (RBF) kernel is represented by:

$$K(x, x_i) = exp(-\gamma||x - x_i||^2), \gamma > 0 \tag{2.5}$$

where parameter γ controls the region of influence of every support vector.

Training an SVM implies the optimization of the α_i and of the so-called hyper-parameters of the model, normally estimated using grid-search and cross-validation [204] as described on next sections. More specifically, for the RBF Kernel the hyperparameters C and γ need be optimized.

2.2.2 Parameter selection: Cross-validation and Grid-search

In RBF kernel there two parameters C and γ which are not known are best for a given problem. Therefore a parameter search method is required in order to select a good combination of C and γ make the classifier reliable and accuracy. It is possible that in certain cases may not be useful to achieve high training accuracy because really the considered classifier is predicting training data whose class labels are already known. A common strategy is to divide the data set into two parts, taking one of these part as unknown. Then prediction accuracy for this unknown set is indicating the performance on classifying an independent data set. However, this version is improved in a procedure known as cross-validation [40].

The cross-validation (CV) is a family of methods can be used to estimate the error of a given meta-model. CV can also be used to select a meta-model taking the model which attains lowest cross-validation error, or choosing the most relevant subset of features. Such error measure estimated for cross validation allows assessing the reliability of the meta-model.

In the basic approach of CV, a starting data set S is considered as $S\{X, Y\}$ with N input-output pairs (x_i, y_i) for $i = 1 \dots N$ where y_i is the output for the input x_i. Hence, N is the number of total samples in the model. In the fist step, the data set is split in two parts and data positions are randomly changed. For example, the first part denoted as $S^1\{X^1, Y^1\}$ with size n^1 is used to fit the model, and the second part $S^2\{X^2, Y^2\}$ is used to calculated the cross validation error. This is the difference between predictions of the model $\hat{y}^2$ and the values of y^2 in the points where x^2 were omitted. The step of cross-validation consists of commutate the data set of the model, fit and predict to obtain an additional estimation of how good is the model predicting a new data set.

In the p-fold cross-validation (p-fold CV), the initial data set is divided in p mutually exclusive and exhaustive subsets, i.e. , $S\{X, Y\} = S^1\{X^1, Y^1\}, S^2\{X^2, Y^2\}, \dots,$ $S^p\{X^p, Y^p\}$. Then, the model is fitted p times, in a time one of the training subsets is leaved out and then it is used to calculate the CV error. In other words, sequentially one subset is tested using the classifier trained on the remaining p subsets. In this way,

the cross-validation procedure can prevent the overfitting problem [40, 140]. There is an extreme implementation of p-fold CV known leave-one-out cross-validation (LOOCV or LOO) which is the same as a p-fold cross-validation but with p being equal to the number of observations (pairs or points) in the original sample (considered set).

However, a common task is to utilise "grid-search" on C and γ using cross-validation, recommended by Chang et al. [40]. This procedure is based on the idea of taking various pairs of C and γ values to be tried in the classifier measuring the cross-validation accuracy, and then one of the pairs C and γ with the best cross-validation accuracy is picked. Chang et al. found that trying exponentially growing sequences of C and γ is a practical procedure to select good parameters. An example of "grid-search" is considering the combination (C and γ) for $C \in \{2^{-5}, 2^{-3}, \ldots, 2^{13}, 2^{15}\}$ and $\gamma \in \{2^{-15}, 2^{-13}, \ldots, 2^{1}, 2^{3}\}$.

Although there are several advanced methods in the literature that can save computational cost (e.g., approximating the cross-validation rate), Chang et al. [40] described two motivations: 1) an exhaustive parameter search is not the "same" as the results obtained by approximations or heuristics and 2) the computational time required to find two good parameters by grid-search is not much more than that by advanced methods. In addition, the grid-search procedure can be easily implemented in a parallel procedure because each combination of C and γ is independent [40].

2.2.3 Clustering

The goal of a data clustering approach is to organize patterns (data) into significative groups using some kind of similarity measure [136]. Consequently a precise definition of similarity is fundamental and can have greatly impact in the result of the clustering algorithm [11]. However, the clustering problem is not well defined except for the resulting classes of samples display certain properties. The selection of properties, i.e, the definition of a cluster, is the key point in the clustering problem. Through an appropriate definition of a cluster, the distinction between good and bad classifications is possible [68].

The clustering algorithms can be classified in supervised and unsupervised algorithms. In supervised clustering, the clustering algorithm are based on a set of given reference classes or vectors. In the case of unsupervised clustering, no predefined set of classes is utilised. Besides hybrid methods are also possible, but an unsupervised approach is followed by a supervised one. In bioinformatics literature, unsupervised methods such as k-means are the most commonly applied, for example in gene expression array experiments [11].

On the other hand, clustering approaches can be divided in parametric and non-parametric approaches. In the case of parametric approaches, normally clustering criteria are defined with which given samples are classified to a number of clusters optimising the criteria. The class separability measures are the most commonly used criteria, and the best considered result is the class assignment which maximises the class

separability. In another parametric approach, a mathematical form is considered to express the distribution of the data. Here, clustering problem is to find the parameter values for this distribution which best fit the data. In the case of non-parametric approach, neither mathematical forms nor criteria are used in the non-parametric approach. Hence, samples are classified according to the valley of the density function. The resulted boundary can be complex and not expressible by any parametric form [68].

2.2.3.1 Hierarchical clustering. Dendrogram

An hierarchical branching procedure can generate clusters. Therefore, there are automatic methods that building a tree using a data given in the form of pair wise similarities. The common and standard algorithm used detailed in Baldi et al [11] recursively computes a dendrogram that gathers all the elements into a tree. It started from the similarity matrix C and N starting points, then at each step of the algorithm is explained as Baldi et al. [11]:

- The two most similar points of the current matrix are computed and a node is created through joining these two elements.
- An expression profile (or vector) is created for the node by averaging the two expression profiles associated with the two points. There is an alternative way, a weighted average of the distances is utilised to estimate the new distance between centres without computing the profile.
- Other smaller correlation matrix is computed using the newly computed expression profile and replacing the two joined elements with the new node.
- This process is repeated N - 1 times or until a single node remains.

This dendrogram algorithm is quite familiar to biologists and experimentalists which has been applied in sequence analysis, phylogenetic trees, and average-linkage cluster analysis [11].

2.2.3.2 k-means

In data mining, k-means clustering [132] is a method of cluster analysis that assigns n observations into k clusters where each observation belongs to the cluster with the nearest mean. Given a set of observations (x_1, x_2,..., x_n), where each observation or point is a d-dimensional real vector. Then n observations are assigned into k sets ($k \leq n$) $S = S_1$, S_2, ..., S_k minimising the within-cluster sum of squares (WCSS) [132]:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 \tag{2.6}$$

where μ_i is the mean of points in S_i.

Therefore, giving a little more detailed "non-mathematically" explanation of the algorithm in the following lines. The typical implementation of the k-means algorithm, the number of clusters K is fixed. Then K representative points (called centroids) are initially selected for each cluster more or less arbitrarily [11]. Then at each step of k-means as Baldi et al.[11] detailed would be:

- Each point in the data is assigned to the cluster which is associated with the nearest representative.
- New representative points are computed after the assignment to the cluster, (e.g. by averaging or taking the center of gravity of each computed cluster).
- The two procedures of previous steps are repeated until fluctuations remain small or the algorithm converges.

Considering all clustering algorithms, k-means has probably the most exemplary probabilistic interpretation as a form of EM (expectation maximization) on the underlying mixture model [11].

2.3 Feature Selection

In pattern recognition theory, patterns are represented by a set of variables (features) or measures. Such pattern is a point in a n-dimensional features space. The main goal is to select features that distinguish uniquely between patterns of different classes. Normally, the optimal set of features is unknown and has commonly a number irrelevant or redundant features. In this way, through a pattern recognition process, these irrelevant or redundant features are filtered out improving greatly the learning performance of classifiers [23][146]. This reduction in the number of features, also known as *feature selection*, allows for a better training efficiency as the search space for most of the parameters of the model is also reduced [23, 146]. Likewise, the less the number of features, less complex model, the easier to understand the final model will be and better visualisation of used data [62].

In many supervised learning tasks the input data are represented by a very large number of features, however only a few of them are relevant for predicting the class (or label). Even the current classification algorithms as SVM are affected by the presence of a large numbers of redundant and weakly relevant features. Normally this problem is associated to "the curse of dimensionality" [15]. Besides, when the dimension is high makes many algorithms computationally intractable [175]. Nonetheless, using a reduced group of good selected features, even the most basic classifiers can attain high performance levels. Thus, feature selection to find a subset of good features is fundamental for efficient learning [175].

Two general approaches are commonly applied in the machine learning literature for feature selection: filter selection methods and wrapper selection methods [62][111]. In the filtering approach, feature selection is performed as a preprocessing step prior to the actual learning algorithm. This preprocessing step measures general characteristics of the training set to select the most important features and to discard the irrelevant ones. On the other hand, wrapper methods use the machine learning classification model itself in order to extract the relevant features, i.e, use the classifier performance to assess the goodness of features subset. This entails a much more expensive computational cost, since for every subset of features selected by the method, the whole classifier has to be optimized and evaluated. Other authors have utilised a combination of filter and wrapper algorithms [166], in fact in this work a combination between filter and wrapper is used. First, a filter methods is applied in order to obtain the relevance of features and subsequently a wrapper method is performed using support vector machine models from the obtained relevance order.

Different criteria have been applied to evaluate the goodness of a feature [62][22]. In this section is presented two kind of criteria: a margin based feature selection criterion and minimal-redundancy-maximal-relevance criterion.

2.3.1 Margin based feature selection criterion methods

In this section, the margin based feature selection methods (Relief, Simba, G-flip) were extracted from works by Ran Gilad-Barchrach, Amir Navot and Naftali Tishby [175, 72] and their tutorial on Internet [71]. The novelty of their work is the usage of large margin principles for feature selection. The authors introduced the idea of measuring the quality of a set of features by the margin it induces. A margin is a geometric measure for evaluating the confidence of a classifier with respect to its decision [175, 72].

Crammer et al. [49] describe two natural ways of defining the margin of an instance with respect to a classification rule: 1) sample-margin (the more common type) measures the distance between the instance and the decision boundary induced by the classifier and 2) the hypothesis-margin that requires the existence of a distance measure on the hypothesis class. The margin of a hypothesis with respect to an instance is the distance between the hypothesis and the nearest hypothesis that assigns an alternative class (label) to the given instance [175, 72].

Ran Gilad-Barchrach et al. worked in margins for the traditional classifier 1-Nearest Neighbour (1-N N[65]), which classifies each point to the class of its nearest neighbour. The 1-NN classifier is defined by a set of training points and the decision boundary is the Voronoi tessellation [7]. Crammer et al. [49] showed that the sample margin for 1-NN can be unstable, then Ran Gilad-Barchrach et al. used hypothesis margin because: 1) the hypothesis-margin lower bounds the sample-margin and 2) the hypothesis-margin is easy to compute [175, 72].

Therefore, in order to describe the used methods of Simba, Relief and G-flip

proposed by Ran Gilad-Bachrach et al. [72], it is required a formal definition of considered hypothesis margin and an evaluation function. This function assigns a score to sets of features according to the margin they induce, considering that a good generalization can be guaranteed if many sample points have a large margin. Nearest Neighbour using a large hypothesis-margin ensures a large sample-margin [175, 72]. Thus, let P be a set of points, x be an instance and w be a weight vector over the feature set, the hypothesis margin (or simply margin) of x is:

$$\theta_P^{\mathbf{w}}(\mathbf{x}) = \frac{1}{2}\left(\|\mathbf{x} - \mathbf{nearmiss}(\mathbf{x})\|_{\mathbf{w}} - \|\mathbf{x} - \mathbf{nearhit}(\mathbf{x})\|_{\mathbf{w}}\right) \tag{2.7}$$

where $\|\mathbf{z}\|_{\mathbf{w}} = \sqrt{\sum_i w_i^2 z_i^2}$, $\mathbf{nearhit}(\mathbf{x})$ and $\mathbf{nearmiss}(\mathbf{x})$ denote the closest point to $\mathbf{x}$ in P with the same and different label (class), respectively. The set of selected features affects the margin through the distance measure [72].

Let $u(\cdot)$ be a utility function. The evaluation function for given a training set S and a weight vector w would be:

$$e(\mathbf{w}) = \sum_{\mathbf{x} \in S} u\left(\theta_{S \setminus \mathbf{x}}^{\mathbf{w}}(\mathbf{x})\right) \tag{2.8}$$

The building blocks of this function are the margins of all the sample points. The margin of each instance $\mathbf{x}$ is calculated with respect to the sample excluding $\mathbf{x}$ (i.e, a "leave-one-out margin") [175, 72].

The utility function manages the contribution of each margin term to the overall score. Gilad-Bachrach et al. [175, 72] considered three utility functions as follows:

1. Linear utility function. This function is defined as $u(\theta) = \theta$. When the linear utility function is applied, the evaluation function is merely the sum of the margins.

2. Zero-one utility function. This function is equals 1 when the margin is positive and 0 in another case. When it is used the utility function is proportional to the LOO (leave-one-out) error.

3. Sigmoid utility function. This function is $u(\theta) = 1/(1 + exp(-\beta\theta))$. The sigmoid utility function is less sensitive to outliers than the linear utility, but does not ignore the magnitude of the margin completely as the zero-one utility does [175].

It is important to take into account that for $\beta \to 0$ or $\beta \to \infty$ the sigmoid utility function becomes the linear utility function or the zero-one utility function respectively [175, 72]. Therefore, in the next subsections, the G-flip (Greedy feature flip algorithm), Simba (Iterative Search Margin Based Algorithm) and Relief methods are explained as Gilad-Bachrach et al. proposed [175, 72].

2.3.1.1 Greedy Feature Flip Algorithm

The Greedy feature flip algorithm (G-flip) is a greedy search algorithm for maximizing directly $e(F)$ considering F is a set of features. G-flip is a repetitive algorithm that iterate over the feature set and updates the set of selected features. Thus in each iteration the current feature is decided to remove or add to the selected set by evaluating the margin function 2.8 with and without this feature. The algorithm converges to a local maximum from of the utility function because of in each iteration increases its value and the number of possible feature sets is finite. The computational complexity of G-flip can be expressed as $\Theta(N^2m^2)$ where N is the number of features and m is the number of instances, considering one pass over all features. Nevertheless, e complexity can be reduced to $\Theta(Nm^2)$ if the updating of the distance matrix is implemented efficiently after each addition or deletion of a feature [175, 72]. The algorithm is described in pseudo-code 2.1.

```
Initialize the set of chosen features to the empty set. F = ∅
for t = 1,2,... do
  select a random permutation s of {1...N}
  for i = 1...N do
    evaluate e1 = e(F ∪ {s(i)})) and e1 = e(F \ {s(i)}))
    if (e1 > e2), F = F ∪ {s(i)}
    elseif (e2 > e1), F \ {s(i)}
  end for
  if no change made in step of evaluation then break
end for
N is the number of features.
```

Algoritmo 2.1: Pseudo-code for Greedy Feature Flip Algorithm (G-flip)

As it can be observer, G-flip is a parameter free algorithm, i.e, there is no necessity of tuning then features or any threshold [71].

2.3.1.2 Iterative Search Margin Based Algorithm

The Iterative search margin based algorithm (Simba) first find the weight vector $\mathbf{w}$ that maximizes $e(\mathbf{w})$ (see equation 2.8) and then utilise a threshold in order to get a feature set. Since $e(\mathbf{w})$ is smooth almost everywhere, Gilad-Bachrach et al. [175, 71] use gradient ascent with the purpose of maximizing it. Then the gradient of $e(\mathbf{w})$ when evaluated on a sample S would be [175, 72, 71]:

$$
\begin{aligned}
(\bigtriangledown e(\mathbf{w}))_i &= \frac{\partial e(\mathbf{w})}{\partial w_i} = \sum_{\mathbf{x}\in S} \frac{\partial u(\theta(\mathbf{x}))}{\partial \theta(\mathbf{x})} \frac{\partial \theta(\mathbf{x})}{\partial \mathbf{w}_i} \\
&= \frac{1}{2} \sum_{\mathbf{x}\in S} \frac{\partial u(\theta(\mathbf{x}))}{\partial \theta(\mathbf{x})} \Big(\frac{(x_i - \mathbf{nearmiss}(\mathbf{x})_i)^2}{\|\mathbf{x} - \mathbf{nearmiss}(\mathbf{x})\|_{\mathbf{w}}} - \frac{(x_i - \mathbf{nearhit}(\mathbf{x})_i)^2}{\|\mathbf{x} - \mathbf{nearhit}(\mathbf{x})\|_{\mathbf{w}}} \Big) w_i
\end{aligned} \tag{2.9}
$$

In this algorithm, a stochastic gradient ascent over $e(\mathbf{w})$ while ignoring the constraint $\|\mathbf{w}^2\|_\infty = 1$ is applied. Then, in each step one term in the sum in equation 2.9 is evaluated and it is added to the weight vector $\mathbf{w}$. The projection on the constraint is done only at the end (step 3) (see algorithm 2.2).

The computational complexity of Simba can be expressed as $\Theta(TNm)$ where N is the number of features, T is the number of iterations and m is the size of the sample S. However, when $T = m$ (i.e. iterating over all training instances), the complexity can be expressed as $\Theta\left(Nm^2\right)$ [175, 72, 71].

Initialize $\mathbf{w} = (1, 1, 1, \ldots, 1)$
for $t = 1 \ldots T$ **do**
  select randomly an instance x from S
  calculate **nearmiss**$(\mathbf{x})$ and **nearhit**$(\mathbf{x})$ with respect to $S \setminus \{\mathbf{x}\}$ and the weight vector $\mathbf{w}$
  for $i = 1 \ldots N$ **do**
    calculate $\triangle_i = \frac{1}{2} \frac{\partial u(\theta(\mathbf{x}))}{\partial \theta(\mathbf{x})} \left(\frac{(x_i - \mathbf{nearmiss}(\mathbf{x})_i)^2}{\|\mathbf{x} - \mathbf{nearmiss}(\mathbf{x})\|_{\mathbf{w}}} - \frac{(x_i - \mathbf{nearhit}(\mathbf{x})_i)^2}{\|\mathbf{x} - \mathbf{nearhit}(\mathbf{x})\|_{\mathbf{w}}} \right) w_i$
  end for
  $\mathbf{w} = \mathbf{w} + \triangle$
end for
if no change made in step of evaluation then break
N is the number of features.

Algoritmo 2.2: Pseudo-code for Iterative Search Margin Based Algorithm (Simba)

2.3.1.3 Relief

For given a dataset S, Relief returns a ranking of features according to an importance weight vector w, and these weights allow us to determine which attributes are relevant and to set an order among them. The pseudo-code of the Relief algorithm as Gilad-Bachrach et al. implemented [175, 72, 71] is shown in algorithm 2.3, where *nearmiss*(x) and *nearhit*(x) denote the nearest point to x in the set of I/O sample data P with the same and different label respectively.

Relief is a well-known, simple and efficient method that has already been successfully used in proteomics [23].

0: $\mathbf{w} = 0$ {Initiate the weight vector to zero}
1: **for** $t = 1...T$ **do**
2: Select randomly an instance $\mathbf{x}$ from S
3: **for** $i = 1...N$ **do**
4: $w_i = w_i + (x_i -$ **nearmiss**$(\mathbf{x})_i)^2$ - **nearhit**$(\mathbf{x})_i)^2$
5: **end for**
6: **end for**
7: The selected feature set is $\{i|w_i > \tau\}$ where τ is a threshold

Note: Relief does not re-evaluate the distances according to the weight vector $\mathbf{w}$. P is a set of points. *nearhit* denotes the nearest point to $\mathbf{x}$ in P with the same label. *nearmiss* denotes the nearest point to $\mathbf{x}$ in P with different label. T is the number of iteractions. N is the number of features.

Algoritmo 2.3: Pseudo-code for Relief Algorithm

2.3.2 Mutual information and minimal-redundancy-maximal-relevance criterion

In this section, the presented filter features selection method is based on mutual information as relevance measure and redundancy between the features through minimal-redundancy-maximal-relevance criterion (mRMR) proposed by Peng et al. [166]. The mutual information of given two random variables is a value that measures the mutual dependence of the two considered variables, and it is described more formally in the next lines.

Let X and Y two random continuous variables with marginal pdfs $p(x)$ y $p(y)$ respectively, and joint probability density function (pdf) $p(x, y)$. The mutual information between X and Y can be represented as [62] [47].

$$I(X,Y) = \iint p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy \tag{2.10}$$

In the case of discrete variables, the integral operation is reduced to a summation operation. Let X and Y two discrete variables with mathematical alphabets $\mathcal{X}$ and $\mathcal{Y}$, marginal probabilities $p(x)$ and $p(y)$ respectively and a joint probability mass function $p(x, y)$. The MI between X and Y is expressed as [62]:

$$I(X,Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{2.11}$$

The mutual information (MI) has two principal properties make it different from other dependency measures: 1) the capacity of measure any relationship between variables and 2) its invariance under space transformations [62] [119].

Now, in order to define mRMR method, the criteria of Max-Dependency (maximal dependency), Max-Relevance (maximal relevance) and Min-Redundancy (minimal

redundancy) are defined. In the context of mutual information, the objective of feature selection is to find a feature set S with m features $\{x_i\}$, which jointly have the largest dependency on the target class c [166]. This scheme is called Max-Dependency (maximal dependency) and it can be defined as [166]:

$$\max D(S, c), D = I(x_i, i = 1, \ldots, m; c). \tag{2.12}$$

However, Max-Dependency criterion is quite difficult to implement, an alternative is to select features based on Max-relevance (maximal relevance criterion). The maximal relevance criterion is to search features fulfilling the equation 2.13, which approximates $D(S, c)$ in equation of Max-Dependency 2.12 with the mean value of all mutual information values between individual feature x_i and class c [166]. The Max-Relevance can be expressed as [166]:

$$\max D(S, c), D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c). \tag{2.13}$$

Therefore, it is likely that selected features using Max-Relevance can have a large dependency among them. In addition, when two features are highly depended on each other, the respective class-discriminative power would not change much assuming that one of them were eliminated. Thus, the Min-Redundancy (minimal redundancy) criteria can be appended to select mutually exclusive features [166]. The Min-Redundancy can be defined as [166]:

$$\min R(S), R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j). \tag{2.14}$$

The criterion resulted of combining the Min-Redundancy and Max-Relevance is called "minimal-redundancy-maximal-relevance" (mRMR). For mRMR, authors considered mutual information based feature selection for both discrete and continuous data [166]. The MI for continuous variables was estimated using Panzer Gaussian windows [166]. Peng et al. show that using a first order incremental search (as a feature is selected in a time) the mRMR criterion is equal to maximum dependence, or in other words, estimating the mutual information $I(C, S)$ between class variable C and subset of selected features S. In Peng el al. [166], for minimizing the classification error in the incremental search algorithm, mRMR method is combined with two wrapper schemes. In a first stage, the method is used with the purpose to find the candidate feature set. In a second stage, backward and forward selections were applied in order to find the compact feature set through the candidate feature set that minimises the classification error.

Given class variable C, the initial set of features F, an individual feature $f_i \in F$ and a subset of selected features $S \subset F$, the mRMR criterion for the first order incremental

search can be expressed as the optimisation of the following condition [62] [166] :

$$I(C;f) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{2.15}$$

where $|S|$ is the cardinality of the selected feature set S, $f_s \in S$.

This filter mRMR method is a fast and efficient method because its incremental nature, showing better feature selection and accuracy in classifier including wrapper approach [166] [62].

2.4 Conclusions

In this chapter, fundamentals of machine learning and pattern recognition methods were explained and used in bioinformatics problems. The algorithms described in this chapter were used during the course of this thesis for the research works summarises in the different chapters.

3

MÉTODOS DE PREDICCIÓN DE INTERACCIÓN PROTEÍNA-PROTEÍNA

3.1 Introducción

Los últimos avances en las metodologías experimentales han permitido obtener una gran cantidad de datos genómicos. Lo cual ha propiciado la iniciativa en la biología actual de la obtención de modelos funcionales de los productos de genes. Todo ello viene suscitado debido a que la mayoría de los procesos biológicos están regulados por interacciones proteína-proteína [80]. Dichas proteínas adquieren su función específica al interaccionar con otras proteínas [240]. Es por ello que en los últimos años se ha invertido un gran esfuerzo en la identificación de las interacciones proteína-proteína con el propósito final de conocer completamente el funcionamiento de los mecanismos celulares permitiendo así el diseño de nuevos enfoques terapéuticos más efectivos [202, 240].

Hasta ahora en la literatura científica sobre estudios de interacción proteína-proteína, podemos encontrar dos tipos de enfoques principales: análisis experimentales [164, 30] y enfoques computacionales [185].

Experimentalmente, hay fundamentalmente dos ramas complementarias de análisis de interacción de proteínas [30]: 1. Análisis de complejos de proteínas usando afinidad de purificación seguido de espectrometría de masas (AP/MS), que identifica interacciones directas e indirectas. 2. Análisis de doble-híbrido de la levadura de alta resolución (high-throughput yeast two-hybrid -Y2H-), que identifica interacciones directas o binarias. Si bien, existen otros métodos experimentales de identificación de interacciones directas que han suministrado pequeños datasets (conjunto de datos)[30].

Pese a los avances en los enfoques experimentales actuales ya desarrollados para diversos organismos, aún está en proceso la obtención de interactomas completos, esto es, la identificación sistemática de interacciones de proteínas dentro de un organismo.

Así, numerosas líneas de investigación actuales se ayudan de métodos computacionales o híbridos, que utilizan postulados teóricos y/o conocimiento experimental para la construcción de modelos estadísticos/inteligentes que ayuden a predecir las interacciones proteína-proteína [80], o que sirvan de guía en la experimentación. Entre estos métodos encontramos: métodos bayesianos [106, 162, 116], Maximum Likelihood Estimation (MLE) [54, 100], Maximum Specificity Set Cover (MSSC) [96], decision trees [238, 129] o Support Vector Machines (SVM) [238, 16, 217, 48, 129, 237], que han sido utilizados para predecir interacciones en organismos diversos, desde el *Saccharomyces cerevisiae* (S. cerevisiae o simplemente levadura de la cerveza) [212, 101, 70, 92], *Caenorhabditis elegans* (C. elegans) [221, 124], *Drosophila melanogaster* (D. melanogaster) [73, 67] hasta el *Homo Sapiens* (H. sapiens) [27, 181, 202] y otros [185]. Entre ellos, la levadura de la cerveza (yeast, baker's yeast) ha sido el organismo más estudiado, aunque hay que resaltar que su interactoma está aún por completar [26, 185]

Existen dos principales inconvenientes en el estudio de la predicción de interacciones proteína-proteína. En primer lugar, pese a que se pueden obtener una gran cantidad de datos sobre interacción proteína-proteína mediante métodos experimentales Large Scale ó High-Throughput (de gran escala o de alto rendimiento), esta información es difícil de contrastar, existiendo numerosos falsos positivos [163, 185]. En todo caso, en algunos trabajos de investigación sobre interactoma se proporcionan diferentes conjuntos de interacciones con distintos niveles de fiabilidad, proporcionándose un conjunto high-confidence (alta confianza) ó conjunto CORE. Por otro lado, en los estudios sobre interacción, rara vez se reportan los non-interacting pairs (los pares que no interaccionan), debido a la dificultad de demostrar la no interacción en todas las condiciones posibles. Todo esto realza la importancia de los métodos computacionales como método adicional en el filtrado de PPIs y como guía en la experimentación de PPIs [163, 96]. A la vez, sin embargo, estos inconvenientes dificultan enormemente la construcción y utilización de metodologías supervisadas de aprendizaje para la predicción de interacciones, que necesitan de conjuntos de datos completos, significativos y fiables.Los investigadores de la estructura de proteínas sugieren que la unidad fundamental es un dominio.

En los últimos años, han surgido en este campo diversos enfoques que tratan de resolver estos inconvenientes. Una línea importante de trabajo en este sentido se basa en la explotación de diversos aspectos de la información proteómica y genómica ya conocida sobre el sistema analizado. Existen varios enfoques como la observación de los dominios de proteínas que tienden a combinarse en una fusión de proteínas (Proteínas de fusión), o la observación de enlaces funcionales en las proteínas que tiende a preservarse o eliminarse en la evolución. Por tanto las proteínas que tienen coincidencias filogenéticas (ver glosario árbol filogenético) tienden a estar enlazadas funcionalmente con gran fuerza [165, 134, 96]. De esta manera, los estudios relativos a la estructura espacial de la proteína sugieren que la unidad fundamental en la interacción es el dominio proteico [96], siendo la arquitectura del dominio la principal

responsable de la funcionalidad biológica de la proteína. Comparando los organismos de los tres reinos, las eucariotas tienden a tener proteínas multidominio, mientras que las bacterias o arqueas suelen poseer proteínas de un único dominio [96]. Así pues, la información relativa al dominio ha propiciado un marco de trabajo ideal para la creación de modelos de predicción de PPIs [54, 96, 238, 129, 162]. Otros trabajos consideran patrones de localización celular como criterio plausible de interacción, ya que las proteínas que interaccionan suelen estar altamente co-expresadas y co-localizadas en el mismo componente subcelular [106]. En otros casos, se toman en consideración que las proteínas que interaccionan muestran altos niveles de similitud funcional, en contraste con la parejas de proteínas seleccionadas aleatoriamente [8]. De igual modo, la introducción del concepto de "interlogs" (parejas ortólogas de proteínas que interaccionan en varios organismos) [221]) revolucionó la forma de abordar el problema de las PPIs permitiendo generar modelos de predicción [52, 162], bases de datos de homólogos [163] o generación de datasets (conjunto de datos) negativos [185].

Aunque dichos enfoques han aportado predicciones que han mejorando la cobertura y precisión de las redes de PPIs[240], las diversas fuentes de información propuestas tienen escaso valor predictivo por sí solas. No obstante, la combinación (o integración) de varias fuentes de información en la construcción de un modelo han permitido realizar predicciones fiables [106, 163, 129, 240], especialmente en los métodos de aprendizaje supervisado. En este caso, de igual relevancia que la selección de fuentes de información o características de las interacciones para nuestro modelo, son los conjuntos de entrenamiento. Para una predicción fiable, dicho conjunto de entrenamiento, normalmente denominado Gold Stardard Set (GSS), debe de estar formado por ejemplos positivos (o comúnmente llamado subconjunto Gold Stardard Positive -GSP) así como por ejemplos negativos (o subconjunto Gold Standard Negative -GSN).

Como anteriormente se ha dicho, la creación de un conjunto de ejemplos negativos fiable plantea un problema difícil de abordar. Tradicionalmente, la obtención de conjuntos negativos se ha realizado emparejando aleatoriamente dos proteínas [238, 173, 232] o seleccionando parejas de proteínas que no compartían el mismo componente celular [106]. Sin embargo, otros estudios sugieren que los conjuntos de negativos creados solamente en la localización celular conducen a estimaciones sesgadas en la precisión de los modelos predictivos de interacciones [16]. Ante este inconveniente, Wu et al. [230] proponen un método de predicción de interacciones negativas mediante el uso de medidas de similitud semántica [222], basada en las ontologías de las anotaciones GO (gene ontology)[42]. No obstante, en dicho trabajo, los autores no determinaron cuál de las tres ontologías contribuía más a la determinación de las interacciones negativas. Todo ello motivó a Saeed y Deane [185] a proponer un método novedoso para generar conjuntos GSN, basándose en datos funcionales, localización, expresión y homología. Estos autores tuvieron en cuenta que una interacción entre dos proteínas debería sólo ser considerada como una interacción negativa, si las dos proteínas que interaccionan no poseen ningún solapamiento en ninguna de las características a

estudio. Por otra parte, en un trabajo muy reciente Yu et al. [233] se demuestra, que los datos de interacción obtenidos de experimentos basados en enfoques cebo-presa (por ejemplo Tandem affinity purification -TAP- y Y2H), las proteínas cebo suelen estar sobre-representadas en las interacciones identificadas. Se comprueba además que muchas proteínas se predisponen a participar en un gran número de interacciones formando "hubs" (como nudos). Así pues, las parejas seleccionadas aleatoriamente que no interaccionan tendrán características cristalográficas diferentes del conjunto positivo, haciéndolas distinguibles en cierta medida sin ninguna otra información. De ahí que Yu et al. planteen un método de selección de ejemplos negativos libre de sesgos usando un esquema de puntuación basado en la frecuencia de aparición de la proteínas en un dataset. Esto permite que el conjunto de negativos quede balanceado al conjunto de positivos, eliminando los posibles problemas planteados en la alternativa de seleccionarlo aleatoriamente.

Sin embargo, nótese que no podemos trabajar con datos 100% fiables, siempre tenemos una tasa de error o un sesgo, los datos tomados bien sean de bases de datos o de publicaciones pueden contener un porcentaje de falsos positivos [148], aunque mediante técnicas experimentales [153] como computacionales [185] en los últimos años se ha mejorado la calidad de los datos.

En este capítulo, se detalla a continuación varios métodos clásicos y de trabajos seleccionados, de los que se tomaron conceptos importantes en el desarrollo de esta tesis, cerrando así el carácter formativo de los capítulos del "estado del arte". Si bien el anterior capítulo se refería a conceptos de aprendizaje supervisado (machine learning) y reconocimiento de patrones (pattern recognition) utilizados en el desarrollo de la tesis, este capítulo se enfoca con mayor profundidad en la aplicación directa sobre el problema de predicción de interacción proteína-proteína. En un primer lugar se detallan dos métodos considerados "clásicos" por varios autores [96, 152, 116] y a continuación se detallan los métodos utilizados en trabajos seleccionados que se han usado como base en el desarrollo de esta tesis desde su comienzo. Así pues, se detalla el método MSSC (Maximum Specificity Set Cover)[96] y otros métodos que utilizan información genómica y proteómica [152] como los enfoques bayesianos de Patil et al. [162] y Jansen et al. [106]. En capítulos posteriores proponemos varias metodologías para la predicción de interacción proteína-proteína utilizando aprendizaje supervisado (SVM), resolviendo también problemas planteados aquí como la selección de un conjunto de negativos fiables o la selección de características relevantes entre otros.

3.2 Métodos clásicos: Metodo Asociativo AM y Estimación de Máxima Verosimilitud MLE

En la literatura, varios autores usan los métodos AM (método de asociación) y MLE (estimación de máxima verosimilitud, Maximum Likelihood Estimation) como base, considerándolos ya "clásicos", para posteriormente realizar un análisis de otros métodos

de trabajos seleccionados [96][152][116]. En dichos trabajos se asume la interacción dominio-dominio como indicador de que el par de proteínas asociadas a esos dominios interaccionan.

Antes de realizar una descripción de los métodos, hay que dar unas medidas de calidad, normalmente se utiliza especificidad y sensibilidad. Por tanto a continuación, se expondrá una serie de definiciones de conceptos y medidas de especificidad, sensibilidad y curva ROC (Receiver Operating Characteristic, o Característica Operativa del Receptor) muy utilizadas en la proteómica.

3.2.1 Especificidad y Sensibilidad

Son medidas estadísticas usadas como test (prueba) de clasificación binaria. La sensibilidad es la proporción de positivos que son correctamente clasificados (es decir, el porcentaje de gente enferma que son identificados bajo la condición); y la especificidad mide la proporción de negativos que son correctamente identificados (es decir, el porcentaje de gente que no está enferma que se identifican no teniendo la condición) [162, 170]. Están muy relacionados con los conceptos estadísticos de error tipo I y II.

Para cualquier test, hay normalmente dos medidas de tipo "trade-off" (si una mejora, la otra empeora). Por ejemplo, en el ajuste de una determinada fabricación en la que se comprueba los fallos, por un lado se puede correr el riesgo de descartar componentes funcionales (baja especificidad), con objeto de incrementar la probabilidad de identificar casi todos los componentes que fallan (alta sensibilidad). Este "trade-off" puede representarse gráficamente usando una curva ROC (ver siguiente sección 3.2.2).

Imaginando un escenario en que a las personas se le realiza una prueba para comprobar si están enfermas, es decir una prueba dicotómica, que se clasifica a cada paciente como sano o enfermo en función de que el resultado de la prueba sea positivo o negativo. En casos como éste, generalmente un resultado positivo se asocia con la presencia de enfermedad y un resultado negativo con la ausencia de la misma. Cuando se estudia una muestra del paciente, nos encontramos que el resultado de la prueba puede ser correcto (verdadero positivo y verdadero negativo) o incorrecto (falso positivo y falso negativo)[170].

Pues la sensibilidad se define como [170]:

$$sensibilidad = \frac{numero\ de\ Verdaderos\ Positivos(TP)}{numero\ de\ TP + numero\ de\ Falsos\ Negativos(FN)} \tag{3.1}$$

Una sensibilidad del 100% quiere decir que la prueba reconoce a toda la gente que está enferma como tal. La sensibilidad por sí sola no nos muestra cómo de bien la prueba predice otras clases (los casos negativos). Necesitamos al menos conocer la especificidad.

La especificidad, en este caso se define como [170]:

$$especificidad = \frac{numero\ de\ Verdaderos\ Negativos(TN)}{numero\ de\ TN + numero\ de\ Falsos\ Positivos(FP)} \tag{3.2}$$

Una especificidad del 100% quiere decir que la prueba reconoce a todas las personas sanas como sanas. Como en el caso anterior la especificidad por sí sola no nos muestra cómo de buena es la prueba reconociendo los casos positivos. Necesitamos conocer al menos la sensibilidad. Para el problema de la interacción proteína-proteína, se traslada la terminología de sano y enfermo a interacción o no interacción de una pareja de proteínas.

3.2.2 Análisis de las curvas ROC (Receiver Operating Characteristic)

Una curva ROC (Receiver Operating Characteristic) es una representación gráfica de precisión de una prueba (test) y expresa el balance "trade-off" entre la sensibilidad y la especificad de la prueba. La sensibilidad de una prueba se define como la habilidad de identificar un verdadero positivo en un conjunto de datos (dataset). La especificidad se define como la habilidad de identificar un verdadero negativo en un dataset [162].

Por tanto definimos,

Sensibilidad (tasa de verdaderos positivos TP): $\frac{TP}{T}$,

Especificidad= $\frac{TN}{F}$

1-Especificidad (tasa de falsos positivos FP)= $\frac{FP}{F}$

donde TP = número de verdaderos positivos, TN = número de falsos positivos, T = número total de positivos, F = número total de negativos. La curva ROC se dibuja con la sensibilidad en el eje-Y y (1-especificad) en el eje X. En la figura 3.1 se presenta un ejemplo de curva ROC. En machine learning (aprendizaje supervisado) entre otros campos, el estadístico AUC (Area Under Curve) de la curva ROC es ampliamente usado para comparar modelos [66, 86]. AUC es igual a la probabilidad de que un clasificador clasifique una instancia positiva mayor que una negativa elegida aleatoriamente. Es decir, éste parámetro AUC se usa para evaluar cómo de buena es la prueba (es decir la bondad), normalmente se tienen en cuenta los valores entre 1 (para una prueba perfecta) y 0,5 (prueba inútil).

3.2.3 AM: Association Method (Método Asociativo)

El método de asociación simplemente asigna una probabilidad de interacción a cada par de dominios (d_m, d_n); I_{mn} es el número de pares de proteínas que interaccionan

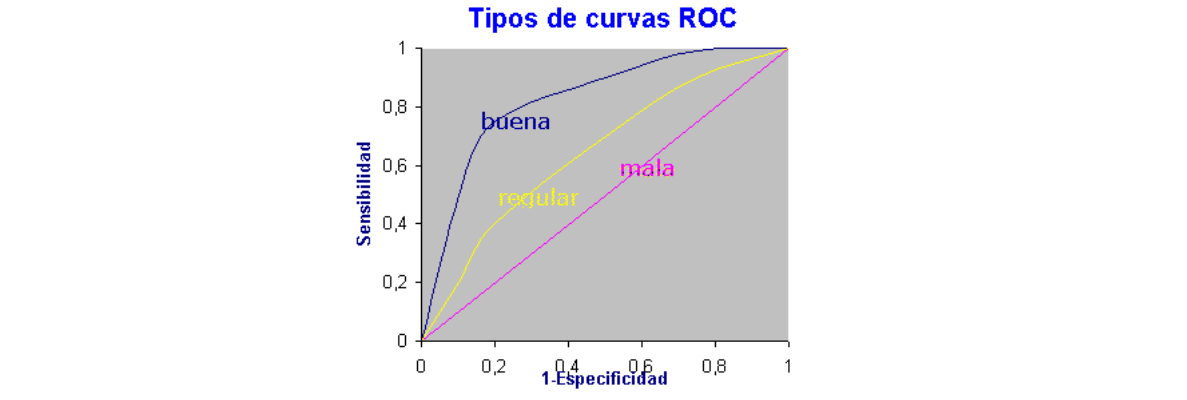


Fig. 3.1: Ejemplo de curva ROC. Figura extraída de un curso de bioestadística del Hospital Universitario Ramón y Cajal de la Comunidad de Madrid en la url http://www.hrc.es/bioest/roc_1.html. Se muestra en la figura tres curvas para tres hipotéticas pruebas con su evaluación.

que contienen (d_m, d_n), y N_{mn} es el número total de pares de proteínas que contienen (d_m, d_n) [96].

$$P_r(d_m, d_n) = \frac{I_{mn}}{N_{mn}} \tag{3.3}$$

3.2.4 MLE: Maximum Likelihood Estimation (Estimación de máxima verosimilitud)

El método de estimación de máxima verosimilitud [54] asume que dos proteínas interaccionan si al menos un par de dominios de las dos proteínas interaccionan. Asumiendo que las interacciones entre diferentes pares de dominios son independientes, la probabilidad de una interacción potencial entre un par de proteínas (P_i, P_j) es [96]

$$P_r(P_{ij} = 1) = 1 - \prod_{(d_m, d_n)\epsilon(P_i, P_j)} (1 - \lambda_{m,n}) \tag{3.4}$$

donde $\lambda_{m,n}$ denota la probabilidad de que el domino d_m interaccione con el dominio d_n. Normalmente se usa el algoritmo EM (Expectation Maximization) para maximizar la probabilidad como ocurre en [54][96]. λ observamos que coincide con el método de asociación AM. En el trabajo de Deng [54], se demuestra que la máxima verosimilitud (likelihood) es [96]

$$L = \prod (E(O_{ij} = 1))^{O_{ij}} (1 - E(O_{ij} = 1))^{1 - O_{ij}} \tag{3.5}$$

Siendo O_{ij} la variable para el resultado de interacción observada para proteínas P_i and P_j: donde $O_{ij} = 1$ si P_i y P_j interaccionan; de lo contrario, $O_{ij} = 0$. La verosimilitud L es una función de $\theta(E(d_i, d_j), f_p, f_n)$ donde $E(d_i, d_j)$ representa la probabilidad de que el dominio d_i y d_j interaccionen mientras que f_p y f_n indican

los índices de interacciones de falsos positivos y negativos en la red de interacciones subyacente, son valores fijados. Por tanto f_p y f_n se pueden definir como [96]:

$$f_p = P_r(O_{ij} = 1|P_{ij} = 0) \tag{3.6}$$

$$f_n = P_r(O_{ij} = 0|P_{ij} = 1) \tag{3.7}$$

La maximización de L por EM. logra un 42.5% de especificidad y 77.6% de sensibilidad para unos datasets determinados por [101][191] en [96].

3.3 Trabajo seleccionado: Predicting Protein-Protein Interactions from Protein Domains Using a Set Cover Approach

El método Método Maximum Specificity Set Cover (MSSC) o Cobertura de conjunto de máxima especificidad ha sido propuesto por Huang et al. [96]. Dicho método ha sido seleccionado en el desarrollo en este capítulo por su valor formativo a la vez que novedoso, sirviendo sus conceptos como referencia a lo largo de ésta tesis. MSSC usa un enfoque basado en la cobertura de conjuntos (en inglés cover, y en matemáticas recubrimiento), tomando aquellos pares de dominios que cubran las interacciones dominio-dominio dadas. Decimos que un par de dominios cubre una interacción proteína-proteína si las dos proteínas que interaccionan contienen los dos dominios respectivamente. Por tanto, el problema de interacción de proteínas se puede definir como el problema de encontrar un conjunto de pares de dominios que represente las interacciones proteína-proteína dadas. Lo ideal es que el conjunto de pares de dominios contenga el número menor posible de falsos positivos. Los falsos positivos por tanto serán las interacciones proteína-proteína predichas que no están incluidas en la red de interacciones de entrada al algoritmo. La información de este método ha sido extraída de la tesis del autor [95] y de su trabajo [96].

3.3.1 El problema de cobertura de conjuntos (Set Cover)

Suponemos que X es un conjunto infinito $\mathcal{F} = \{S_i, 1 \leq i \leq t\}$ es una familia de subconjuntos de X que pueden cubrir X, es decir, $X = \bigcup_{S \in \mathcal{F}} S$. El problema de la cobertura de conjuntos consiste en encontrar un subconjunto $\mathcal{C}$ perteneciente a $\mathcal{F}$ que cubra X,

$$X = \bigcup_{S \in \mathcal{C}} S \tag{3.8}$$

y C también se requiere para satisfacer ciertas condiciones dependiendo de lo específico del problema. Por ejemplo, el problema de conjunto de mínima cobertura (MSC, Mininum set-cover) es encontrar una $\mathcal{C}$ con la mínima cardinalidad $|\mathcal{C}|$ (el número de elementos en $|\mathcal{C}|$), y el problema del conjunto de cobertura mínima exacta

(MESC, Mininum exact set-cover) requiere que $\Sigma_{S\in\mathcal{C}}|S|$ sea minimizado. MSC y MSEC son problemas NP-completos [96].

3.3.1.1 Generalizando el problema de la cobertura de conjuntos

Generalizamos el problema de la cobertura de conjuntos encerrando X en un conjunto mayor Y, como la imagen 3.2. Suponemos que Y es un conjunto finito, $X \subseteq Y$ y $\mathcal{F}$ es una familia de subconjuntos de Y que pueden cubrir X, es decir, $X \subseteq_{S\in\mathcal{F}} S$. El problema de cobertura de conjuntos generalizado consiste en encontrar un subconjunto $\mathcal{C}$ perteneciente a $\mathcal{F}$ que cubra X [96],

$$X \subseteq_{S\in\mathcal{C}} S \tag{3.9}$$

y C también se requiere que satisfaga ciertas condiciones dependiendo del problema, como ocurría anteriormente.

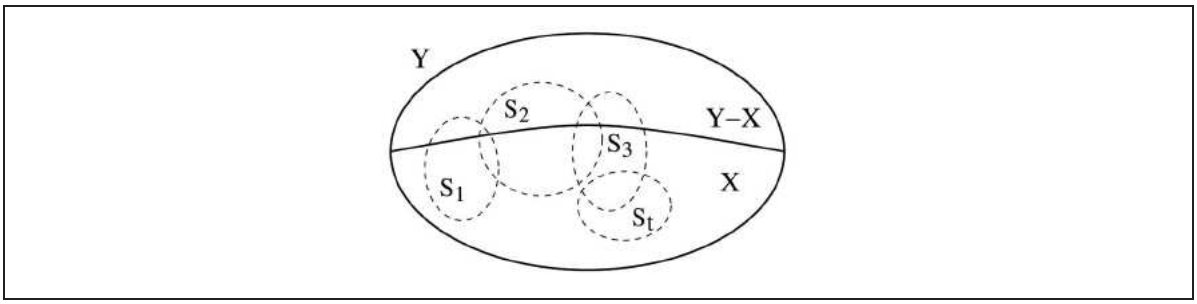


Fig. 3.2: Problema de cobertura de conjunto generalizado: X es un subconjunto de Y, y $\mathcal{F} = \{S_i, 1 \leq i \leq t\}$ es una familia de subconjuntos de Y. Figura extraída del trabajo de Huang et al.[96]

3.3.1.2 Adaptando las Interacciones de Proteínas como un problema de cobertura de conjuntos

Experimentalmente se sabe que el árbol de interacciones de proteína-proteína puede ser modelado como un grafo $G = (P, E)$, donde P es el subconjunto de proteínas y E es el subconjunto de relaciones (interacciones). Las proteínas son los vértices (nodos) de G. Hay una relación entre dos proteínas si y solo si interaccionan las dos. Formalmente, la red de interacción de proteínas se representa como un problema de cobertura de conjuntos definiendo [96]:

$$Y = \{todas\ los\ pares\ de\ proteinas\ (P_i, P_j)\ |P_i, P_j \in P\} \tag{3.10}$$

$$X = \{pares\ de\ proteinas\ (P_i, P_j)\ |P_i\ interacciona\ con\ P_j\ en\ G\} \tag{3.11}$$

y $\mathcal{F}$ es el conjunto de todas los pares de dominios (d_m, d_n) (ver la figura 3.3 para una representación esquemática de estas relaciones). Un par de dominios (d_m, d_n) se

ve como un subconjunto de Y. Especialmente, si un par de proteínas (P_i, P_j) (un elemento en X) contiene (d_m, d_n), entonces (P_i, P_j) pertenece al subconjunto (d_m, d_n). En la figura 3.4 se muestra cómo todos los pares de dominios posibles pueden cubrir una interacción de proteínas, elementos en un par de dominios. Algunos elementos (pares de proteínas) son interacciones dadas. Se supone que estamos buscando un subconjunto $\mathcal{C}$ de $\mathcal{F}$ que cubra cada elemento (P_i, P_j) en X. Un elemento de $\mathcal{C}$ corresponde a un par de dominios (d_m, dn). Si (d_m, dn) cubre (P_i, P_j), entonces dos proteínas P_i y P_j contienen d_m y d_n, respectivamente; así que (dm, dn) pueden representar la interacción entre P_i y P_j. Por lo tanto, tenemos un conjunto de pares de dominios que representa a la red de proteínas G. Por otro lado, suponiendo que hay un conjunto D de pares de dominios para representar G. Para cada elemento (P_i, P_j) en X, hay un par de dominios (d_m, d_n) de D para presentar su interacción. Como (d_m, d_n) puede ser visto como un elemento en $\mathcal{F}$, la colección $\mathcal{C}$ de todos los pares de proteínas de D es un subconjunto de $\mathcal{F}$, y $\mathcal{C}$ cubre por tanto a X [96].

Téngase en cuenta que, en matemáticas, cuando se tiene una colección de subconjuntos A de un conjunto X es un recubrimiento (en inglés cover, y que en esta sección se usa cubrir en lugar de recubrir) de X, o también llamado una cubierta de X, si la unión de los elementos de la colección A es igual a X. Y por tanto, un subrecubrimiento (en inglés subcover) de C es un subconjunto C (formado entonces por elementos de C) que todavía cubre X [149].

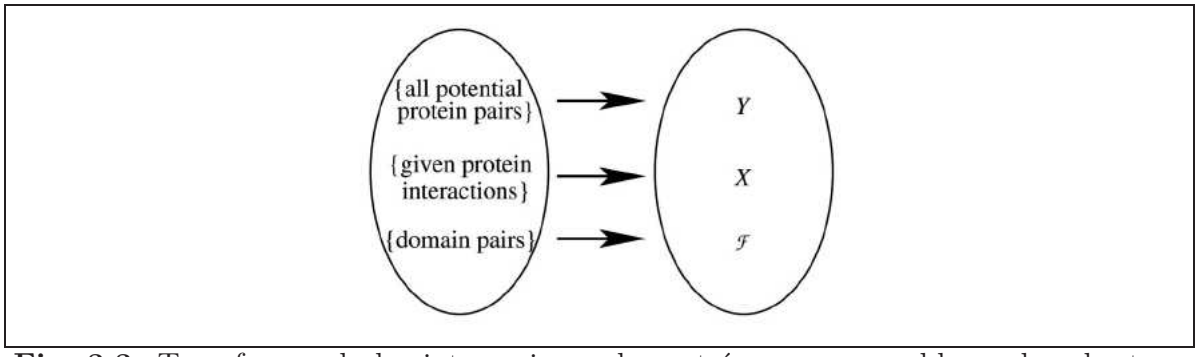


Fig. 3.3: Transformando las interacciones de proteínas en un problema de cobertura de conjuntos: El gran conjunto Y se toma como el conjunto de todas los posibles pares de proteínas, y el subconjunto X se toma como el de interacciones proteína-proteína dado, y la familia F se toma como el conjunto de todas los pares de dominios. Figura extraída del trabajo de Huang et al.[96]

3.3.1.3 Algoritmo Maximum Specificity Set Cover (MSSC)

Hay mucho modos de elegir pares de dominios para representar la red de interacción de proteínas. *AM* simplemente usa todos los posibles pares de dominios para explicar la interacción proteína-proteína, es decir, usa $\mathcal{F}$ para cubrir X, resultando

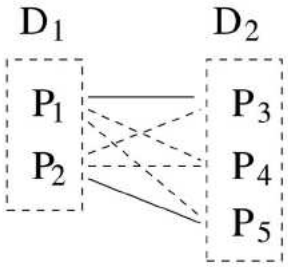


Fig. 3.4: Cada dominio en el par de dominios (D_1, D_2) está contenido en una lista de proteínas. (D_1, D_2) es el subconjunto de Y, el conjunto de todas los pares de proteínas. Como tal, (D_1, D_2) cubre estos pares de proteínas, donde hay de hecho interacción (líneas sólidas, no punteadas). Nuestro algoritmo busca la mejor cobertura que explica las interacciones de la forma más extendida, lo más extenso posible.

baja especificidad [54]. Estamos interesados en encontrar un subconjunto de pares de dominios que nos permita presentar la red de interacción de proteína-proteína maximizando la especificidad y la sensibilidad en el conjunto de entrenamiento. El problema MSSC es encontrar un subconjunto $\mathcal{C}$ de $\mathcal{F}$ que cubra X mientras que [96]

$$m(\mathcal{C}) = \Sigma_{S \in \mathcal{C}} |S - X| \tag{3.12}$$

sea minimizado.

Si observamos la figura 3.5, MSSC permite que el subrecubrimiento (subcover) de $\mathcal{C}$ cubra el solapamiento con X, mientras que el solapamiento $Y - X$ (fuera de X) sea minimizado, una restricción de optimización que permite a MSSC maximizar la especificidad de la cobertura seleccionada porque los falsos positivos se consideran que aparecen sólo en $Y - X$. En [96] optaron por implementar un algoritmo voraz ("Greedy") como rutina básica MSSC 3.1. Se considera que los falsos positivos aparecen en $Y - X$, donde Y sería el conjunto de todas las posibles proteínas y U representa la parte no cubierta de X, $\mathcal{E}$ es el subconjunto de $\mathcal{F}$ que no ha sido elegido por el algoritmo.

Por tanto, en este algoritmo, en cada paso cuando el subconjunto necesita ser seleccionado, seleccionamos uno cuyo ratio se encuentre entre la parta de fuera de X y la parte de dentro de U sea minimizado (3.5). El algoritmo "greedy" es una aproximación a la solución pero tiene relación con la solución óptima (para demostración, ver en [96]).

En [96] podemos observar la aplicación del algoritmo a diversos datasets (cinco en total) entre los cuales encontramos a Bader[10], complejos de proteínas e interacciones físicas (que es posible extraerlos de la base de datos de MIPS), todos de la levadura de la cerveza (*S. cerevisiae*). Han usado, en los mencionados datasets, dos conjuntos disjuntos de entrenamiento (training) y prueba (test), fruto de la división de los datasets (80% de training y 20% de test). Resumiendo, en general, para los dataset mencionados

GREEDY_MSSC($Y, X, \mathcal{F}$)
$U \leftarrow X$
$\mathcal{E} \leftarrow \mathcal{F}$
$\mathcal{C} \leftarrow \emptyset$
while $U \neq \emptyset$ **do**
  select an $S \in \mathcal{E}$ con el minimo $\frac{|S-X|}{|S \cap U|}$ (la interacion se rompe por $|S \cap U|$)
  $U \leftarrow U - X$
  $\mathcal{E} \leftarrow \mathcal{E} - \{S\}$
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{S\}$
end while

Algoritmo 3.1: Algoritmo MSSC propuesto por Huang et al. [96]

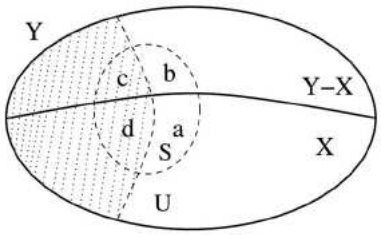


Fig. 3.5: El área sombreada está cubierta por $\mathcal{C}$. U es el área no sombreada en X. El conjunto candidato $\mathcal{C}$ se divide en cuatro partes, a, b, c y d. MSSC elige un conjunto S con el mínimo $\frac{b+c}{a}$. El algoritmo "greedy" para MSSC permite el solapamiento de subrecubrimiento (subcover) dentro de X, de hecho, incrementando la probabilidad de interacción para un par de proteínas. Figura extraída del trabajo de Huang et al.[96]

MSSC y MLE consigue mejor sensibilidad que AM. Se observa que MLE no es capaz de obtener buena especificidad en muchos casos, pero en general MLE y MSSC obtienen resultados similares, diferenciándose fundamentalmente en el tiempo de ejecución, resultado MLE un algoritmo que por su naturaleza iterativa obtiene resultados en horas mientras que MSSC en segundos [96]. Los otros dos datasets que usaron, son derivados de la 26S Proteasoma y RNA Polimerasa II de la mencionada levadura y no se obtiene buenos resultados debido, según los autores [96], a la falta de información sobre estos complejos. Para la comparación en la ejecución de los diferentes métodos, se han usado coeficientes de correlación de Pearson y uso de una distancia GO que considera términos comunes y distintos (la fórmula basada en la de "Czekanowski-Dice", ver [138]). Además, también se ha hecho uso de una prueba t-Student. Para más información leer el trabajo doctoral de Huang [95, 96].

3.4 Otros trabajos seleccionados: Enfoques de predicción de interacción proteína-proteína usando información genómica y proteómica

Como hemos mencionado anteriormente, en la literatura aparecen varios enfoques de predicción de interacción proteína-proteína, tomando varias fuentes de información genómica. Según el trabajo de Jansen et al. [106], se demuestra que las fuentes de información genómica, tomadas por sí solas, son predictores débiles de interacciones y sólo puede dar predicciones fiables cuando combina varias fuentes. El trabajo de Jansen et al. [106] ha sido referencia para muchos artículos, incluso para la creación de bases de datos [162], Patil y Nakamura crean una base de datos de homologías (Hint) extendiendo el trabajo de Jansen et al.[106]. Aunque los dos artículos mencionados usen aproximaciones basadas en redes bayesianas, existen otros tipos de enfoques, por ejemplo [152] que usa programación lógica como aproximación al problema; si bien Nguyen comenta que Jansen et al. usa características genómicas débilmente predictivas, Nguyen toma hasta 22 características de diferentes fuentes genómicas.

Antes de comenzar a realizar una descripción de las metodologías que tomaremos como referencia, al usar aproximaciones bayesianas, daremos un ligera descripción de redes bayesianas vinculadas al uso de características genómicas.

3.5 Aplicación de redes bayesianas usadas en trabajos seleccionados de predicción de proteína-proteína

Las redes bayesianas pueden ser usadas para combinar evidencias de diferentes fuentes y calcular la probabilidad posterior (posterior odds) de un evento basado en la evidencia anterior (prior evidence). La relación entre la probabilidad posterior (posterior odds) y la probabilidad anterior (prior odds) de encontrar una verdadera interacción está dada por una regla de Bayes como sigue [162]:

$$O_{posterior} = L(g_1, g_2, g_3, ..., g_N)O_{prior} \tag{3.13}$$

donde $g_1, g_2, g_3, ..., g_N$ son características genómicas de una interacción. O_{prior} es la probabilidad anterior de que una interacción sea verdad, y $O_{posterior}$ es la probabilidad posterior de que una interacción, considerando sus N características genómicas, sea verdad. Por tanto, $L(g_1, g_2, g_3, ..., g_N)$ es lo que llamamos razón de verosimilitud (likelihood ratio) de una interacción con características genómicas [162].

Así pues, podemos definir la probabilidad anterior como [162]:

$$O_{prior} = P(true)/1 - P(true) \tag{3.14}$$

donde $P(true)$ es la probabilidad de que una interacción sea verdad.

La probabilidad posterior podemos definirla mediante probabilidad condicionada

$$O_{posterior} = \frac{P(true|g_1, g_2, g_3, ..., g_N)}{P(false|g_1, g_2, g_3, ..., g_N)} \tag{3.15}$$

donde $P(true|g_1, g_2, g_3, ..., g_N)$ es la probabilidad de que una interacción con N características genómicas sea verdad.

Mediante la ecuación 3.13, la verosimilitud (Likelihood ratio) podemos escribirlo como [162]:

$$L(g_1, g_2, g_3, ..., g_N) = \frac{P(g_1, g_2, g_3, ..., g_N|true)}{P(g_1, g_2, g_3, ..., g_N|false)} \tag{3.16}$$

donde $P(g_1, g_2, g_3, ..., g_N|true)$ es la probabilidad de que una interacción verdadera tenga N características genómicas.

Si las N características genómicas, $g_1, g_2, g_3, ..., g_N$ son condicionalmente independientes, entonces la red resultante bayesiana se le llama red bayesiana tipo naïve y su verosimilitud (likelihood) puede ser dada como el producto de las razones de verosimilitud (likelihood ratios) de cada característica, por tanto podemos escribir [162]:

$$L(g_1, g_2, g_3, ..., g_N) = \prod_{i=1}^{N} L(g_i) \tag{3.17}$$

$$O_{posterior} = \prod_{i=1}^{N} L(g_i) O_{prior} \tag{3.18}$$

$$L(g_i) = \frac{P(g_i|true)}{P(g_i|false)} = \frac{TP_i/T}{FP_i/F} \tag{3.19}$$

donde T son todas las interacciones que son verdad y F son todas las interacciones que son falsas.

TP_i sería el número de interacciones que son verdad en el conjunto de datos con la característica i-ésima. FP_i sería el número de interacciones que son falsas en el conjunto de datos con la característica i-ésima.

Por tanto, para cualquier organismo que cumpla $L(g_1, g_2, g_3, ..., g_N) > 1$, los resultados serán $O_{posterior} > O_{posterior}$. Esto es debido a que en la ecuación 3.13, O_{prior} es una constante y depende del número de interacciones en cualquier organismo. Por consiguiente, $O_{posterior}$ es directamente proporcional a $L(g_1, g_2, g_3, ..., g_N)$. Así, que la probabilidad posterior siendo verdad, si tiene una o más características genómicas, se incrementa tanto como se incrementa $L(g_1, g_2, g_3, ..., g_N)$, es decir, cuanto mayor es $L(g_1, g_2, g_3, ..., g_N)$ mayor será la probabilidad de que una interacción sea verdad [162].

3.5.1 Enfoques bayesianos

En esta sección hablaremos fundamentalmente de dos enfoques bayesianos planteados por Jansen et al. [106] y de Patil et al. [162] que han servido de referencia en esta tesis. Comenzando por Jansen et al. [106], elige un enfoque bayesiano para integrar la información de interacción ya que permite la combinación probabilística de múltiples datasets [57] y demuestra su aplicación a la levadura de la cerveza. La idea básica es medir cada fuente de evidencia para interacciones comparándolo contra muestras de ejemplos conocidos positivos y negativos ("gold-standars"), mostrando fiabilidad estadística. Las redes bayesianas tienen varias ventajas: permite la combinación de tipos de datos muy diferentes (es decir, numéricos y categóricos), permite acomodarse a datos ocultos (missing data), por su naturaleza proporciona fiabilidad.

Como medida de fiabilidad, el solapamiento de varias fuentes de información con el gold-standard puede expresarse en términos de razón de verosimilitud (likelihood ratio, L). Tanto Patil et al.[163] como Jansen et al. [106] hacen uso del enfoque de redes bayesianas comentado en el apartado 3.5. Se predice que un par positivo si la razón de verosimilitud (likelihood ratio) excede un valor de corte particular ($L > L_{cut}$), y lo contrario es un par negativo.

Para dar una medida de cómo la predicción funciona, tanto Jansen et al. [106] como Patil et al. [163], usan un método de validación cruzada (CV) p-fold que ya se explicó en el apartado 2.2.2 ([106] CV p-fold p=7, [163] CV p-fold, p=10), evaluando el número de TP y de FP predichos en el conjunto de prueba.

Jansen et al. [106], se habla de resultados probabilísticos en interactomas, dando a cada par de proteínas una medida de probabilidad, pero se divide los datos en dos grupos, datos de interacción experimentales (objetivos experimentalmente, que se llaman PIE) y datos de predicción "de novo" de complejos de proteínas (que se llaman PIP). Como PIE toma datos de experimentos de alto-rendimiento (high-thoughput experiments) que comprende dos tipos: datos de muestras de alta-escala doble-híbrido (large-scale two-hybrid screen, 2YH)[101][212], datos de experimentos "in vivo" (pull-down) [70][92]. Al PIE se le aplica una red bayesiana completamente conectada ya que existe evidencia correlada (correlated evidence, se solapan los datasets). PIP está formado por datos de expresión de ARNm (compendio Rosetta [179] y ciclo celular), funciones biológicas (catálogo funcional de MIPS 4 y procesos biológicos de GO) y datos sobre si las proteínas son esenciales (ver [106]). Para PIP se le aplica una red bayesiana tipo naïve ya que no existe correlación entre los datos.

Al final combina PIP, PIE y los datos gold-standard en un grupo PIT, para dar una visión global al problema usando redes bayesianas tipo naïve ya que PIP y PIE muestran evidencia no correlada. Cuando PIE y PIP son comprobados contra el gold-standard, se observa que el índice TP/FP se incrementan monotónicamente con L_{cut}, confirmando que L es una buena medida para pedir la probabilidad de interacción. Se observa como hemos comentado anteriormente, que para poder dar una predicción

fiable hay que combinar las características genómicas ya que por sí solas obtenemos predictores muy pobres. En este caso [106], se obtiene mayor sensibilidad con PIP que con PIE comparando las razones TP/TP.

Mediante procedimiento de voto (cada característica aporta un voto a favor o en contra para la clasificación final, si un par de proteínas interaccionan o no), con esto se pretende comparar la red bayesiana con el rendimiento de los métodos machine learning complejos, comprobando con ello que la red Bayesiana logra mayor sensibilidad con comparables índices TP/FP. La precisión que logramos con PIP es comparable a la que se obtiene con PIE, y en cuanto se combinan simultáneamente se consigue una mayor cobertura de la red de interacciones. Para mayor información leer [106], en el que en el material adicional que explica todo con mayor detalle los experimentos llevados a cabo, realizando incluso comprobaciones de sesgo, de los falsos positivos al crear el grafo de interacción, incluyendo una distribución conjunta de las interacciones proteína-proteína (PPIs), y usando para una comprobación mejor de las predicciones en PIP mediante información de experimentos "TAP-tagging" (ver apéndice D).

Como bien se comenta en Jansen et al. [106], se han tomado características débilmente predictivas que combinándolas en una red bayesianas se transforma en un predictor fiable de interacción de proteínas. Pero también es posible añadir más características que mejoren mucho más los resultados, incluso extendiendo la exploración a más organismos además de la levadura.

Patil et al. [162], realiza un trabajo a partir de Jansen et al., aplica una aproximación bayesiana para varios organismos: S. *cerevisiae, C. elegans, D. melanogaster y H. sapiens*. Para la levadura toma los mismos datos que Jansen et al: los de sistema doble híbrido [101][212], y los datos inferidos de espectrometría de masas por coimmunoprecipitación (co-IP, ver apéndice D), [70][92]. Crea un gold standard positive y negative. Y toma tres características genéticas, por tanto, a cada interacción proteína-proteína le asigna una probabilidad, y será una interacción verdadera si cumple algo o todas las condiciones siguientes [162]:

- Las proteínas que interaccionan tienen homólogos con los que interaccionan
- Cada proteína que interacciona en un par, posee un dominio Pfam que interacciona con otro dominio en PDB.
- Las proteínas que interaccionan tienen al menos una anotación GO idéntica.

Para el cálculo de la fiabilidad de cada característica genómica, se ha usado interacciones proteína-proteína de datasets HT (high-throughput) de la levadura de la cerveza (yeast en inglés). El objetivo es maximizar las interacciones identificadas en el conjunto gold positivo (alta sensibilidad) y al mismo tiempo maximizar el número de interacciones identificadas en el conjunto gold negativo (alta especificidad). Para cada característica genómica, los TP (verdaderos positivos) son aquellas interacciones que

están en el gold positivo (GSP), y son FP (falsos positivos) aquellas que estaban en un conjunto gold negativo. Usando estos valores (ver sección 3.5). se calcula la razón de verosimilitud (L, likelihood ratio). Una $L > 1$ indica la habilidad de la característica genómica de identificar más verdaderos positivos que falsos positivos [162].

Para poder aplicar la red bayesiana, es necesario que las características genómicas sean independientes, de ahí que se suela aplicar el coeficiente de correlación de Pearson para comprobar la independencia. Un valor de L mayor que 1 representa la probabilidad posterior (posterior odds) de una interacción que es verdadera mayor que la probabilidad anterior (prior odds).

Los autores de los datasets de alto-rendimiento (HT, High-Throughput) suelen asignar un nivel de confianza a las interacciones[162], las interacciones que o se confirman experimentalmente o tienen una alta probabilidad de que sean verdad se las estima como altamente confiable (high confidence) mientras que el resto poseen una confianza baja (low confidence). En el trabajo de Patil et al. [162] se observa cómo se prueba el solapamiento entre las interacciones verdaderas solapadas y los datasets de confianza baja y alta dadas por los autores. La mayoría de las interacciones de alta confianza se predicen como verdad en todos los datasets excepto en el de humanos (H. sapiens).

El método de Patil et al.[162] obtiene alta sensibilidad (89.7%) y una buena especificidad (62.9%). Con respecto a la información de estructura (dominios Pfam respecto PDB) se obtiene alta especificad (número bajo de falsos positivos FP), pero la sensibilidad (número de verdaderos positivos) está limitada por el pequeño número de estructuras en PDB, la sensibilidad aumenta conforme más número de estructuras tengamos. La segunda característica que muestra mayor fiabilidad es la de anotaciones GO (identifica el mayor número de verdaderos positivos). La información de secuencia en forma de interacciones homólogas (base de datos HINTdb [163]) no proporciona la fiabilidad esperada, ya que quizás [162] no está limitada a interacciones ortólogas o parálogas. Aunque cada característica permite independientemente predecir si una interacción es verdad, la combinación de dos o más características mejora la predicción.

3.5.2 Comprobación de las predicciones

En caso de Patil et al. [163], el conjunto de entrenamiento (training) se ha usado para el cálculo de las razones de verosimilitud (likelihood ratios) y el conjunto de prueba (test) se ha usado para el cálculo de sensibilidad y especificidad partiendo del mismo conjunto de datos de la levadura de la cerveza -*S. cerevisiae*- (datos obtenidos por métodos de alto-rendimiento, yeast high-throughtput data set). Para realizar un cálculo de comprobación de las predicciones, se ha usado validación cruzada p-fold (p-fold cross validation) con p = 10; esto es, dividimos el conjunto de gold standard positivos y negativos en 10 conjuntos aproximadamente iguales. Se usan 9 de esos conjuntos para calcular las verosimilitudes (likelihood ratios) para cada característica genómica. Se identifican entonces los verdaderos positivos y los falsos negativos en el

conjunto que quedó fuera usando las razones de verosimilitud (likelihood ratios). Se repite en cada turno hasta que cada uno de los 10 conjuntos haya sido conjunto de prueba (test) y los 9 restantes hayan sido de entrenamiento (training). Entonces se suman el número de verdaderos positivos y el de falsos positivos entre los 10 conjuntos de prueba obteniendo la sensibilidad y especificidad y pintando la curva ROC. Para Jansen et al. [106] la validación cruzada p-fold fue realizada con p=7.

3.5.3 Selección de ejemplos positivos y ejemplos negativos

En el trabajo de Jansen [106] se nos menciona que los datos "gold-standard" idealmente (para su enfoque bayesiano):

- independientes de las fuentes de datos que sirven como evidencia.
- deben haber suficientes datos para que estadísticamente el cálculo sea significativo, fiable.
- deben estar libres de sesgo sistemático.

Podemos generalizarlo a enfoques de machine-learning, ya sean redes neuronales, árboles de desición o incluso Support-Vector Machines (SVM, máquinas de soporte vectorial).

Jansen et al. [106] (ver material adicional) comenta que los conjuntos gold-stardard no son ideales, pero se busca una buena aproximación, incluso puede existir una pequeña intersección entre el gold-stardard positivo y el negativo. De ahí que tomen como gold-standard positive el catálogo de complejos de la base de datos MIPS (cap. 4). Pero como es habitual, el gold-standard negative es más difícil de definir, pero es esencial para el entrenamiento. Por tanto, sintetizaron los negativos de una lista de proteínas que se encontraban en diferentes componentes (compartment) subcelulares[120].

Según Patil et al. [163], el gold standard positive está formado por:

- Todas las interacciones físicas de MIPS de sistemas doble-híbrido de levadura (2YH) (excluyendo las interacciones de Uetz [212] et al. e Ito et al [101]).
- Datos de complejos MIPS (excluyendo los datos de Gavin et al [70] y Ho et al. [92]).
- Datos experimentales 2YH (doble-híbrido de levadura) de las bases de datos DIP e Intact (ver capítulo 4).
- Interacciones encontradas en más de un dataset de alto-rendimiento (high-throughput datasets).

El gold-standard negative se forma basándose en la localización de proteína dentro de las células de la levadura [97]. Las proteínas que no existen dentro del mismo

compartimento subcelular se asume que no interaccionan ya que la mayoría de las interacciones tienen lugar en el mismo compartimento celular [106, 105, 191]

Saeed y Deane. [185] han conseguido un conjunto de ejemplos positivos y negativos muy fiables para la levadura de la cerveza (en inglés yeast), proporcionando más de 4 millones de negativos. Para la interacción de gold-standard negativo ha usado información de homología además de localización, expresión y función. Más concretamente, los conjuntos GSP (ejemplos positivos) se crean tomando interacciones de proteína de múltiples especies y devolviendo el subconjunto de interacciones que han sido verificadas en la existencia de una interacción homóloga. Para la verificación de los datos se ha tomado el esquema propuesto por Deng et al. [54] (MLE) y se ha comprobado mediante una curva ROC.

Para la construcción de los ejemplos negativos (GSN) para la levadura se ha tomado en consideración las características de función, expresión, localización celular y datos de homología. Las anotaciones funcionales para la levadura se han tomado de la base de datos GO. La notación de localización se tomo de MIPS y la información de expresión se tomó de Young Lab [94]. Las proteínas que cumplían esa notación se ha bajado de la base de datos CYGD de MIPS.

Las interacciones negativas se han predicho seleccionando un par aleatorio de proteínas que no comparten ninguna característica en común. Por tanto dos proteínas que fueron seleccionadas aleatoriamente y se clasifican como interacción negativa si cumplen los siguientes requisitos:

- No comparten ningún proceso funcional en común.
- No se solapan en ninguna localización celular.
- Su coeficiente de co-expresión se encuentra entre 0.3 y -0.3.
- No hay interacciones homólogas en ningún conjunto de datos (datasets) experimental.

Para analizar la influencia de cada característica de proteína en las interacciones negativas se generaron 100 subconjuntos de interacciones para cada característica de proteína individual y el método propuesto integrado. La exactitud de cada método fue comprobada midiendo el solape medio entre un conjunto de datos extraídos de la base de datos DIP (ver capitulo 4) y 100 subconjuntos de cada conjunto GSN (ejemplos negativos).

Se han generado los conjuntos gold-standard positivos (GSP) como subconjuntos de información de interacción de experimentos e información de secuencia. Así pues, a lo largo de todo el trabajo de Saeed y Deane demuestra que gracias a buenos ejemplos positivos y negativos mejora notablemente la predicción de los algoritmos para la interacción proteína-proteína. Así pues, se ha conseguido un conjunto de ejemplos negativos (GSN) fiables de más de 4 millones de proteínas ([185]).

Las interacciones homólogas pueden dividirse en dos subcategorías: interacciones parálogas y ortólogas. Las interacciones parálogas se consideraron por simplicidad, aquellas que existen entre proteínas homólogas dentro del mismo organismo y como ortólogas aquellas que existen entre proteínas homólogas de diferentes organismos. Con ello, se ha comprobado que las interacciones parálogas son más prevalentes y más influyentes en los métodos que Saeed y Deane [185] han usado en su comprobación que la inmensa mayoría de las ortólogas. Por consiguiente, esto indica que los métodos que predicen interacciones de proteína putativa y se basan en la conservación en otras especies, podría arrojar más información si sus análisis incluyeran información de interacciones parálogas.

Saeed y Deane. [185] propone usar el solapamiento entre las proteínas ortólogas de la mosca de la fruta (D. melanogaster, eucariota) y la E. coli (procariota), se espera encontrar más interaccionadas verificadas de la levadura de la cerveza (eucariota) ya que son especies más cercanas; se ha verificado un 73% de interacciones ortólogas verificadas de la levadura en la mosca de la fruta y un 13% de las interacciones ortólogas en la E. coli.

3.5.4 Selección de características

Como ya hemos mencionado, [106] se toma dos grupos de datos, unos experimentales que se llaman PIE y otros que no lo son llamados PIP. . Como PIE toma datos de experimentos de dos tipos: datos de muestras de alta-escala doble-híbrido [101][212] y datos de experimentos "in vivo" (pull-down) [70][92]. PIP está formado por datos de expresión de ARNm (compendio Rosetta [179] y ciclo celular), funciones biológicas (catálogo funcional de MIPS 4 y procesos biológicos de GO) y datos sobre si las proteínas son esenciales.

Con respecto a las funciones biológicas, las proteínas que interaccionan normalmente realizan el mismo proceso biológico [191, 123, 214], es por ello que dos proteínas que actúen en el mismo proceso biológico son más propensas a que interaccionen que otras que no lo están. De ahí que Jansen tomase como datasets información de catálogo funcional de MIPS y datos de procesos biológicos de GO (ver capítulo 4). Para cuantificar la similitud funcional entre dos proteínas, primero se considera cuál conjunto de clases funcionales comparte dos proteínas; y después se cuenta de todos los millones de pares de proteínas de las que disponemos (aprox. 18 millones en Jansen et al. [106] en la levadura de la cerveza -*S. cerevisiae*-) que compartan la misma clase funcional también. En general, cuanto más baja es la cuenta, más similar y específico es la descripción funcional de dos proteínas, mientras que cuanto más grande sea la cuenta, la relación entre las proteínas no es muy específica funcionalmente. Cuanto más baja sea la cuenta, Jansen et al [106] mayor oportunidad de que las dos proteínas estén en el mismo complejo.

En el trabajo de Patil et al. [162], con objeto de predecir si una interacción es verdadera, se han tomado las siguientes características genómicas basadas en:

- Interacciones homólogas. Usando su propia base de datos Hint [163], se han identificado todas las interacciones a partir de los datasets de alto-rendimiento (high-throughput, HT) que tenían interacciones homólogas, incluyendo ortólogas e interacciones homólogas. Una interacción se considera como homólogo a una interacción dada cuando cada una de sus proteínas que interaccionan tiene homologas que se encuentran interaccionando en las bases de datos DIP o IntAct. Hint usa una versión BLAST llamada PSIBlast con 5 interacciones y un e-value de corte de 10^{-8}.

- Anotaciones GO. Usando datos de una base de datos GO, identificamos todas las interacciones de los datasets de HT (High-Throughtput), donde las proteínas que interaccionan comparten al menos un término GO, con lo que las proteínas que interaccionan generalmente comparten una función en común [191].

- Dominios Pfam que interaccionan. Se han identificado todas las interacciones de los datasets de HT, donde cada proteína que interacciona tiene uno de los dominios Pfam que se encuentran interaccionando en la estructuras PDB en la base de datos 3Did (ver capítulo 4).

Una vez tomadas las características genómicas. La correlación entre cada característica genómica ha sido calculada con el coeficiente de correlación de Pearson (es común usarlo, ya se vió en [96]) para 100 interacciones aleatorias de los datasets HT. Posteriormente, la importancia de cada coeficiente de correlación fue comprobado usando un t-test (ver glosario t-student) usando 98 grados de libertad. Todas las características genómicas fueron independientes unas de otras.

Huang [96] tomó como datos los de la base de datos DIP, lo divide en dos partes, un 80% para entrenamiento y el resto para test, la selección la realiza de forma aleatoria. Estos datos son un subconjunto de los de Bader et al. [10], introduce un método para el cálculo de la calidad de las interacciones utilizando otras fuentes de información, incluyendo expresión ARNm, interacciones genéticas y anotaciones de bases de datos.

Con Nguyen [152] propone un método basado en dominios para predecir interacciones proteína-proteína usando programación de lógica inductiva (inductive logic programming ILP). Las dos características principales son las fusiones de dominio y las interacciones dominio-dominio. Toma como información características genómicas y proteómicas relevantes de las interacciones extraídas de cinco bases de datos de genómica y proteómica muy conocidas. Para comprobar la predicción del método, utiliza una validación cruzada 10-fold usando como datasets obtenidos de diversas fuentes de datos, por ejemplo extraídos de MIPS, Ito et al y Uetz et al. [101][212], de DIP, etc. en los que se ha estimado un índice llamado EPR [52] (se basa en la puntación de distancia entre todos los pares que interaccionan en un conjunto dato).

Más concretamente Nguyen [152], ha usado 22 características, que nombramos a continuación:

- **Fusión de dominio**. Se han tomado los datos de la base de datos de fusión de dominio (Domain Fusion Database [208]). Los dominios de proteínas que interaccionan tienen más posibilidad de fusionarse juntos en los dominios de proteínas que no interaccionan. De ahí que se haya tomado información para cada pareja de proteínas información sobre la fusión de dominios (más información en glosario, Proteínas de fusión). Esta es la característica 1.

- **Información de interacción de dominios**. Se ha extraído información sobre la interacción dominio-domino (DDI) de la base de datos Pfam, gracias a un recurso iPfam que describe las interacciones dominio-dominio que han sido observadas en las entradas de PDB. Cuando dos o más dominios aparecen en la misma estructura, se analizan los dominios para comprobar si forman una interacción considerando los enlaces formados y tomando una medida. Por tanto se ha tenido en cuenta si hay interacción DDI y también el número de DDI relacionada con una proteína De aquí se obtienen las características 2 y 3.

- **Base de datos UnitProt**. Se ha tomado de esta base de datos:

  – La información de los campos KW (keyword) que proporciona información funcional, estructural o de otras categorías.

  – Regiones o lugares de interés en las secuencias (línea FT).

  – Enzima codificada (línea EC)

  – Punteros a información relacionada de otras entradas de otras bases de datos: GO (término GO), PIR (identificador ID), PROSITE (identificador ID), Pfam (identificador ID), Intepro (identificador ID).

  Ahí tenemos por tanto las características de la 4 a la 11.

- **CYGD de MIPS**. Comprenhensive Yeast Gnome Database (CYGD) de MIPS que nos proporciona información detallada sobre la levadura de la cerverza (*S. cerevisiae*). Se han tomado información sobre los catálogos de proteínas siguientes:

  – Catálogo de funciones

  – Catálogo de localizaciones subcelulares

  – Catálogo de complejos

  – Catálogo de fenotipos

  – Catálogo de proteínas

  Así se mina la relación entre los catálogos, y se toman en consideración que proteínas que están en el mismo catálogo tienen una mayor posibilidad de interaccionar. Con ello suman desde la característica 12 hasta la 16.

- **Base de datos Interpro**. Una base de datos de familias de proteínas, dominios y sitios funcionales. Se ha considerado la asociación entre los identificadores Interpro y los términos GO. Es la característica 17.

- **Base de datos GO**. Se organiza en tres ontologías: función molecular, proceso biológico y componente celular. Los términos en una ontología se enlazan mediante dos relaciones: *is_a* y *part_of*. Por tanto se han tomado estas dos relaciones como características a tener en cuenta. (características 18 y 19).

- **Expresión de genes**. Las proteínas en el mismo complejo normalmente están co-expresadas, y por tanto esta característica es muy útil en la predicción de la interacción proteína-proteína. Se ha tomado la información de expresión de Jansen et al. [106]. Por tanto se ha utilizado como característica el coeficiente de expresión de genes. Aquí tenemos entonces la característica 20.

- **Características extra**. Se ha considerado también que el número de interacciones proteína-proteína dadas dos proteínas es una característica útil (característica 21). Y como última característica (la número 22) a tener en cuenta, se ha tomado la generalidad de interacción que es el número de proteínas que interaccionan con las dos proteínas compañeras, es decir, que forman una interacción.

Nguyen [152] ha comparado su método ILP con un método AM y un SVM usando un kernel lineal y los parámetros por defecto de SVMlight [110], en el que en la comparativa, su método mejora en la predicción a los dos comparados. Para ello ha usado ejemplos positivos y 1000 ejemplos negativos para la comparación y la construcción de las curvas ROC. Se ha tomado el corazón de los datos del dataset de Ito et al. [101] como ejemplos positivos, y como ejemplos negativos, se han seleccionado 1000 pares de proteínas aleatoriamente cuyos elementos se encontraban en componentes (compartments) subcelulares separados.

3.6 Conclusiones

La predicción de la interacción de proteína-proteína es muy importante, ya que la detección en laboratorio es muy laborioso y necesita una gran dedicación de tiempo, la predicción fiable tiene especial transcendencia en medicina o farmacia, ya que nos permite encontrar una cura para una enfermedad, o sino al menos poder pronosticarla y someterla a un tratamiento.

Con ello, en este capítulo, se ha realizado un estudio del estado de arte de métodos de predicción proteína-proteína centrándonos detalladamente en ciertos trabajos ([163, 96, 185, 106, 152]) en el que sus metodologías para afrontar este problema merecían ser mencionadas. Todos los conceptos desde las características tomadas hasta herramientas de los artículos principales mencionados que han servido para comprender los conceptos

fundamentales y que han sido tomadas como referencia en el trabajo desarrollado en esta tesis.

4

BASES DE DATOS EN PROTEÓMICA: INTERACCIONES PROTEÍNA-PROTEÍNA

4.1 Introducción

Como se ha podido observar a lo largo de los primeros capítulos de este trabajo, el volumen que ocupa toda esta información, incluso a veces es inmanejable, haciendo necesario la intervención de algún sistema que relacione toda esa información, para que no sea un conjunto de datos sin sentido.

Con este propósito surgen los denominados LIMS (Laboratory Information Management System), que son sistemas informáticos basados en bases de datos, generalmente relacionales, que como su nombre indica relaciona todo el flujo de información y control que se establece en cualquier laboratorio para cualquier experimento que se realice [99].

Muchos de los motores de búsqueda ponen a disposición del usuario, descarga directa de los datos consultados o de las bases de datos (ya sea totales o parcial). En General, el formato de descarga puede ser en texto plano dividido en columnas (TAB, es decir, tabulado) o en formato XML, el más común es PSI-MI 2.5 (también suele estar disponible la versión anterior 1.1) para la interacción de proteínas. Para más información acerca del formato PSI-MI consulta:

http://psidev.sourceforge.net/mi/xml/doc/user/

4.1.1 Interacciones de Proteínas

Actualmente existen una gran variedad de bases de datos de interacción de proteínas. Muchas de ellas como MINT [38], MIPS [160], BIND [9] recogen interacciones de proteínas. PREDICTOME [141] y STRING [143] recogen enlaces funcionales entre proteínas, derivadas de conjuntos de doble híbrido (ver apéndice D métodos de doble híbrido Y2H), eventos de fusión de dominios, historia filogenética y proximidad de

genes. En contraste con otras bases de datos, GRID es una compilación de BIND[9], MIPS [160] y otros datasets, como la base de datos DIP, que nos suministra conjuntos de datos de interacciones proteína-proteína de la levadura -S. cerevisiae- (que usaremos en este trabajo) comprobadas manualmente en laboratorio (curated). La mayoría de las entradas en DIP se obtienen de combinar, datos no solapados recolectados sistemáticamente mediante análisis doble-híbrido [95]. Sin embargo existen muchas otras bases de datos, a continuación se realizará una descripción de las bases de datos que se consideran más importantes o más usadas en el campo de la proteómica, y de las cuales, algunas de ellas han sido usadas para la realización de esta tesis.

4.2 Bases de datos

La información para la descripción de las bases de datos de estas secciones han sido extraídas fundamentalmente de las publicaciones y páginas webs asociadas a dichas bases, usando además como apoyo la documentación técnica aportada por Integromics S.L. [99] para dar un enfoque más práctico y completar información a este capítulo.

4.2.1 Gene Ontology (GO)

Gene Ontology (GO) es un proyecto del Gene Onlotogy Consortium [42] que parte de las bases de datos de tres organismos: Drosophila (mosca), MGI (Mouse Genome Informatics) [36] y SGD (Saccharomyces Genome Database) [59]. La meta del consorcio es generar un vocabulario estructurado, bien definido y controlado para descubrir los papeles que desempeñan los genes y las proteínas dentro de cualquier organismo. Con este objetivo se organizó la información recapitulado en torno a tres ontologías independientes [99]:

1. Proceso biológico en el que pueden estar presentes los genes y proteínas.

2. Función molecular o función propia de cada molécula.

3. Componente celular o localización celular.

Cada nodo de la ontología GO enlaza con la información almacenada en las principales bases de datos de genes y proteínas, como son, EMBL [203], SwissProt [24], MIPS [160] e InterPro [4].

4.2.1.1 Procesos biológicos

Hacen referencia al objetivo biológico al que contribuyen los genes o proteínas. Un proceso implica que actúe de forma ordenada una o varias funciones moleculares, por tanto, los procesos implican a veces transformaciones químicas o físicas, en el sentido que algo entra dentro de un proceso y algo distinto es lo que sale. Hay que tener en cuenta la jerarquía de los términos, de alto nivel como “crecimiento celular” y de bajo nivel como “biosíntesis cAMP” [99].

4.2.1.2 Función molecular

Es definida como una actividad bioquímica (incluyendo enlaces específicos a estructuras) de los productos de los genes. Esta definición se aplica también a la capacidad de los productos de los genes potenciales. Esto describe sólo cuando una actividad se ha llevado a cabo sin especificar dónde o cuándo el evento ha ocurrido. Los términos que están relacionados son enzimas o ligandos [99].

4.2.1.3 Componente celular

Se refiere al lugar en la célula donde el producto del gen es activo. Como ocurre en otras ontologías, no se pueden aplicar todos los términos a todos los organismos. El componente celular incluye términos con ribosoma o proteosoma, que especifica lugares en donde numerosos productos génicos se pueden encontrar [99].

4.2.1.4 Productos de genes y términos GO

El proceso biológico, la función molecular y componente celular son atribuidas a los genes, productos de los genes (proteínas). Cada uno de estos puede ser asignado independientemente y, de hecho el proceso biológico, la función molecular y la localización celular representa atributos independientes que pueden ser suficientes en muchos casos para la anotación de los datos de expresión genética. Las relaciones entre productos de los genes y el proceso biológico, la función molecular y la localización celular son una de las principales, conociendo que una simple proteína puede realizar su función en distintos procesos, contiene dominios (unidades funcionales de las proteínas) que pueden llevarse a cabo en distintas funciones moleculares y participar en múltiples alternativas interacciones con otras proteínas, orgánulos o localizaciones en la célula [99].

Go se ha diseñado para células eucariotas genéricas; por tanto, órganos especializados o partes del cuerpo no se han incluido. Los términos GO son nodos conectados dentro de una red, de manera, que las conexiones entre padres e hijos son conocidas y forman grafos acíclicos directos. Aunque las ontologías son dinámicas, en el sentido que su existencia con una red va cambiando según la información se va acumulando, GO, como tal no evoluciona de forma autónoma [99].

Los términos de GO son aplicados en la anotación de productos de genes en bases de datos biológicas. Las anotaciones GO son asociaciones hechas entre productos de genes y los términos GO que las describen. Un producto gen es un producto ARN o de proteína codificado por un gen. Porque cada gen único puede codificar diferentes productos con atributos muy diferentes, GO recomienda asociar los términos GO con objetos de bases de datos que representan productos de genes antes que un gen. Si los identificadores no están disponibles para distinguir productos de genes individuales, términos GO estar asociado con un identificador para un gen. Un objeto gen están asociados con todos los términos GO aplicables a cualquiera de sus productos.

Los identificadores GO son puramente identificadores, no codifican ninguna información sobre un término o su posición relativa a otros términos en el árbol. Las ontologías son "especificaciones de vocabulario relacional". En otras palabras, son conjuntos de términos definidos en orden como si fuera un diccionario, pero los términos están dados en relaciones jerárquicas de uno a otro. Los términos de un vocabulario dado son probablemente restringidos a un campo particular o un dominio, y en el caso de GO, los términos son todos biológicos.

Cada anotación debe ser atribuida a una fuente, la que la que refiere en la literatura, otra base de datos o un análisis computacional. La anotación debe indicar qué clase de evidencia es encontrada en la cita que apoya la asociación entre el producto de gen y el término GO. Un vocabulario simple controlado es usado para grabar evidencias; y los códigos de evidencias son simplemente códigos de 3 letras usado para significar el tipo de evidencia citada.

Los términos GO pueden estar enlazados por 5 tipos de relaciones: is_a, part_of, regulates, positively_regulates and negatively_regulates.

Para realizar las búsquedas usar el motor AMIGO; GO como otras bases de datos, posee una implementación en XML.

Para más información visitar:
http://www.geneontology.org/

Introducción:
http://www.geneontology.org/GO.doc.shtml

Motor de búsqueda:
http://amigo.geneontology.org/cgi-bin/amigo/go.cgi

Descargas:
http://www.geneontology.org/GO.downloads.shtml

Ejemplo de GOA:

```
1 UniProtKB        P38903   2A5D_YEAST      contributes_to  GO:0004722      PMID:10329624    TAS
            F        RTS1, SCS1, YOR014W, OR26.04: Serine/threonine-protein phosphatase 2A 56 kDa
      regulatory subunit delta isoform           protein taxon:4932      20060105         SGD

3 UniProtKB        P38903   2A5D_YEAST              GO:0000159      GOA:interpro|GO_REF:0000002      IEA
          InterPro:IPR002554      C       RTS1, SCS1, YOR014W, OR26.04: Serine/threonine-protein
      phosphatase 2A 56 kDa regulatory subunit delta isoform           protein taxon:4932
      20080521         UniProtKB

5 UniProtKB        P38903   2A5D_YEAST              GO:0000159      PMID:10329624    TAS             C
           RTS1, SCS1, YOR014W, OR26.04: Serine/threonine-protein phosphatase 2A 56 kDa regulatory
      subunit delta isoform           protein taxon:4932      20010118         SGD

7 UniProtKB        P38903   2A5D_YEAST              GO:0000780      PMID:16541024    IDA             C
           RTS1, SCS1, YOR014W, OR26.04: Serine/threonine-protein phosphatase 2A 56 kDa regulatory
      subunit delta isoform           protein taxon:4932      20060525         SGD

9 UniProtKB        P38903   2A5D_YEAST              GO:0005515      PMID:11283351    IPI      UniProtKB:
      P47135         F       RTS1, SCS1, YOR014W, OR26.04: Serine/threonine-protein
```

Se observa por tanto, la relación UniProt [43] de sus AC (números de acceso) con sus ID (identificadores) a los términos GO (GO:number) entre otros. Las referencias (GO_REF) señala un conjunto de "abstracts" que pueden ser citados en las anotaciones GO y ficheros de anotación. Para más información visitar GOA:
.http://www.ebi.ac.uk/GOA/
Documentación GOA y descripción de ficheros de datos:
http://www.ebi.ac.uk/GOA/goaHelp.html
Referencias GO:
http://www.geneontology.org/doc/GO.references

Ejemplo de GO:

```
1 SGD     S000007287      15S_RRNA              GO:0003735      SGD_REF:S000073642|PMID:6261980 ISS
                F       Ribosomal RNA of the small mitochondrial ribosomal subunit      15S_rRNA
      |15S_RRNA_2     gene    taxon:4932      20030723        SGD

3 SGD     S000007287      15S_RRNA              GO:0005763      SGD_REF:S000073642|PMID:6261980 ISS
                C       Ribosomal RNA of the small mitochondrial ribosomal subunit      15S_rRNA
      |15S_RRNA_2     gene    taxon:4932      20040202        SGD

5 SGD     S000007287      15S_RRNA              GO:0006412      SGD_REF:S000073643|PMID:6280192 IGI
                P       Ribosomal RNA of the small mitochondrial ribosomal subunit      15S_rRNA
      |15S_RRNA_2     gene    taxon:4932      20060630        SGD

7 SGD     S000007287      15S_RRNA              GO:0042255      SGD_REF:S000051605|PMID:2167435 IGI
                P       Ribosomal RNA of the small mitochondrial ribosomal subunit      15S_rRNA
      |15S_RRNA_2     gene    taxon:4932      20030723        SGD
```

4.2.2 DIP

DIP (Database of Interacting Proteins) [186] cataloga experimentalmente determinadas interacciones entre proteínas. Combina la información de una variedad de recursos para crear un único conjunto consistente de interacciones proteína-proteína. Los datos almacenados en la base de datos DIP son comprobados empíricamente, manualmente por expertos y también automáticamente usando enfoques computacionales que utilizan el conocimiento de las redes de interacción proteína-proteína extraídas de los datos de subconjuntos principales (core) más fiables de DIP.

Es posible descargase un subconjunto CORE (el más fiable) de la levadura: contiene los pares de proteínas identificados en el momento del inicio del crecimiento de la levadura (budding yeast). También se dispone de un plugin para un software de visualización de redes de interacción de proteínas (cytospace http://www.cytoscape.org/ [195]). Necesita un registro para poder acceder a todas las funcionales del motor de búsqueda y para las descargas. Es posible descargase los datos en formato texto plano tabulado (TAB) o en formato XML (formato llamado por DIP XIN).

Los datos descargados se refieren a redes de interacción, los nodos (N) y enlaces (E) van unidos a la palabra DIP como prefijo. Por ejemplo:

```
DIP:2551N AAC1 YMR056C DIP:1189N APG12 YBR217W DIP:11374E
```

Web:
http://dip.doe-mbi.ucla.edu/dip

4.2.3 Pfam

Es una base de datos de dominios y familias de proteínas comunes, que incluye alineamientos y modelos ocultos de Markov (HMM) [107], y valora el contexto y dependencias posicionales. Proporciona un punto de partida tanto para secuencias de ácidos nucleicos como de proteínas [64]. Pfam ofrece una colección fiable de alineamientos múltiples de familias de proteínas y un perfil de modelos ocultos de Markov. Se divide en tipos de familias: Pfam-A y Pfam-B. Pfam-B suministra clúster de alineamientos de secuencias de SWISSPROT y TrEMBL [24] generado por Prodom [44, 192] que no han sido modelados en Pfam-A [64, 96].

A lo largo de este trabajo de tesis, con objeto de esclarecer la arquitectura de dominios de Pfam, se tomó el compendio "swisspfam", una compilación de estructuras de dominio de SWISSPROT y TrEMBL [24] según PFAM.

Su web es:
http://pfam.sanger.ac.uk/

El ftp donde descargar swisspfam
ftp://ftp.sanger.ac.uk/pub/databases/Pfam/current_release

Ayuda:
http://pfam.sanger.ac.uk/help

Este fichero "swisspfam" relaciona los Identificados (ID) y números de acceso a la base de datos (AC) de las proteínas de UNITPROT con sus dominios, y éstos dominios aparecen con sus IDs y ACs de PFAM. Ejemplo del fichero:

```
1 >2A5D_YEAST       |=============================================================| P38903.2  757 a.a.
  B56               1                  ------------------------      (197) PF01603.11 Protein
      phosphatase 2A regulatory B subunit (B56 family)  276-713
3 Pfam-B_64176      1          --------                                (3) PB064176    154-275
  >2AAA_YEAST       |=============================================================| P31383.2  635 a.a.
5 HEAT              12       --- --   --- -- -- --- -- -- -- --   --- --  (6644) PF02985.13 HEAT repeat
      71-107 110-146 188-224 234-269 275-311 315-351 355-391 394-430 433-469 472-508 552-588 597-633
  Pfam-B_2473       1   ---                                             (39) PB002473    30-70
7 Pfam-B_2601       1            --                                     (37) PB002601    147-169
```

Hay que tener especial cuidado porque en ciertas bases de datos, y en versiones anteriores, no aparecen los ACs de los dominios y de las proteínas con número sufijo (p.ej. P38903 en lugar de P38903.2 ó PF02985 en lugar de PF02985.13).

Manual de usuario y descripción de los ACs e IDs:
ftp://ftp.sanger.ac.uk/pub/databases/Pfam/current_release/userman.txt

4.2.4 UnitProt

Universal Protein Resource (UniProt) [43] es un recurso global para datos de secuencia de proteínas y anotación. Las bases de datos Unitpro es la UniProt Knowledgebase (UniProtKB), y UniProt Reference Clusters (UniRef), y la UniProt Archive (UniParc). La UniProt Metagenomic and Environmental Sequences (UniMES) database es un

repositorio especificamente desarrollado para metagenómica y datos medioambientales [43].

UniProt es una colaboración entre European Bioinformatics Institute (EBI), el Swiss Institute of Bioinformatics (SIB) y el Protein Information Resource (PIR). Entre los tres institutos cerca de 150 personas están dedicadas en diferentes tareas como comprobación experimental de la base de datos (database curation), desarrollo de software y soporte.

En SWISSPROT [24], el fichero relaciona identificadores y números de acceso de diferentes bases de datos, nombres de genes, ORFs que están relacionados con la proteína, incluyendo en qué cromosoma está el gen que tiene producto esa proteína. En la cabecera del fichero está la descripción del fichero, en nuestro caso, hemos tomado el fichero "yeast.txt" ya que en este trabajo tomaremos la levadura como organismo de estudio; he aquí el ejemplo:

Gene designations	Swiss-Prot AC	Swiss-Prot Entry	SGD X-ref.	Sequence length	3D	CH
a2; YCR096C; YCR96C	P01367	MAT2_YEAST	S0000692	210	(3)	3
AAC1; YMR056C; YM9796.09C	P04710	ADT1_YEAST	S0004660	309		13
AAC2; PET9; YBL030C; YBL0421	P18239	ADT2_YEAST	S0000126	318		2

Para más información acerca de los formatos de números de acceso (AC), identificadores (ID) y otros, por favor, leer el siguiente manual:
http://www.uniprot.org/docs/userman.htm

Para la descargar de archivos visitar los ftps:

ftp://ftp.ebi.ac.uk/pub/databases/swissprot/release/
ftp://ftp.ebi.ac.uk/pub/databases/trembl/

La web es:
http://www.uniprot.org

UnitprotKB está formada por dos secciones: SwissProt que manualmente anotada y revisada, y por TrEMBL que es automáticamente anotada y no está revisada.

4.2.5 MIPS

El Institute of Bioinformatics and Systems Biology (IBI) es parte de HZM - Centro de investigación alemán para la salud medioambiental y los anfitriones del Munich Information Center for Protein Sequences (MIPS) [160]. Su foco principal es la bioinformática orientada al genoma, en particular, los análisis sistemáticos de la información del genoma incluyendo el desarrollo y la aplicación de métodos bioinformáticos en una

anotación de genoma, análisis de expresión y proteómica. MIPS soporta y mantiene un conjunto de base de datos genéricas como el análisis comparativo sistemático de genomas microbiales, fúngico (hongos) y de plantas. Dispone catálogos de proteínas según función, complejo, fenotipo e incluso componente celular.

Dispone de un compendio de dataset de interacciones actuales o de mayor uso últimamente de diferentes organismos:
http://mips.gsf.de/proj/yeast/tables/interaction/
Web:
http://mips.gsf.de/
FTP:
ftp://ftpmips.gsf.de/

4.2.6 Diccionarios de secuencias

Se observa que es de vital importancia, una vez identificada una proteína, conocer qué posibles funciones puede realizar, lo que determina con qué puede interaccionar, ya sean otras proteínas o sustratos. Por tanto, tiene especial incidente en los campos bromatológico (referente a la manipulación y conservación de los alimentos), patológico, farmacológico o biotecnológico, lo que supone al menos contrastar con las moléculas más interesantes o contra todos los compuestos comerciales o contra todos los posibles. No hay que olvidar un aspecto importante en la evolución: el que sea al azar implica que puede discurrir en cualquier dirección y así tenemos múltiples casos de evolución convergente (genes distintos con funciones similares) y divergente (secuencias similares con funciones distintas).

La secuencia determina la estructura y ésta la función, por tanto, funciones similares deberían proceder de secuencias similares ¿o no? por ello se comparan secuencias localizando las regiones conservadas y consiguiendo con ello, asociar esas "palabras" o "frases" con su significado. Podemos recopilar estas palabras y frases en bases de datos para facilitar su uso y consulta. El mecanismo básico del análisis de secuencias consiste en comparar secuencias: seleccionamos las secuencias de interés y producimos un alineamiento múltiple de las mimas. Dependiendo de lo que se busque se puede detectar qué secuencia se asocia a una función. Una vez detectadas las regiones se procede a cribarlas comprobando como son de específicas. Las que pasan de forma satisfactoria se recopilan en una base de datos.

A veces, la variabilidad es muy grande y por ello se usa para representar una secuencia una matriz de frecuencias indicando para cada posición los aminoácidos que aparecen y la frecuencia con la que suele aparecer cada uno. Es obvio, que la función no siempre está asociada a una secuencia claramente localizada: en muchos casos depende de una conformación determinada o del contexto que rodea a la secuencia, o viceversa, la secuencia correspondiente a la parte funcional depende del entorno. El siguiente paso consiste en incluir esta información contextual.

Para representar las dependencias posicionales se utilizan "modelos ocultos de

Markov" (Hidden Markov Models, HMM) [107] y redes neuronales. Estas técnicas permiten detectar y capturar dependencias posicionales complejas no evidentes de una forma precisa y automatizable.

4.2.6.1 PROSITE

Prosite [194] es un diccionario de motivos (patrones de secuencia) conservados derivados por comparación de secuencias guiada por datos experimentales. También se han incorporado reglas y matrices. Cada entrada de la base de datos va acompañada de una amplia cantidad de información detallada que incluye descripciones, referencias bibliografías e índices de fiabilidad de motivo, regla o matriz.

Web:
http://expasy.org/prosite/

4.2.7 Otras bases de datos

Como ya hemos comentado, existe una cantidad de bases de datos desorbitada, y en este trabajo no sólo se pretende presentar las bases de datos que se usarán en este tesis, por ello caben destacar algunas bases de datos adicionales de cierta importancia, y que son útiles en el desarrollo de este trabajo y de posibles trabajos futuros.

4.2.7.1 HUPO-PSI: Human Protome Organisation - Proteomics Standards initiative

HUPO-PSI [87] se creó con la intención de definir estándares para la representación de los datos par facilitar la comparación, el intercambio y la verificación de datos. Posee dos áreas vitales: espectrometría de masas (ver apéndice. D) e interacciones proteína-proteína. Hay que añadirle una área más, de propósito general, que pretende dar un esquema general de la proteómica. Las bases del esquema general para proteómica son las que se presentan:

- Conjunto mínimo de conceptos: Existen unos mínimos datos necesarios para poder representar un experimento. Estos componen el conjunto mínimo de datos que tienen que estar presentes en todos los resultados que aparecen en los informes.
- XML: La base de este esquema es el lenguaje estándar XML.
- Ontologías: Establece relaciones entre distintos conceptos para describir relaciones semánticas entre ellos. En muchas ocasiones, sobre todo cuando se trabaja con equipos informáticos, se tienen una serie de conceptos que aún significando lo mismo no se puede especificar que existe una similitud semántica entre ellos.
- Otras actividades: En donde se coordinarán la generación de comparaciones entre los distintos motores de búsqueda.

Para más información visitar:
http://www.psidev.info/

4.2.7.2 ProDom

ProDom [44, 192] se trata de un conjunto exhaustivos de dominios identificados por familias de proteínas. Usa información estructura (dominios 3D) pero usa secuencias.

Su web:
http://prodom.prabi.fr/prodom/current/html/home.php

4.2.7.3 InterPro

InterPro [4] es una base de datos de familias de proteínas, dominios y sitios funcionales que surgió como un proyecto para integrar distintas bases de datos (o diccionarios) existentes en un recurso, único. Actualmente reúne información de bases de datos como Prosite [194], ProDom [44, 192],...

Su web es:
http://www.ebi.ac.uk/interpro/

4.2.7.4 IntAct

IntAct [5] es una base de datos libremente disponible que contiene interacciones de proteínas además de herramientas para el análisis de los datos de interacción de proteínas.

Su web es:
http://www.ebi.ac.uk/intact/site/index.jsf

También tiene la opción de bajarse los dataset completos en formato plano TAB (tabulado), introducción el identificador (referencia) de la publicación en PUBMED. Para ello, visitar:
http://www.ebi.ac.uk/intact/search/do/search

4.2.7.5 iHOP y PubMed

Cuando el investigador quiere averiguar información sobre la proteína, normalmente realiza una búsqueda en PubMed. PubMed es un motor de búsqueda de libre acceso a la base de datos MEDLINE de citaciones y "abstracts" de artículos de investigación biomédica. Ofrecido por la Biblioteca Nacional de Medicina de los Estados Unidos. MEDLINE tiene alrededor de 4.800 revistas publicadas en Estados Unidos y en más de 70 países de todo el mundo desde 1966 hasta la actualidad.

iHOP (Information Hyperlinked over Proteins) [93] es una base de datos en la que se puede identificar las interacciones recogidas por PubMed de una proteína de interés pero que tiene un uso más eficiente que las búsquedas que ofrece PubMed.

Web de PubMed:
http://www.ncbi.nlm.nih.gov/pubmed/

Web de iHOP:
http://www.ihop-net.org/UniPub/iHOP/

4.2.7.6 BIND

La base de datos BIND [9] ha pasado a ser BOUND (Biomolecular Object Network Databank), y necesita realizar un registro para poder usarla. Ciertas partes del portal son de pago, pero podemos hacer consultas. Bind sigue existiendo como tal dentro del BOUND.

Dispone incluso de una serie de bibliotecas en perl, c/c++ para la lectura de su base de datos.

Además dispone de un algoritmo (BIND PICKS) que nos da una medida de confianza. Por tanto, la puntuación que nos proporciona BIND PICKS representa una medida de confianza numérica para las interacciones de *Saccharomyces cerverisiae* en BIND. BIND Sugieren que las puntuaciones de interés 1 constituyen interacicones verdaderas adecuada para una investigación y análisis en profundidad. Un subconjunto de interacciones positivas y negativas de *Saccharomyces cerevisiae* fueron tomadas para entrenar un Support Vector Machine (SVM) usando un conjunto de características para generar BIND Protein Interaction Confidence Kernel Scores (PICKS). Las características incluyen número de publicaciones comprobadas (curated) y publicaciones en las que se le ha aplicado (minería de datos), existencia de interacciones homólogas, anotaciones de GO (Gene Ontology) comunes y relacionadas, perfiles de fenotípicos y composición de dominios. El sistema BIND PICKS supera los otros métodos de puntuación y clasifica 36% de las interacciones HTP (High-Throughtput Procedures) de forma fiable.

Web:
http://bond.unleashedinformatics.com

4.2.7.7 InParanoid y HintDB

La base de datos InParanoid [19] proporciona información acerca de secuencias ortólogas (genes homólogos de distintas especies que descienden de un gen único de un ancestro común) para *S. cerevisiae* y el conjunto de proteínas completas de *H. sapiens, D. musculus, C. elegans and A. thaliana.*

Web:
http://inparanoid.sbc.su.se/cgi-bin/index.cgi

En el caso de Hintdb (Homologous Interactions Database) [163] proporciona una colección de interacciones proteína-proteína determinadas experimentalmente y que se agrupan en base de la homología de secuencia. A día de hoy, almacena 176983 interacciones de 9 especies (*S.cerevisiae, H.sapiens, D.melanogaster, E.coli, S.pombe, C.elegans, A.thaliana, M.musculus, R.norvegicus*) de las cuales 87026 tiene homólogos.

Web:
http://hintdb.hgc.jp/hint/

4.2.7.8 PDB

El Protein Data Bank (PDB) (Banco de Datos de Proteínas) [20] es una base de datos que almacena información 3-D (tridiminesional) sobre estructura de las proteínas y ácidos nucleicos. Estos datos, generalmente obtenidos por Cristalografía de rayos X o Resonancia Magnética Nuclear, son enviados por biólogos y bioquímicos de todo el mundo. Están bajo el dominio público y pueden ser usados libremente.

La RCSB PDB (Protein Data Bank) suministra una variedad de herramientas y recursos para el estudio de estructuras de macromoléculas biológicas y sus relaciones con secuencia, función y enfermedad.

Web de PDB:
http://www.rcsb.org

4.2.7.9 3DID

La base de datos de interacción de dominios 3D (3D Interaction Domains -3did-) [201] es una colección de interacciones dominio-dominio en proteínas para las cuales las estructuras tridimensionales de alta resolución son conocidas. 3did explota la información estructural para suministrar detalles moleculares críticos necesarios para comprender cómo la interacción tiene lugar. También ofrece una visión general de cómo de similares son las interacciones en la estructura entre miembros diferentes de la misma familia de proteínas. La base de datos contiene anotaciones funcionales basadas en GO e interacciones entre proteínas de la levadura (en inglés yeast) procedentes de estudios de interacción de gran escala.

Web de 3did:
http://gatealoy.pcb.ub.es/3did/

4.2.7.10 BioGrid

BioGrid [199] es una base de datos de interacciones de proteína-proteína. Permite búsqueda por identificador de SWISSPROT. Es útil para formar datasets (conjunto de datos). Los resultados pueden mostrarse de diversos modos, permitiendo darnos un índice (score o puntuación) siendo datos fiables lo que toman valores 1. o permite la opción "ALL DATA" que nos muestra interacciones proteícas y genéticas. Permite la descarga en varios formatos: TAB (texto plano separado en columnas), PSI-MI (formato XML) en versiones 1.1 y 2.5. Además dispone de un software de visualización (Osprey) de redes de interaciones.

Web:
http://www.thebiogrid.org/

4.2.7.11 STRING

STRING [143] es una base de datos de interacciones proteína-proteína conocida y predichas. Las interacciones incluyen asociaciones directas (físicas) e indirectas

(funcionales). Se basan en 4 fuentes: contexto genómico, experimentos HT (High-throughput), co-expresión (conservada) y conocimientos previo (de publicaciones). Integra cuantitativamente datos de interacción de estas fuentes para un gran número de organismos, y transfiere información entre estos organismos cuando es aplicable.

La base de datos STRING dispone de la descarga de toda la base de datos en formato SQL (lenguaje de consulta de base de datos) para el sistema de gestión de base de datos relacional orientada a objetos PostgreSQL (http://www.postgresql.org/). La descarga no es gratuita, es de pago, sin embargo la versión anterior está disponible para fines académicos realizando un registro (enviándolo por fax).

Web:
http://string.embl.de/

4.2.7.12 Domain Fusion Database

La finalidad de esta base de datos (Domain Fusion Database [208]) es repartir información sobre los eventos de fusión de dominios detectados en la levadura y en la secuencias de humano en la base de datos SWISSPROT [24] usando una técnica de álgebra relacional. Los eventos de fusión de dominios son indicados con sus posibles uniones funcionales. (Más información en el glosario, Proteínas de fusión)

Web:
http://calcium.uhnres.utoronto.ca/pi/flash.htm

Los identificadores (ID) usado son los de SWISSPROT [24] para las proteínas y los de PFAM [64]para los dominos.

4.3 Conclusiones

Con este capítulo se ha proporcionado un breve compendio describiendo las bases de datos más importantes del ámbito de la proteómica, si bien existen otras muchas más. Con ello, se pretende dar una idea de los formatos de las bases de datos, aportando también información técnica y de la variada información de la que se puede disponer referente a los genes y sus productos. De todas ellas algunas serán usadas a lo largo del desarrollo de la tesis.

5

NEW METHOD FOR PREDICTION OF PROTEIN-PROTEIN INTERACTIONS IN YEAST USING GENOMICS/PROTEOMICS INFORMATION AND FEATURE SELECTION

5.1 Introduction

As commented in the previous chapters, protein-protein interactions (PPIs) play a great importance role in almost any biological function carried out within the cell [56][91]. In fact, an enormous effort has already been made to study biological protein networks in order to understand the main cell mechanisms [181] [96] [80]. The development of new technologies have improved the experimental techniques for detecting PPI [202]. However, computational approaches have been implemented for predicting PPIs because of cost and time requirements associated with the experimental techniques [80].

Although different computational methods have been applied in PPI prediction [106] [163] [116]. In this chapter, a predictor of protein-protein interactions in *Saccaromices cerevisiae* (Yeast) based on Support Vector Machines (SVM) that can be used as a validator is proposed. A high-quality GSP is used to build the SVM classifiers and a GSN is generated from data published by Saeed et al. [185]. Furthermore, a total of 26 genomic/proteomic features are extracted from well-known databases and datasets and the similarity measure proposed in this chapter is used to calculate some of these features. The most relevant features are subsequently obtained by means of the proposed feature selection method in this chapter derived from the methods studied in Gilad-Bachrach et al. et [175, 72]. Finally, SVM predictors are constructed that return a confidence score in classification of each pair of proteins and it is obtained

results higher than 90% of specificity and sensibility in the prediction of PPIs. In order to check that the selected features calculated by the proposed similarity measures in this chapter help to improve the predictive power of SVM, a comparison was analysed between the three features based on the proposal by Patil et al. [162] (i.e, using the implementation of those features in this chapter) and the three highest weighted features obtained with the proposal of this chapter. This study confirms that it is possible to use the proposed SVM predictors as validatory methodology to filter the PPI datasets and that the models also can return a general score that can be used as a confidence measure for each PPI. A more universal confidence scoring approach is preferable because it is free of biases resulting from the influence of a particular experimental setup [30].

The rest of the chapter is organized as follows. The section 5.2 explains how features were extracted from the genomic/proteomic databases, the section 5.3 describes the variable selection approach based on the well-known margin-based algorithms [175, 72] followed in this chapter, the section 5.4 describes which kernels for SVM were used and the proposed confidence score in PPI prediction problems and the section 5.5 describes the results obtained upon applying the presented approach. Finally, the conclusions that can be drawn from this study are set out in the section 5.6.

5.2 *Feature extraction and similarity measures using databases in bioinformatics*

In this section, the used databases are described, from which it was extracted the genomic and proteomic information for all yeast proteins used in this chapter. Then, 26 features are presented, they were selected from these databases, as a prior step to create the proposed model in this chapter. Although some of these databases are described in chapter 4, a briefly description is added to this section, and it is also included additional information and some new material as mRNA co-expression dataset from Jansen et al. [106] focused on developing the specific approach presented in this chapter. The databases used, all well-known in Bio-informatics, are the following ones:

- Gene Ontology Annotation (GOA) Database [37]: It provides high quality annotation of Gene Ontology (GO) [42] (version May 2008). The GO project was developed to provide controlled vocabularies for the annotation of molecular attributes in different model organisms. These vocabularies are classified in GOA into three structured ontologies, which are used to describe molecular function (F), biological process (P) and cellular component (C). Each ontology is organized as a directed acyclic graph.

- MIPS Comprehensive Yeast Genome Database (MIPS CYGD, version June 2008) [84]: It gathers information on molecular structure and functional network in yeast.

In this study all catalogues were considered: functional, complexes, phenotype, proteins and sub-cellular compartments.

- Homologous Interactions database (HINTdb, version 13 June 2007) [163]: It is a collection of protein-protein interactions and their homologous in one or more species. Using this database, it is included homology information for a given pair of proteins. Homology refers to any similarity between characteristics that is due to their shared ancestry.

- 3D Interacting Domains database (3did, version 25 May 2008) [201]: It is a collection of domain-domain interactions in proteins for which high-resolution three-dimensional structures are known in the Protein Data Bank (PDB) [20]. 3did exploits structural information to provide critical molecular details necessary for understanding how interactions occur. It also offers an overview of how similar in structure are interactions between different members of the same protein family. The database also contains Gene Ontology-based functional annotations and interactions between yeast proteins from large-scale interaction discovery studies.

- Swisspfam (version 22.0) [24] from Unitprot database [43]: It is a compilation of domain structures from SWISSPROT and TrEMBL [24] according to Pfam [64]. Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models.

- In Jansen et al. [106] used the Rosetta Compendium, a list of mRNA co-expression values which is considered in this analysis.

As it was stated above, one of the main ideas of the work in this chapter is to find an approach capable of integrating distinct protein features to design a reliable classifier of PPIs. Taking into account the importance of protein domains in predicting PPIs [54, 96, 162], it is more than justified the use of SwissPfam and 3did databases. Besides, GO data have been successfully used in previous papers supporting classification models [180] and the use of similarity measures [230]. The MIPS CYGD catalogues (functional, complexes, phenotype, proteins and sub-cellular compartments) provide very precise information about yeast and it is very well-known in proteomics [106, 163], making it a very good complement to the other proposed features. As previously mentioned in Introduction section, the "interlogs" concept helps to develop new approaches in proteomics, in fact homology information has already been used to predict, classify and build reliable PPI databases [221, 163, 185]. Therefore, paralogous data is included using the Hintdb database in this study. Moreover, due to the fact that interacting proteins are generally co-expressed, it was incorporated as a genomic feature, the information about expression profiles that can be obtained from [106, 163].

Using all this genomic and proteomic information, the next step was to extract a set of features that can be associated to all possible combinations of pairs of proteins, and

can be used to predict whether each and every one of these pairs of proteins are likely to interact. The key idea concerning feature extraction here consists of computing how many common terms are shared between both proteins in the pair in any given database.

In the case of querying databases, the gene ID was replaced by that of the protein. it is used the Uniprot database for mapping gene to protein. An example of the process of extracting data from the database is as follows:

- from the GOA database [37] The IDs of the GO terms associated to each protein are extracted. Depending on each feature, All ontologies together or each one separately (F,P,C) are considered.

- from the MIPS database [84] The IDs in the classification of each catalogue associated to each protein are considered separately.

- from the HINTdb homologous [163] The number of homologous for each pair of proteins are considered; the query is accessed by protein interaction A-B.

- from the 3did database [201] The Pfam domains [64] from the Swisspfam catalogue (version 22.0) [24] are extracted. These domains are associated to the proteins in a pair (A,B) and it is checked whether these domains are found in 3did. Then it is counted the number of domains founds for each A-B pair.

- from the Rosetta Compendium [106] mRNA co-expression values are extracted.

In spite of this plausible extraction process, these features in themselves, without being combined, do not provide enough information to decide whether any two given proteins are very likely to interact [106]. Thus, to improve the discriminative capabilities of the features, and by reinforcing the predictive power of models through a specific combination of features, the feature extraction process is extended by incorporating two new similarity measures (local and global). Therefore, the definition of these similarity measures would be:

Let A be the set of all terms associated in a specific database for protein *protA* and B the set of terms associated for protein *protB* in the same database. The local similarity measure for both proteins is defined as:

$$sim_{local} = \frac{\#(A \cap B)}{\#(A \cup B)} \tag{5.1}$$

where $\#(A \cap B)$ represents the number of common terms for a specific database between both proteins and $\#(A \cup B)$ represents the total number of all terms in the union of sets A and B.

Similarly, the global similarity measure takes into account the ratio of common terms shared by a given pair of proteins with respect to the sum of all terms in a database. It can be calculated as:

$$sim_{global} = \frac{\#(A \cap B)}{\#C} \tag{5.2}$$

where C represents the total number of terms in the entire database.

Once these local and global measures are defined, now the specific 26 features extracted from the selected databases that will represent every protein-protein interaction are described. It is also indicated between parenthesis the type of data (real or integer) and the position in this feature list. For the sake of clarity, this information is also represented in table 5.1:

- Common GO annotation terms between both proteins in the three ontologies (P,C,F) (1^{st} integer) and their local and global similarity measures (13^{th} real, 14^{th} real). Moreover it is considered each separately ontology (5^{th} P integer, 6^{th} C integer, 7^{th} F integer) and their respective local (16^{th} real, 17^{th} real, 18^{th} real) and global similarity measures (19^{th} real, 20^{th} real, 21^{st} real).

- Number of homologous between both proteins obtained from HintDB (2^{nd} integer).

- Common Pfam domains between both proteins, extracted from SwissPfam, which are found in the 3did database (3^{rd} integer) divided by the total Pfam domains that both proteins have (15^{th} real).

- mRNA co-expression value from Rosetta Compendium, extracted from Jansen et al. [106] (4^{th} real).

- Considering each MIPS catalogue separately, it is computed also the number of common terms (catalogue identifier) between both proteins (functional 8^{th} integer, complexes 9^{th} integer, proteins 10^{th} integer, phenotypes 11^{th} integer, sub-cellular compartments 12^{th} integer). Furthermore, their local similarity measures were considered (22^{nd} real, 23^{rd} real, 24^{th} real, 25^{th} real, 26^{th} real).

Table 5.1: Description of the 26 extracted features

Number	Description	Type
1^{st}	$\#(A_{GOA} \cap B_{GOA})$ from GOA DB taking 3 ontologies together (P,F,C)	integer
2^{nd}	Number of homologous for $(protA, ProtB)$ from HintDB	integer
3^{rd}	$\#[(A_{SPFAM} \cap 3DID) + (B_{SPFAM} \cap 3DID)]$, A and B are domains extracted form SwissPfam, 3DID is 3did database	integer
4^{th}	mRNA co-expression value extracted from Jansen et al. [106]	real
5^{th}	$\#(A_{GOA-P} \cap B_{GOA-P})$ from GOA DB taking Biological Process ontology	integer
6^{th}	$\#(A_{GOA-C} \cap B_{GOA-C})$ from GOA DB taking Cellular Compartment ontology	integer
7^{th}	$\#(A_{GOA-F} \cap B_{GOA-F})$ from GOA DB taking Molecular Function ontology	integer
8^{th}	$\#(A_{MIPS-F} \cap B_{MIPS-F})$ from functional MIPS catalogue identifiers	integer
9^{th}	$\#(A_{MIPS-C} \cap B_{MPIS-C})$ from complexes MIPS catalogue identifiers	integer
10^{th}	$\#(A_{MIPS-P} \cap B_{MIPS-P})$ from proteins MIPS catalogue identifiers	integer
11^{th}	$\#(A_{MPIS-FE} \cap B_{MPIS-FE})$ from phenotypes MIPS catalogue identifiers	integer
12^{th}	$\#(A_{MPIS-FCC} \cap B_{MIPS-FCC})$ from sub-cellular compartments MIPS catalogue identifiers	integer
13^{th}	Local similarity of 1^{st} feature	real
14^{th}	Global similarity of 1^{st} feature	real
15^{th}	$\#[((A_{SPFAM} \cap 3DID) + (B_{SPFAM} \cap 3DID))]/\#(A_{SPFAM} \cup B_{SPFAM})$	real
16^{th}	Local similarity of 5^{th} feature	real
17^{th}	Local similarity of 6^{th} feature	real
18^{th}	Local similarity of 7^{th} feature	real
19^{th}	Global similarity of 5^{th} feature	real
20^{th}	Global similarity of 6^{th} feature	real
21^{th}	Global similarity of 7^{th} feature	real
22^{th}	Local similarity of 8^{th} feature	real
23^{th}	Local similarity of 9^{th} feature	real
24^{th}	Local similarity of 10^{th} feature	real
25^{th}	Local similarity of $11th$ feature	real
26^{th}	Local similarity of 12^{th} feature	real

The symbol # indicates the number of elements in a set. See equations 5.1 and 5.2.

5.3 Emsemble Feature Selection Approach using margin based feature selection criterion methods

Feature selection is an essential pre-processing in problems such as classification or forecasting with machine learning, as a first phase to guarantee high accuracy and efficiency [196] as it was already commented in chapter 2. After the analysis of different methodologies for the PPI problem presented in this thesis, in this section a novel method is proposed based on 3 feature selection method suitable for the protein-protein interaction forecast: G-flip, Simba [71] and Relief [117] already explained in chapter 2. They use a margin based criterion to measure quality of sets of features. G-flip and Simba assign a score to sets of features according to the induced margin from an evaluation function. This G-flip and Simba evaluation function requires a utility function; the three functions are utilised in this chapter: linear, sigmoid and zero-one. Please remember that the linear utility function is a simple sum of margins. The Zero-one utility (not for Simba) equals 1 when the margin is positive and 0 otherwise. And finally, the sigmoid utility function is proportional to the leave-one-out (LOO) error. Relief is simple and efficient, returning weights that allows determining which attributes are relevant and to set an order among them. Given a dataset, Relief returns a ranking of features according to an importance weight. In order to obtain a subset including the most relevant features, the method selected the features whose weight was above the mean weight of all features.

The proposed feature selection method in this chapter consists of normalizing the weights between [0,1] of all previous methods results (Gflip, Simba and Relief). Subsequently, the final weight for each feature is calculated as a mean of all weights of all methods. The feature with a final weight above the mean are the features selected to perform the classification (see figure 5.1).

For the problem of PPI presented in this chapter, the weights of the 26 used features for yeast, are presented in the figure 5.2 for each of the six methods analysed, and the mean of all weights of all methods in 5.3

Finally, it is important to emphasize that the prediction of PPI is classification problem, the each point in a set (dataset) represents a pair of proteins they are classified in two classes or labels: interacting or non-interacting proteins.

5.4 Support Vector Machines and proposed confidence score for this problem

Support Vector Machines (SVM) are a classification and regression paradigm first developed by Vapnik [46, 88], the SVM paradigm already described in section 2.2.1.

In the work for this chapter, it was adopted a SVM approach to construct the classifiers using linear and RBF kernels. In order to perform the predictors, data set was divided randomly in two groups: training and test. Here it is proposed a

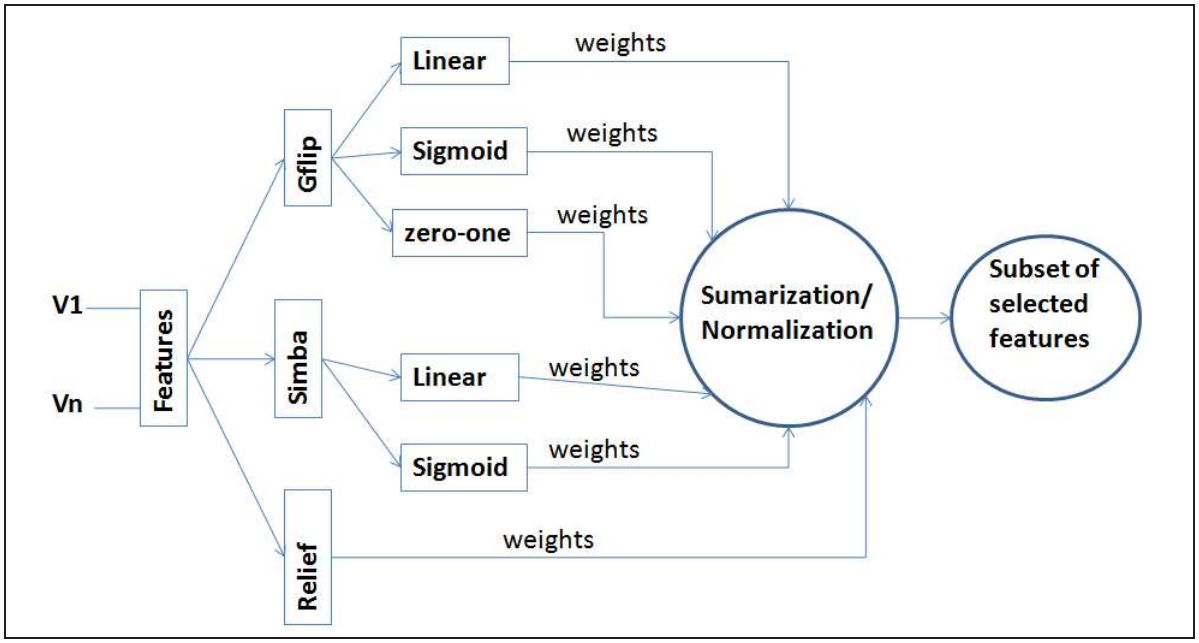


Fig. 5.1: The proposed feature selection method in this chapter

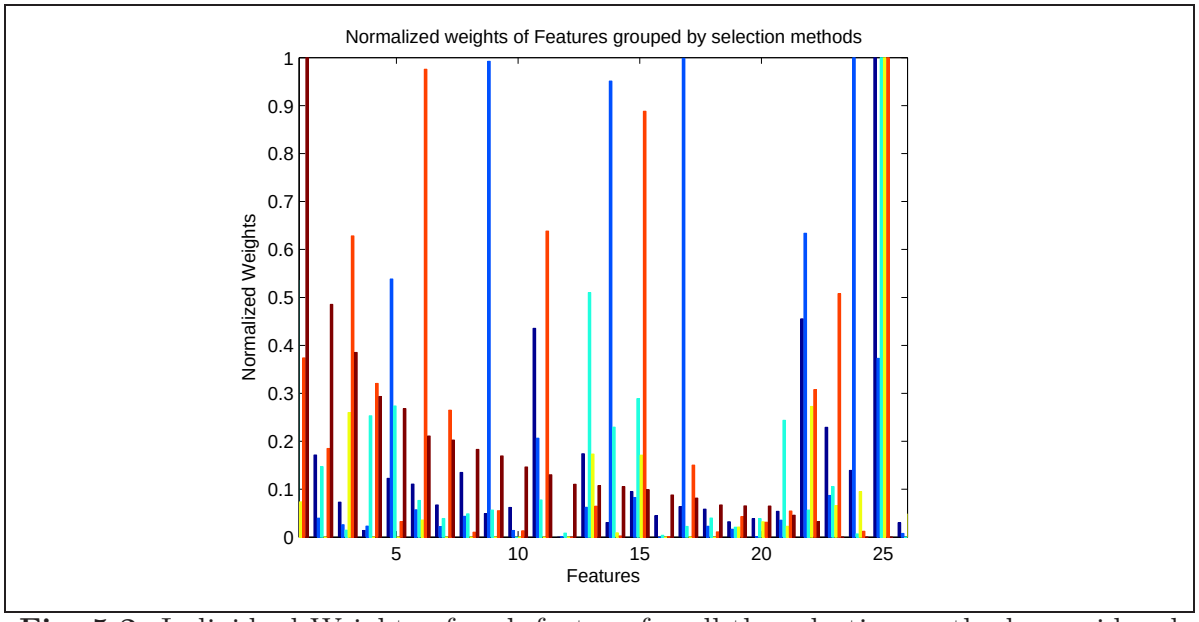


Fig. 5.2: Individual Weights of each feature for all the selection methods considered

score in prediction which is the difference of probabilities in absolute value returned by SVM for each pair of proteins, that could be used as measure of confidence. SVM classifies the pairs returning two probability values that represent the chance to be an interacting or non-interacting pair. A pair of proteins will be classified in positive or

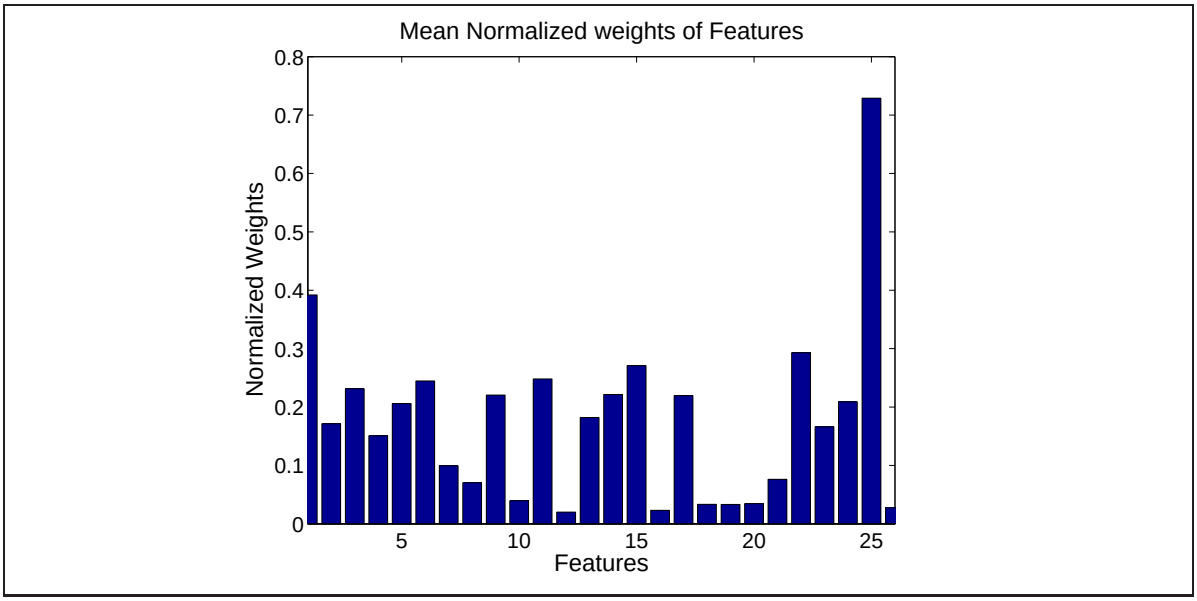


Fig. 5.3: Mean Weights of the 26 features presented

negative interaction according to best probability. These probabilities are obtained by the particularisation of multi-class classification methodology presented in [229] for the problem of PPI (binary classification). In a general problem, given the observation $\mathbf{x}$ and the class label y, it is assumed that the estimated pairwise class probabilities $\mu_{ij} = P(y = i|\ y = i\ or j,\ \mathbf{x})$ are available. Following the setting of the one-against-one approach for the general problem of multi-class problem with k classes, first, the pairwise class probabilities are estimated by r_{ij} with

$$r_{ij} \approx P(y = i|\ y = i\ or j,\ \mathbf{x}) \approx \frac{1}{1 + e^{A\widehat{f}+B}} \tag{5.3}$$

where A and B are estimated by minimising the negative log-likelihood function using known training data, and $\widehat{f}$ are their decision values for these training data. In [239] it is recall that SVM decision can be easily clustered at ± 1, making the estimate probability in 5.3 inaccurate. Therefore, and five-fold cross-validation to obtain decision values has been carried out in the experimental results. The next step is obtaining p_i from these r_{ij}, solving the following optimisation problem presented in [229]:

$$\min_{p} \frac{1}{2} \sum_{i=1}^{k} \sum_{j:j\neq i} (r_{ji}p_i - r_{ij}p_j)^2 \quad subject\ to\ \sum_{i=1}^{k} p_i = 1,\ p_i \geq 0, \forall i \tag{5.4}$$

5.5 Results

Following the steps of the proposed experiment, the extraction of all features were carried out from the given problem and its set of examples, and applied the proposed

feature selection method to obtain the subset of relevant features. The dataset in this section has been obtained from Saeed et al. [185], all methods were applied to this dataset. They provide a set of 4 million negative pairs of proteins and a set of 4809 interacting pairs of proteins for yeast. All positive examples (4809) have been used and a random subset of 4895 negative examples has also been used, giving a total of 9704 pairs. This random subset has been obtained according with other papers presented in the field of bioinformatics for PPI [109, 185, 234].

Due to computational needs it was used randomly 70% of data for feature selection methods. Simba and Gflip were executed over 10 iterations applying all their utility functions for evaluation function. Finally, features that were selected 10 times in Gflip were chosen. On the other hand, in Simba the features were selected according to mean weights in all executions. Relief is deterministic. The proposed method in this chapter calculated a mean weight per feature derived from the weights of these methods. Table 5.2 shows the selected features by method, and figure 5.3 shows the final features weights for the proposed method.

Table 5.2: Table of selected features by different selection methods

Method	Selected Features
Gflip Linear	1 2 3 5 6 7 8 10 11 13 15 16 17 18 19 22 23 24 25
Gflip Sigmoid	1 2 3 4 5 6 7 8 9 10 11 13 15 16 17 18 20 21 22 23 24 25
Gflip zero-one	1 2 25
Relief	25 22 11 1 13 5 23 2 24
Simba Linear	3 15 22 25
Simba Sigmoid	1 3 4 6 11 15 22 23 25
The selected features by the proposed feature selection method	1 3 5 6 9 11 13 14 15 17 22 24 25

In the next step, the data again was randomly divided in two groups: 70% pairs to train the SVMs and so getting the classifiers, and 30% pairs to test the models and to evaluate them. Two types of kernels were used for SVM classification: linear and RBF (Radial-Basis Function) [178]. A single combination of calculated parameters γ and C for RBF Kernel and C for linear kernel was employed. The performance of the test set was evaluated by sensitivity and specificity already described in chapter 3 but a small description for this problem is written in the next lines. Sensitivity of a test can be defined as the aptitude to identify a true positive in a data set. Mathematically, sensitivity is represented by equation 5.5:

$$Sensitivity = \frac{TP}{TP + FN} \tag{5.5}$$

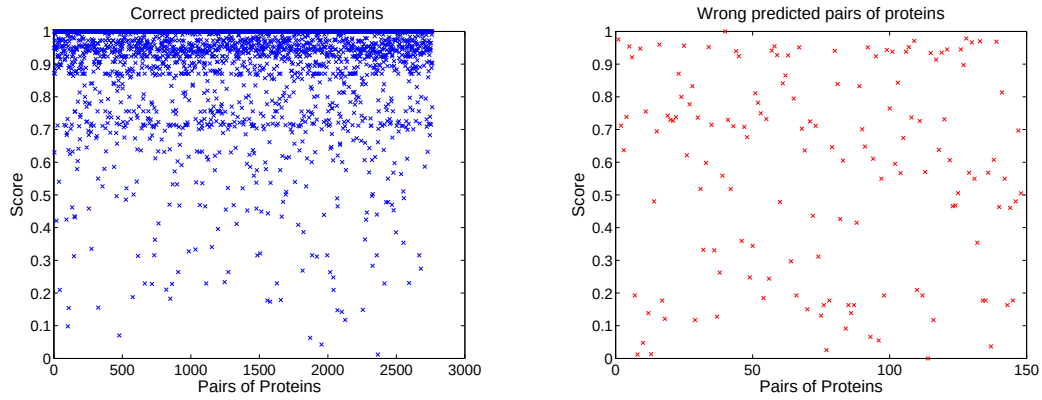
Specificity is defined as the aptitude to identity a true negative in a data set, represented by equation 5.6:

$$Specificity = \frac{TN}{TN + FP} \tag{5.6}$$

Where TN represent the number of true negatives, TP the number of true positives, FP the number of false positives and FN the number of false negatives.

The predictors were executed three times with three different randomly groups for training and test. The linear-SVMs obtained: classification error 6.63 % ± 1.37, specificity 89.17 % ± 2.5 and sensitivity 97.4 % ± 0.37. For RBF-SVMs the results are: classification error 5.03 % ± 0.25, specificity 91.5 % ± 0.88, sensitivity 98.3 % ± 0.36.

In figures 5.4(a) and 5.4(b), it can be observed that the classification score of a test group for correct and wrong predicted pairs, determining clearly two groups. Correct predicted pairs obtain score 0.905 ± 0.149, and as expected the wrong predicted pairs obtain 0.584 ± 0.302. So it is possible to affirm that the proposed SVM predictor is reliable, returning high score for correct classified ones (around to 0.9) and for the not correct predicted pairs (150 protein-protein interaction pairs of a total number of 2911 for this simulation), the mean of the probability is around 0.60, and therefore has bigger uncertainty.



(a) Scores (probabilities) for correct predicted pairs

(b) Scores (probabilities) for not correct predicted pairs

Fig. 5.4: Scores for correct and not correct predicted pairs

To check that the selected features calculated by the proposed similarity measures help to improve the predictive power of SVM, it was performed a comparison between the three features (but using the implementation of this chapter) based on the proposal by Patil et al. [163] (d,g,h) and three highest weighted features (25^{st}, 1^{st}, 22^{st}) with the approach of this chapter (see figure 5.3). Table 5.3 shows the results of this comparison.

LINEAR SVM	The proposed predictors with best features based in the bibliography[163]			Results by Patil[163]			The proposed predictors with the 3 best features		
Features	Error (%)	Sp. (%)	Se. (%)	Sp. (%)	Se. (%)	Features	Error (%)	Sp. (%)	Se. (%)
d+g+h	**7.03±0.47**	92.53±0.52	93.40±0.42	99.40	12.30	1+2+3	11.63±0.05	95.13±0.09	81.23±0.19
d+g	15.73±0.61	92.60±0.57	75.50±0.57	99.30	14.50	1+3	**14.00±0.14**	99.83±0.05	71.47±0.52
d+h	20.50±0.85	99.90±0.00	58.00±1.27	99.20	14.70	1+2	**12.70±0.00**	92.60±0.57	81.73±0.75
d	34.30±0.42	67.57±0.66	63.77±0.24	99.20	14.80	1	**14.37±0.05**	99.83±0.05	70.67±0.38
g+h	**7.03±0.47**	92.53±0.52	93.40±0.42	94.00	44.10	2+3	14.70±0.42	95.13±0.09	74.9±0.42
g	**15.73±0.61**	92.60±0.57	75.50±0.57	74.30	86.70	3	34.43±0.38	38.93±0.05	93.7±0.00
h	20.50±0.85	99.9±0.00	58.00±1.27	62.90	89.70	2	**15.73± 0.61**	92.60±0.57	75.50±0.57
RBFs SVM	The proposed predictors with best features based in the bibliography[163]			Results by Patil[163]			The proposed predictors with the 3 best features		
Features	Error (%)	Sp. (%)	Se. (%)	Sp. (%)	Se. (%)	Features	Error (%)	Sp. (%)	Se. (%)
d+g+h	**7.03±0.47**	92.53±0.52	93.40±0.42	99.40	12.30	1+2+3	7.53±0.24	98.90±0.00	85.67±0.66
d+g	15.37 ± 0.80	91.10±0.85	77.80±2.26	99.30	14.50	1+3	**8.77±0.19**	99.73±0.05	82.27±0.52
d+h	20.50±0.85	99.90±0.00	58.00±1.27	99.20	14.70	1+2	**8.47±0.33**	98.47±0.09	84.27±0.80
d	34.37±0.47	67.57±0.66	63.67±0.24	99.20	14.80	1	**10.83±0.38**	99.83±0.05	77.97±0.94
g+h	**7.03±0.47**	92.53±0.52	93.40±0.42	94.00	44.10	2+3	14.00±0.42	95.57±0.19	75.97±0.47
g	**15.73±0.61**	92.60±0.57	75.50±0.57	74.30	86.70	3	24.73±0.09	99.90±0.00	49.23±0.75
h	20.53±0.83	99.90±0.00	58.00±1.27	62.90	89.70	2	**15.73±0.61**	92.60±0.57	75.50±0.57

Table 5.3: Comparison table with results of Selected Features

Se: sensitivity, Sp: Specificity. 1: (25^{th} feature) Phenotype MIPS calculated by the presented "local similarity" measure in this work, 2:($1^{st} feature$) similar GO annotations, 3: (22^{th} feature) Functional MIPS calculated by the "local similarity" measure .d: interacting Pfam domains (3^{rd} feature); g: similar GO annotations (1^{st} feature) ; h: homologous interactions (2^{st} feature). More than one features are indicated by listing the features separated by a '+' sign.

As it can be observed in table 5.3, in general, the proposed selection of three features does not improve the obtained results with the proposed features by Patil et al. [163]. However, please note that more than 3 features are necessary to provide a correct classification, giving higher values of specificity and sensibility.

It is important to graphically represent the accuracy of a test and expresses the trade-off between the sensitivity and the specificity of the test though a ROC curve (Receiver Operating Characteristic). In figure 5.5 a ROC curve (graphical plot of the sensitivity, vs. (1 - specificity)) using the proposed score as measure of prediction, is presented for the results of the binary RBF-SVM classifier using the 14 selected features by the presented feature selection approach (red line and dots) versus the best results of proposed RBF-SVM approach with the "d+g+h" features proposed by Patil et al. [163] (blue line and cross). It can be observed that better levels of sensibility and specificity using the selected 14 features. Hence, using a reasonable subset of features from the 26 extracted characteristics, the prediction is improved rather than using only 3.

Two of the three most important features (25^{th}, 22^{th}) considered by the proposal of this chapter are derived from the similarity measure for phenotype and functional MIPS catalogue. It can be found that features related with domains have not a important weight in the presented selection and the features related to GO annotations have a important weight. Anyway, it is confirmed that a combination of a subset of relevant features permit the construction of a reliable classifier. The designed SVM classifiers with the selected subset of features show to be good predictors.

Finally, it is mentioned that the linear kernel and the RBF kernel obtain a reliable prediction in the test set. However as expected, the RBF kernel is slightly more reliable.

5.6 Conclusions

The elucidation of protein-protein interactions is a important goal in the current proteomics research. There is a large number of approaches for PPI prediction that use several genomic/proteomic features, utilizing a set of reliable examples and applying SVM [16, 48].

In this work of this chapter, it was proposed the usage of features based on a new similarity measures for yeast extracted from known databases: SwissPfam [24], GOA [37], MIPS CYGD [84], 3did [201], Hintdb [163]. The data set that it was used is a set of examples extracted from a reliability source for yeast [185]. Through a combination of the outcomes of tree feature selection methods, it was performed a characteristic relevance weighting, finally selecting the features with relevance above the mean. With the selected subset, a linear-SVM and a RBF-SVM were constructed as effective predictors, which may provide a confidence score in classification. Hence, it has been checked that the features derived from our similarity measure improved the predictive power of the classifiers, obtaining high specificity and sensibility in prediction

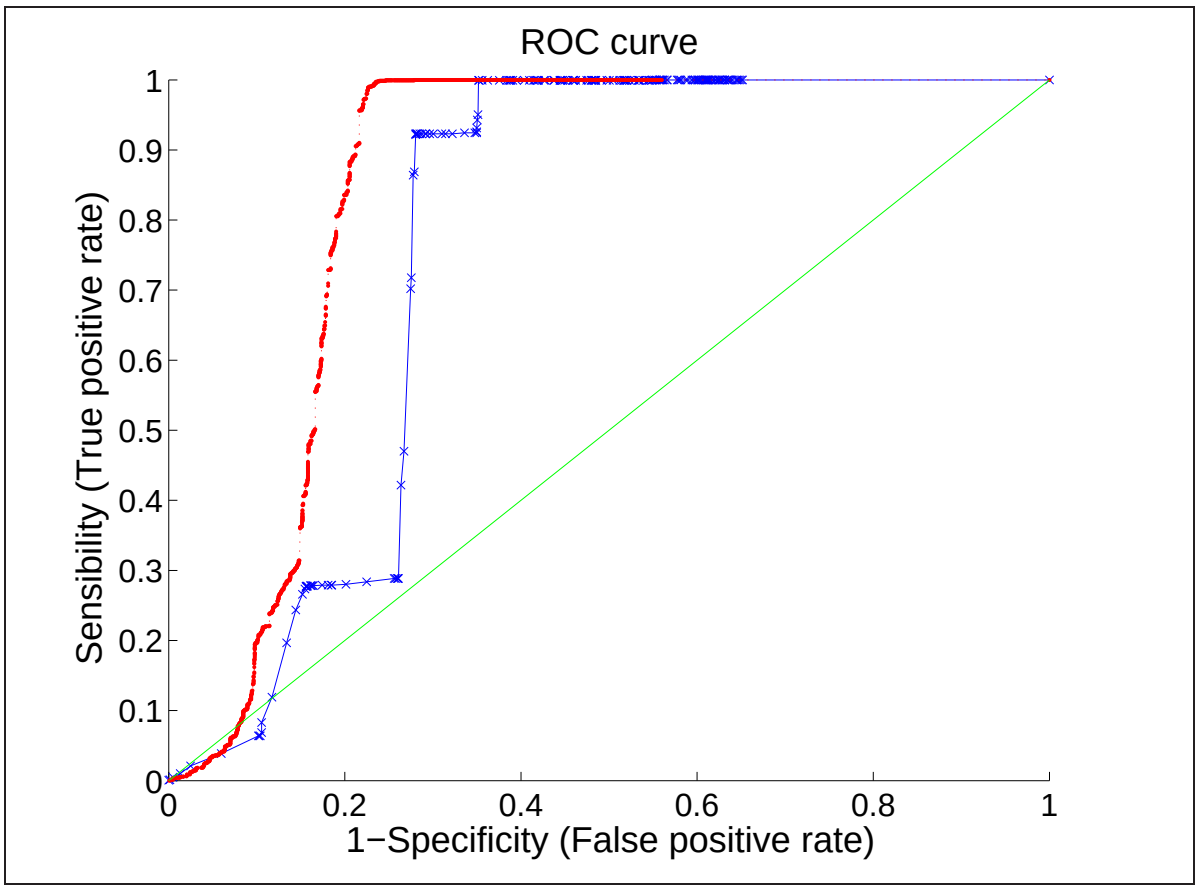


Fig. 5.5: ROC curves. The proposed method in this chapter (red) and a RBF-SVM with features (using the implementation of this work) appeared in Patil et al.[163] (blue)

of PPIs.

For extending the work of this chapter, it could be interesting an analysis with several and a larger size of data since Saeed et al. [185] provide a set of more than 4 million negative examples. Therefore, a clustering approach to select the most relevant samples of that negative set is carried out in chapter 7. The consideration of using different training sets may be a key point, at least different negative sets that are not extensive reported by experimentalists due to the difficult to obtain. In order to test this approach, other negative datasets obtained from different approaches are utilised in next chapters 6 and 7.

Other feature selection approaches may be also interesting to apply with the purpose of find a more suitable subset of features, then a new filter/wrapper feature selection approach is implemented in the next chapter 6. Because of high computational needs of these algorithms, the possibility of applying parallelism to this approach is also

considered. In fact, in chapter 7 it is also implemented an approach applying parallelism in order find a better subset of features to train a more accurate SVM model.

For further work, it is considering to apply this approach to other organisms. Moreover, it is possible to apply other variable selection methods as well as using other machine learning methodologies such as fuzzy systems that provide interpretable solutions [89].

6

USING MACHINE LEARNING TECHNIQUES AND GENOMIC/PROTEOMIC INFORMATION FROM KNOWN DATABASES FOR DEFINING RELEVANT FEATURES FOR PPI CLASSIFICATION

6.1 Introduction

In this chapter, it is a presented a new work as the continuation of the previous chapter 5. Then, other feature selection approach is presented here, including a new "balanced" negative dataset for training classifiers. Besides, external datasets are used to measure the accuracy in classification of the proposed SVM predictor. Instead of using linear kernel for SVM models, RBF kernel is applied because this kernel obtained generally better performance as previously shown (see section 5.5).

Thus, here it is proposed a novel method for constructing a PPI classifier, based on the new approach of feature selection from well known databases, applied specifically to a yeast organism model. This approach uses similarity semantic measures applied to these proteomic features and demonstrates that their use improves the predictive power of constructed classifiers. In a first stage, the proposed method is based on the selection of the most relevant extracted features via the filter-wrapper technique. The RELIEF algorithm [175] is used as filter (see chapter 2). Likewise, the construction of classifiers is based on support vector machines (SVM) using a GSP set and a GSN set. GSP extracted from Saeed et al. [185] is a high-reliability set built using a homologous verification method. The GSN set is a subset of non-interaction pairs randomly selected, as commonly are taken in literature, from a high-quality 4 million non-interaction pairs set from the approach proposed by Saeed et al. [185] with the purpose of increasing the reliability of the model. Additionally, this classifier may

return a confidence score for each prediction by means of a modification of the SVM implementation.

In a second stage, to validate the general applicability of the model, it is used to classify a group of highly reliable external datasets from [234]. Such model is a SVM classifier built using the most relevance selected features that characterize the protein-protein interaction. The used GSP is also extracted from Saeed et al.[185] as previously. However, GSN is a set balanced to GSP, i.e, this negative set is constructed using the method proposed by Yu et al. [234]. Yu et al. [233] expressed the opinion that as experimental protein interaction data is usually obtained via a bait-prey approach (e.g. TAP and Y2H) bait proteins are over-represented in identified interactions. Furthermore, most known protein interactions are predisposed to form clusters, many of them tending to be involved in a large number of hub-forming interactions. Randomly picked non-interacting pairs will therefore display very different typographic characteristics from the positive set, making them distinguishable to a certain degree without any other information. In this way, Yu et al. provided a method for the selection of unbiased negative examples based on the frequency of the proteins in the dataset. Therefore, the negative set is balanced against the positive set, solving the possible problems of randomly paired proteins for negative sets.

In addition, referring these external datasets to validate, there were obtained using computational and experimental approaches together with information from the literature. The datasets were filtered for assessment to avoid biased results, i.e, without any overlapping between the datasets used during the training stage.

The rest of this chapter is organized as follows. The section 6.2 refers to section 5.2 how features describes were extracted from the genomic/proteomic databases used also here. The section 6.3 explains the proposed filter/wrapper variable selection in this chapter, the section 6.4 refers to section 5.4 which kernels for SVM were used and the proposed confidence score, they are also applied here. The section 6.5 describes the results obtained upon applying the proposed approach of this chapter including also a discussion about the results. Finally, the conclusions that can be drawn from this work are set out in the section 6.6.

6.2 Databases and Feature Extraction

In the work of this chapter, the same feature extraction approach using identical databases is applied as described in section 5.2. Therefore, 26 features were extracted from the proposed databases, as a prior step to create the model as previously explained.

6.3 Proposed filter/wrapper feature selection method

As it was already describe in chapter 2, in machine learning theory, filtering out irrelevant or redundant features greatly improves the learning performance of classifiers.

This reduction in the number of features, also known as *feature selection*. As previously explained, two general approaches are commonly applied for feature selection: filter selection methods and wrapper selection methods (see chapter 2). In this study, the a hybrid methodology has been adapted, which is capable of combining the advantages of both approaches.

In the first step of this hybrid methodology, a filter feature selection algorithm is applied called Relief [117] which has proved to be very efficient for estimating feature quality. Relief has been already describe in chapter 2 using the implementation of Relief from Gilad-Bachrach et al. [175]. They implemented a margin-based criterion to measure the quality of a set of features and which can also be used for multi-class categorization problems. This algorithm keeps a weight vector over all features and updates this vector according to the given I/O sample data. Under some assumptions, Kira and Rendell [117] demonstrated that the expected weight is large for relevant features and small for irrelevant ones.

With the purpose of finding out the most suitable number of features to build the proposed model, in a second step, the whole set of possible classifiers are evaluated making use of different feature sets selected according to the ranking of features returned by Relief. In total, 26 sets of features were evaluated, created by starting from the most relevant feature and adding one feature to the previous set every step, until reaching the final whole set. For every set the classifier is trained and computed the classification error in order to select the optimal subset of features from the performance point of view. The section 2.2.1 presents the classifier model used (support vector machines) and deals with the details of its optimization process.

6.4 Support Vector Machines and Proposed Confidence Score

The description of support vector machines is already explained in section 2.2.1 of chapter 2. However, in the work of this chapter, RBF kernel is applied because of this kernel offers generally better performance as previously shown (see section 5.5). In addition, by the means of a modification of SVM model, the model is able to return a confidence score as it is already explain in section 5.5 of the former chapter.

Thus, with the explanation of the above sections, the results section is drawn in the next section.

6.5 Results and Discussion

The experimentation carried out in this chapter can be divided into two parts. First, the proposed filter-wrapper selection approach is applied to obtain an optimized SVM-based PPI classifier with the most important input features. The main idea is to provide a sufficient number of features to the problem, then, in a pre-construction stage of the model, apply a variable selection method. In this way, the model is trained with only those features selected. Second, the validity of the proposed approach is tested against

a set of experimental, computational and literature-collected datasets.

In the filter-wrapper selection approach, a highly reliable dataset has been formed using a gold standard positive and negative set. Saeed et al. [185] provided a set of positive-interaction samples of 4809 pairs of proteins that will be used here as GSP set. Additionally, they provided a reliable set of negative-interaction pairs of proteins formed by 4 million pairs. For the sake of efficiency and computational cost in the design of the models, the GSN set used to build the models was formed by a random subset of samples of similar size to the available GSP set. This is of key importance, as classifiers generalise better when the training data is well balanced. From this dataset, 26 features were extracted, which are given in table 5.1 (in the former chapter 5). The data obtained from the feature extraction stage were normalized in the interval $[0, 1]$ for the application of the proposed approach.

With the objective of evaluating the relevance of these features, the complete reference dataset was subdivided into 70% for training and 30% for testing. Next, the Relief algorithm was applied to the training data, obtaining a ranking of the 26 features in ascending order, according to their estimated relevance (see figure 6.1). Using the proposed filter-wrapper approach, a total of 26 SVM models were obtained for each possible optimal subgroup of features. Hyper-parameters C and γ of each SVM model were optimized by using 10-fold cross-validation. The test set was used to assess the performance of those 26 models. This filter-wrapper approach was repeated four times to obtain a more accurate estimation of the behaviour of the proposed approach, each using a different training-test random subdivision of the reference dataset. All classification performances obtained, including sensitivity and specificity, are graphically shown in figure 6.2. It is recalled that *sensitivity* is the capacity to properly classify an interacting pair, and *specificity* is the capacity to properly classify a non-interacting pair. In other words, sensitivity measures the proportion of actual positives (interacting pairs of proteins) which are correctly identified, and specificity measures the proportion of negatives (non-interacting pairs) which are correctly identified.

In figure 6.2 it can be seen that, for all cases, both accuracy and sensitivity increase as the interaction information augments, i.e., when adding more features to the models. The specificity is kept around 99% with little significant variations.

It can be observed that the performance stops improving significantly beyond the models trained with the eight most relevant features (they obtain a sensitivity and accuracy just slightly lower than those models trained using the complete set of features). It can be therefore considered that those eight features form a suboptimal subset of features for this problem. Table 6.1 shows the specificity, sensitivity and classification error obtained for the four random subdivisions of the reference dataset using the selected and the complete set of features; it can be observed that both models share similar classification accuracy. However, as mentioned before, the models with eight features are slightly more specific and less sensitive than those with 26 features.

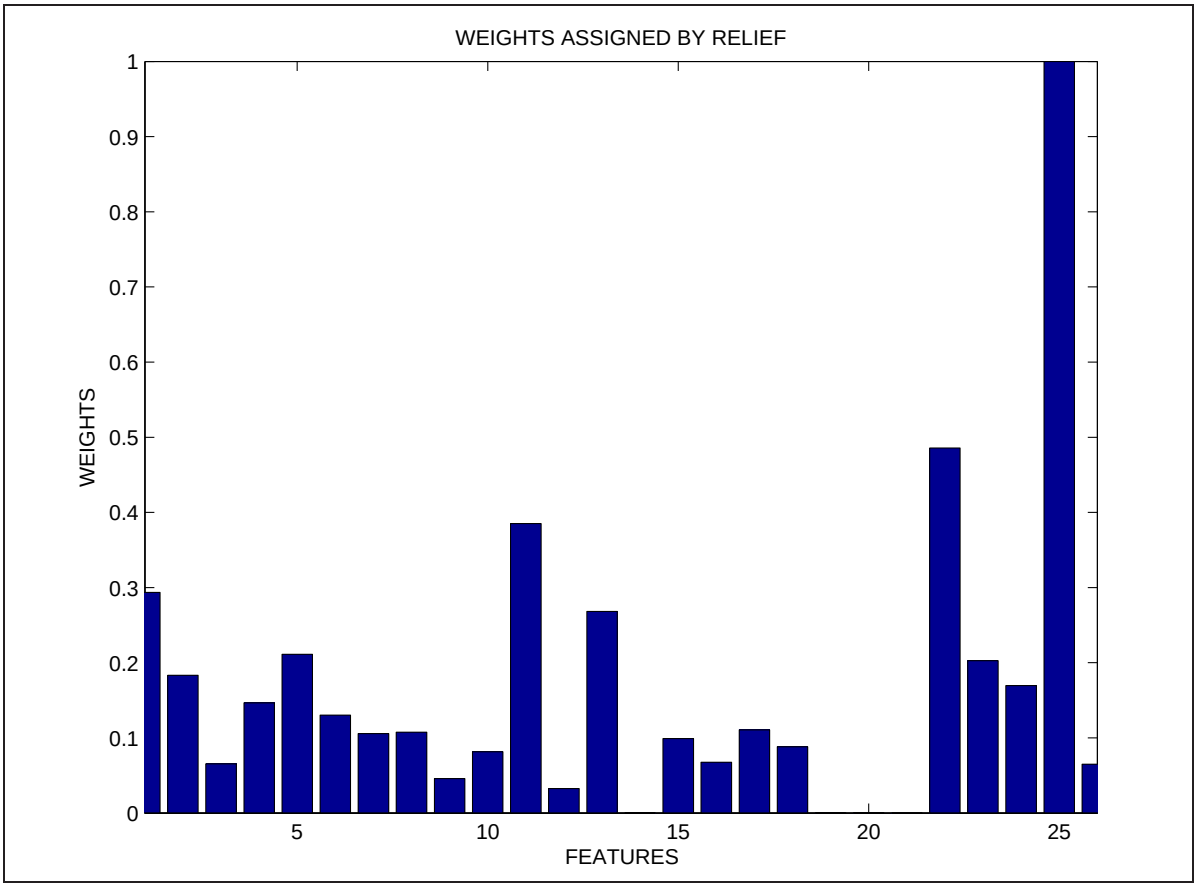


Fig. 6.1: Normalized weights of features obtained by Relief (range [0,1]).

Table 6.1: Classification Results with 4 Randomly Partitioned Datasets using 8 features and 26 features

Training & Test Group	RBF Kernel SVM 8 features				RBF Kernel SVM 26 features			
	Test Acc. (%)	Test Error (%)	Sp. (%)	Se. (%)	Test Acc. (%)	Test Error (%)	Sp. (%)	Se. (%)
1^{st}	96.54	3.46	99.18	93.75	96.32	3.68	98.08	93.89
2^{nd}	96.36	3.64	98.99	91.49	96.50	3.5	97.38	94.80
3^{rd}	95.95	4.05	99.07	92.60	97.630	2.37	99.14	93,66
4^{th}	95.98	4.02	99.05	92.85	96.77	3.23	98.44	95,07
Mean	96.21	3.79	99.07	92.67	96.81	3.19	98.42	94.36
Std. Deviation	0.25	0.25	0.07	0.81	0.5	0.5	0.65	0.59

Se: Sensitivity. Sp: Specificity. Acc.: Accuracy.

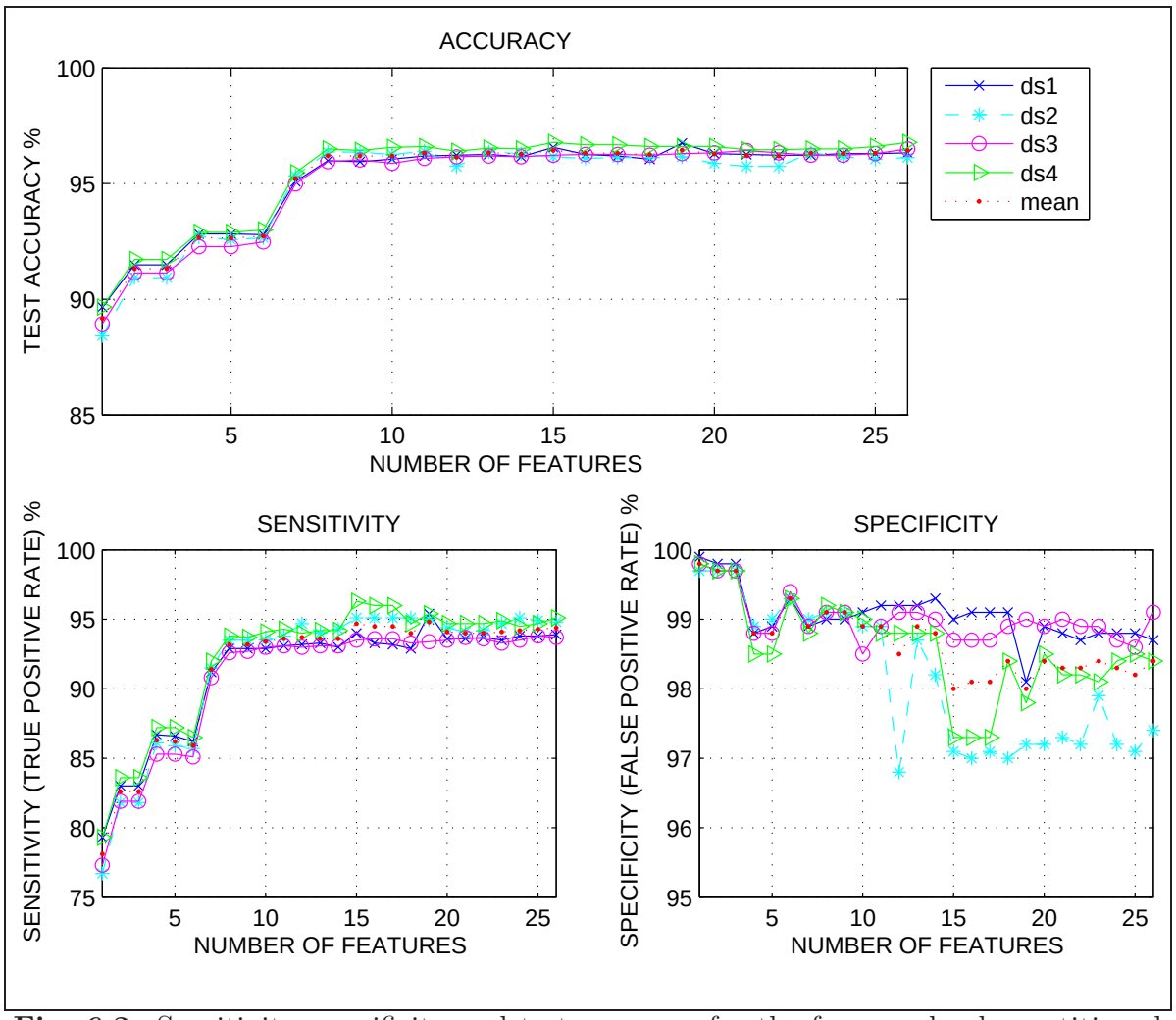


Fig. 6.2: Sensitivity, specificity and test accuracy for the four randomly partitioned datasets and their average values.

The eight features composing the suboptimal selected set are the following (see Table 6.2): 25^{th} referring to *local similarity measure for MIPS phenotype catalogue*, 22^{nd} referring to *local similarity measure for MIPS functional catalogue*, 11^{th} referring to *common terms for the two proteins in MIPS phenotype catalogue*, 1^{st} referring to *number of common GO terms*, 13^{th} referring to *local similarity measure for the number of common GO terms*, 23^{rd} referring to *local similarity measure for MIPS complexes catalogue*, 5^{th} referring to *number of common GO terms in the Biological Processes Ontology* and 2^{nd} referring to *number of shared homologous proteins between a pair of proteins*.

In this sub-optimum set the features concerning protein complex, phenotypes and functional information from MIPS catalogues and GO ontologies have already been

used successfully in other investigations into prediction [120, 106, 230, 34, 240, 35] and proved themselves to be reliable. Note that similarity measures were also included to this set in order to improve the prediction power of the model. In addition, the second feature is included in the sub-optimum set of features, thus supporting the relevance of homology in prediction works, as has been shown in other publications [52, 162, 163, 185].

Table 6.2: Sub-optimal set of selected features

Selected features by Relief
25^{th}, 22^{nd}, 11^{th}, 1^{st}, 13^{th}, 23^{rd}, 5^{th}, 2^{nd}

With the intention of formally determining whether the trained SVM models with 8 features are similar to those with 26 features, it is represented a ROC (Receiver Operating Characteristic) curve using the designed SVM confidence score for the four random subdivisions of the reference dataset. The ROC curve represents the sensitivity with respect to (1-specificity). In machine learning, the AUC (Area Under Curve) statistic of the ROC curve is widely used for model comparison [66, 86]: AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Figure 6.3 shows the ROC curves obtained for both types of models and table 6.3 shows the AUC results. It is observed that, on average, models trained with 8 features are similar to those with 26 features. This reduction in the number of features implies an important saving in information calculation, memory and other computational requirements, obtaining a similar classification performance. This is of vital importance especially in this type of problems, in which there is an overwhelming quantity of information available in different databases, and the programming, access and execution of mathematical calculations for feature extraction can entail a large amount of time.

Table 6.3: Results for R.O.C.: Area Under Curve (AUC)

Training & Test Group	8 features SVM	26 features SVM
1^{st}	0.895	0.879
2^{nd}	0.892	0.865
3^{rd}	0.902	0.883
4^{th}	0.887	0.877
Mean	0.894	0.876
Std. Deviation	0.005	0.007

In the second part, in order to confirm that using just eight selected features, the system is able to correctly classify the presented PPI problem, the training set was changed from previous part. The GSP set extracted from Saeed et al. [185]

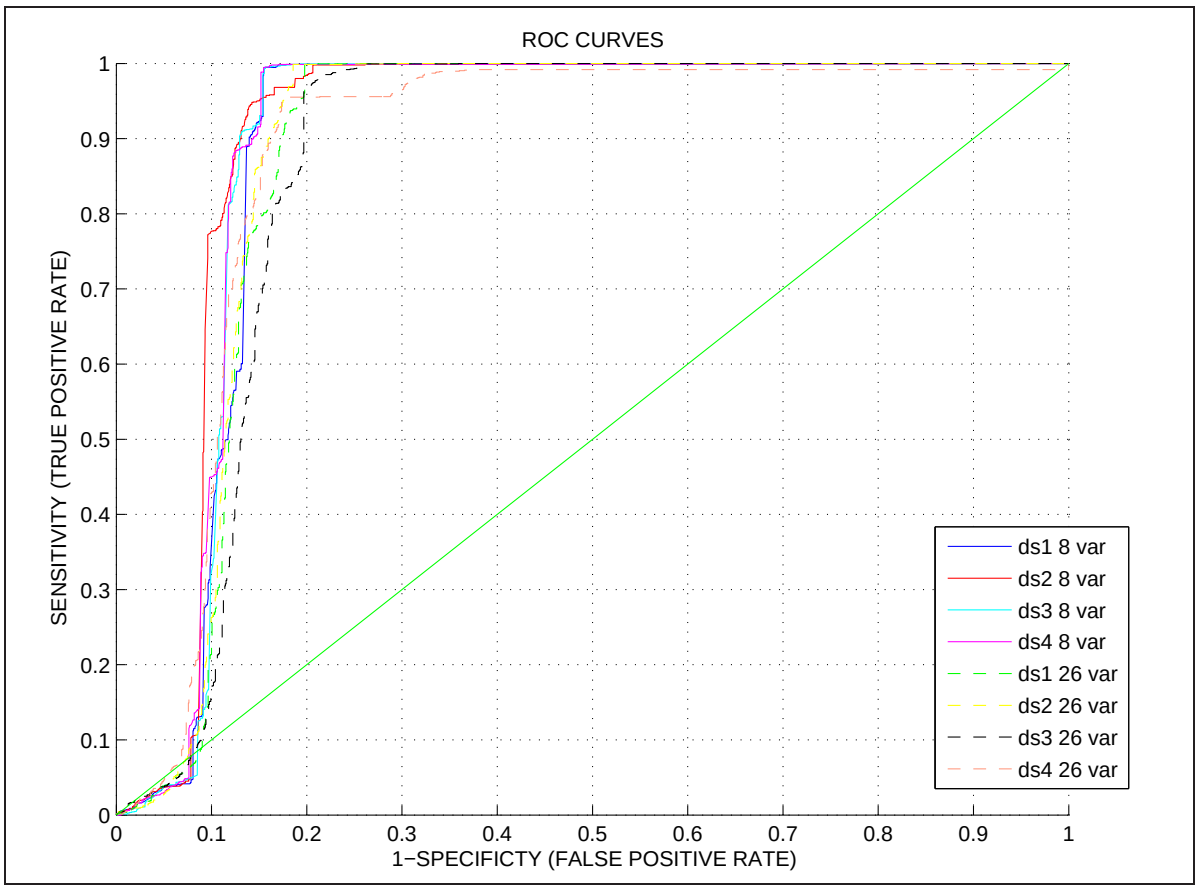


Fig. 6.3: ROC curve for the four randomly partitioned groups (8 and 26 features).

is conserved as part of training set. Now, the GSN set is changed using the the novel approach proposed by Yu et al [233], the new GSN set is *balanced to* GSP set. Thus so, the negative set obtained is a unbiased set of negative examples based on the frequency of the proteins in the GSP set. Therefore, it is an effort to avoid the problems with randomly selected negative pairs that normally display very different typographic characteristics from the positive set, making them distinguishable to a certain degree without any other information.

After the training phase, the behaviour of this approach was tested using the selected subset with the eight most relevant features against a series of high-quality binary interaction datasets taken from Yu et al. [232]: LC-multiple, Binary-GS, Ito-core and Uetz-screen. For negative examples the authors also supplied the RRS dataset. These datasets have been obtained in several different ways (experimentally, computationally and from the literature) and can be freely downloaded from http://interactome.dfci.harvard.edu.

For the sake of reliability, previous to the evaluation of the proposed model, those

interactions from every testing dataset were filtered out to avoid overlapping with the training set (see figure 6.5 for example as proof of overlapping between datasets). The description of the testing datasets would remain:

- The LC-multiple dataset is composed of literature-curated interactions supported by two or more publications. There are 2855 positive interactions. Filtering the overlapping, the dataset contains 2468 interactions.

- Binary-GS dataset. It is a binary gold standard set that was assembled through a computational quality re-examination that includes well-established complexes, as well as conditional interactions and well-documented direct physical interactions in the yeast proteome. There are 1253 positive interactions. Filtering the overlapping, the dataset contains 987 interactions.

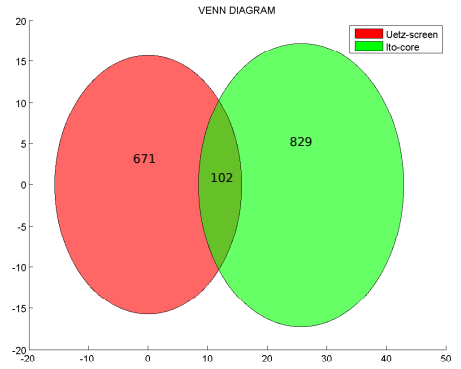
- Uetz-screen: The union of sets found by Uetz et al. in a proteome-scale all-by-all screen [212]. There are 671 positive interactions, 594 after filtering the overlapping.

- Ito-core: interactions found by Ito et al. that appear three times or more [101]. There are 829 positive interactions, 700 after filtering the overlapping.

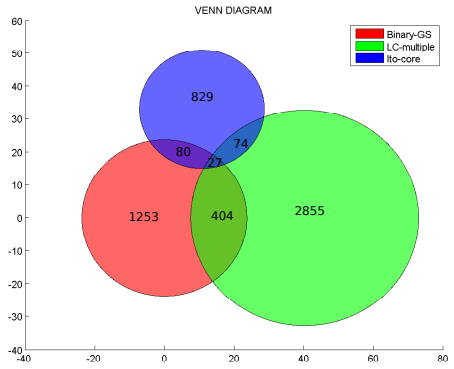
- RRS dataset (Yeast Random Reference Set). This set is composed of random paired proteins that the authors considered that are extremely unlikely to be interacting. There are 156 pairs of non-interacting proteins. There is no overlapping.

- *Dataset Negative 1 and 2.* Due to the low number of non-interacting protein data within the RRS set, two negative subsets of similar size of the available GSP have been used (4897 and 4898 respectively). These set are denoted: Dataset Negative 1 and Dataset Negative 2, and were extracted from Saeed et al. [185].

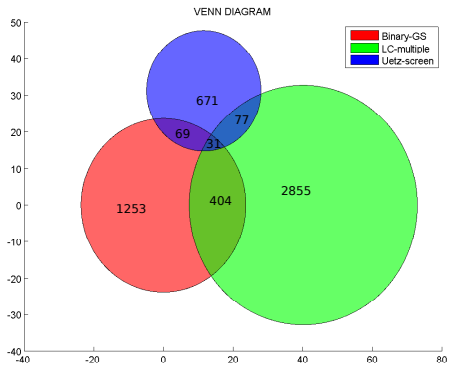
Table 6.4 shows the results obtained. It can be observed that the proposed approach generally attains results with low classification error. Specifically, better accuracies are obtained in the computational and literature-collected datasets. The model classifies the literature-collected dataset "LC-multiple" with a 83% of accuracy. For the computationally-obtained "Binary-GS" data, the classifier attains an accuracy of 84%. For the experimental datasets "Ito-core" and "Uetz-screen" there is a noteworthy difference in performance among them. The model obtains a low error rate of 2.77% for 'Ito-core". However, the obtained accuracy for "Uetz-screen" is 66.78%. That difference may be because of the "Ito-core" dataset contains the interactions found by Ito et al. that repeat three times or more [101] but not with "Uetz-screen". Considering the nature and complexity of the filtering of experimental data, the outcome is still satisfactory, as it is able to validate more than 66% of the interaction pairs in the worst case. Finally, it can be seen that for the negative datasets, the presented model



(a) Overlapping between Uetz-screen and Ito-core datasets



(b) Overlapping between Binay-GS, LC-multiple and Ito-core datasets



(c) Overlapping between Binay-GS, LC-multiple and Uetz-screen datasets

Fig. 6.4: Venn Diagram. Several examples of overlapping between testing datasets

Table 6.4: Prediction Accuracy for Other Experimental and Computational Datasets

Datasets	RBF Kernel SVM Accuracy (%)	Error (%)
LC-multiple	83.403	16.597
Binary-GS	84.312	15.688
Ito core	97.230	2.770
Uetz screen	66.781	33.219
RRS	100	0
Dataset Negative 1	80.188	19.812
Dataset Negative 2	65.591	34.409

is able to detect more than 65% of the non-interactions, in some cases even the 100%, demonstrating that not only is it highly sensitive, but also it proves to be highly specific.

These obtained results using the proposed approach with datasets that are completely independent of the model show its suitability for PPI validation in yeast, making this methodology applicable to other organisms such as Drosophila Melagonaster or Homo Sapiens.

Although it is difficult to make a rigorous comparison with other contributions due to the lack of "gold standard" universal datasets in this field [109], in the following comparison of the proposed approach will be carried out with other techniques existing in the literature to build PPI prediction and validation models.

First, for yeast, in Patil and Nakamura [162], the authors used a Bayesian approach, previously proposed by Jansen et al. [106] with three genomic features to filter out high-throughput PPI datasets of *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens*. Their model was capable of obtaining a sensitivity of 89.7% and a specificity of 62.9%, with a prediction accuracy of 56.3% for true interactions of the Y2H datasets, which were external (i.e. independent) to the model. For two of these datasets, "Ito" and "Uetz" (see table 6.4), the proposed model attained a classification rate of almost 98% for "Ito" and 66% for "Uetz". It should be noted that the homology information (2^{nd} feature) provided by the authors has been included in the presented model and finally selected as one of the most important features. Jiang et al. [109] proposed a mixed framework combining high-quality data filtering with decision trees for PPI prediction, basically using the notation of all GO ontologies, obtaining an accuracy in a range of 65-78%. In the work of this chapter, this information is incorporated in combination with other features to improve the generalisation of this approach.

Finally, it may be worthy to mention that other similarity measures have been proposed in the literature, mainly based on GO annotations, e.g. Wu et al. [230] who were able to detect 35% of the cellular complexes from the MIPS catalogues, or the paper of Wang et al. [222] for the validation of gene expression analysis. However, they did not take into account the Cellular Component ontology as they considered that this information could lead to error due to lack of accurate annotation. In this case, it was decided on proposing a set of similarity measures that allowed for their easy generalisation to a wide range of databases in the attainment of the prediction model. The results presented in this section have proved this decision to be right.

6.6 Conclusions

In this chapter, it is proposed a new approach to PPI dataset processing based on new similarity measures for the extraction of genomic and proteomic information from well-known databases in proteomic analysis of yeast organism model (SwissPfam [24], GOA [37], MIPS CYGD [84], 3did [201], Hintdb [163]) and the application of data mining

techniques for feature selection, model selection and model optimization. This new approach is an "extension" of the previous approach presented in the former chapter 5. A hybrid filter-wrapper feature selection approach has been designed in order to find out the most relevant features from the 26 initial candidates. A sufficient number of features is provided to the problem, thus, in a previous stage of the model, a variable selection method is applied to select a minimum set of characteristics. By means of this approach, 8 features turned out to be the most important ones to create an RBF-SVM classifier with a confidence score which yielded high levels of sensitivity and specificity in the classification of PPIs for the used dataset. By using a ROC analysis, it has been shown that the SVM models trained with the 8 proposed features have a similar behaviour to those trained with the complete set of features. This reduction in the number of required features can also lead to a very important saving in preprocessing time, memory and other computational requirements, which is of special interest in bioinformatics as the number of data that has to be handled can be very vast. Finally, with the objective of evaluating the proposed approach, an analysis of the behaviour of the SVM model with 8 features is performed, this model trained using the proposed gold-standard set, against a collection of different external binary datasets (experimental, computational and literature-collected). The model attained excellent classification results on the validation of datasets obtained by computational approaches and from literature compendium. However the classification error for some experimental datasets and negative datasets was larger. This might be due to the lack of specific information about the nature of those experiments. We can state that it would be advisable to incorporate information relative to the specific experiment of the dataset to be validated, in order to make a more appropriate filtering process and increase the accuracy of the model.

In summary, it can be concluded that by combining information from different databases, using reliable gold standards, providing a set of widely applicable similarity measures to the feature extraction process, using machine learning techniques for feature selection, model selection and model optimization, and an RBF-SVM model with a confidence score, a highly reliable approach to the validation of PPI datasets has been presented in this chapter.

However, as previously commented in the previous section 5.6, it still could be interesting an analysis with several and a larger size of data since Saeed et al. [185] provide a set of more than 4 million negative examples. Therefore, a clustering approach to select the most relevant samples of that negative set is carried out in the next chapter 7. Because of high computational needs of these algorithms, it is also implemented an approach applying parallelism in order find a better subset of features to train a more accurate SVM model to this work in the next chapter 7.

7

SELECTING NEGATIVE SAMPLES FOR PPI PREDICTION USING HIERARCHICAL CLUSTERING METHODOLOGY

7.1 Introduction

In this chapter, a new approach is described solving some issues of the previous chapters 5, 6. Thus, a new feature selection approach using mutual information is presented here. In this case, RBF kernel for training SVM models is also applied because this kernel offers generally better performance as previously shown (see section 5.5). Besides, external datasets as in the chapter 6 are also used to measure the accuracy in classification of the proposed SVM predictor.

Hence in this chapter, a novel method is presented for constructing a SVM classifier for PPI prediction, selecting negative dataset through clustering approach applied to 4 million negative pairs from Saeed et al. [185]. This clustering approach is applied in a effort to avoid the impact of negative dataset on the accuracy of the classifier model. This new method is based on a new of feature extraction and selection using well known databases, applied specifically to a yeast organism model, since yeast is the most widely analysed organism and easiest to find data. A new similarity semantic measures calculated from the features are proposed and demonstrates that their use improves the predictive power of trained classifiers. In addition, this classifier may return a confidence score for each PPI prediction through a modification of the SVM implementation. First, features are extracted for positive and negative samples, then a clustering approach is performed in order to obtain high-reliable non-interacting representative samples. Subsequently a parallel filter-wrapper feature technique select the most relevance extracted features in order to obtain a reliable model. The algorithm called mRMR (minimal-redundancy-maximal-relevance criterion) [166] is used as filter,

it is based on the statistical concept of mutual information. This reduction in the number of features allows for a better training efficiency as the search space for most of the parameters of the model is also reduced [23, 146].

In a second part, with the purpose to validate the generalisation capability of the proposed model, a group of highly reliable external datasets from [234] were classified using the presented method. These datasets to validate were extracted using computational and experimental approaches together with information from the literature. The used models are SVM classifiers built using the most relevance selected features that characterise the protein-protein interaction as explained. They were trained using three training sets, the positive examples were kept and but negative set was changed, each negative set was obtained by a specific method: 1) hierarchical clustering method presented in this chapter, 2) randomly selection and 3) using the approach proposed by Yu et al. [233].

The testing datasets were filtered for assessment to avoid biased results, i.e, without any overlapping between the datasets used during the training stage. High sensitivity and specificity are obtained in both parts using this proposed approach, i.e., the model trained using negative set by the proposed hierarchical clustering method. The presented approach leads to the possibility to be guides for experimentation, being usefully tool to save money and time.

The rest of the chapter is organized as follows. The section 7.2 describes the datasets used to develop the presented approach. The section 7.3 refers to section 5.2 how features describes were extracted from the genomic/proteomic databases, however in this case, co-expression data is not included. The section 7.4 explains the proposed filter/wrapper variable selection using mutual information in this chapter. The section 7.5 refers to section 5.4 which kernels for SVM were used and the proposed confidence score which have been also utilised here. The section 7.6 describes the hierarchical approach used in this chapter to select a representative negative dataset. The section 7.7 explains the master-slave parallel approach applied to the proposed feature extraction approach. The section 7.8 describes the results obtained upon applying the proposed approach of this chapter and a discussion about the results is also included. Finally, the conclusions that can be drawn from this chapter are set out in the section 7.9.

7.2 Material

Two types of datasets were used: training dataset to construct the models and testing datasets to assess the goodness of predictions. A supervised learning classifier as SVM requires positive and negative samples for training data as previously explained in former chapter 2. The positive and negative examples were extracted from Saeed et al. [185], where authors provide a positive dataset is composed of 4809 high-reliability interacting pairs of proteins and a high-quality negative set formed by more than 4

millions of non-interacting pairs. Two negative subsets of the similar size of positive dataset were extracted from this negative set: one dataset is composed by randomly selected non-interaction pairs (4894) and the other one is created by the means of the proposed hierarchical clustering approach presented in this chapter in order to select the most representative negative samples (4988). The main goal of this negative dataset of clustered samples is to represent the whole negative space of more than 4 millions examples avoiding biased results in PPI prediction. The third negative set used in this chapter is created using the method proposed by Yu el at. [233] which is "balanced" to the taken positive set. A comparison of the PPI classification results training three models using these negative datasets is shown on Results section. In training phase, the positive dataset is called Gold Standard Positive (GSP) set and used negative dataset is called Gold Standard Negative (GSN) set.

In the case of testing datasets, these were selected for the sake of validating the generalisation capability of the proposed approach in PPI prediction. Although some of these datasets are already described in section 6.5, but here it is included a briefly description for the sake of clarity and because new information is also added. In this way, a group of reliable binary interaction datasets (LC-multiple, Binary-GS, Uetz-screen and Ito-core) were taken from Yu et al. [232]. These datasets have been obtained using several approaches from experimentally, computationally and grouping datasets well known in the literature. These datasets can be downloaded freely from website http://interactome.dfci.harvard.edu. Besides, other group of used negative testing dataset are also described here. So all proposed testing dataset are:

- The LC-multiple dataset is composed of literature-curated interactions supported by two or more publications. There are 2855 positive interactions.

- Binary-GS dataset. It is a binary gold standard set that was assembled through a computational quality re-examination that includes well-established complexes, as well as conditional interactions and well-documented direct physical interactions in the yeast proteome. There are 1253 positive interactions.

- Uetz-screen: The union of sets found by Uetz et al. in a proteome-scale all-by-all screen [212]. There are 671 positive interactions.

- Ito-core: interactions found by Ito et al. that appear three times or more [101]. There are 829 positive interactions.

- *Random Negative Dataset 1, 2.* Due to the low number of non-interacting protein data within the RRS set, three negative subsets of similar size of the proposed GSP have been utilised. These set are denoted: Random Dataset Negative 1 (4896 pairs) and Random Dataset Negative 2 (4898 pairs) were also randomly selected from Saeed et al. negative set [185].

- *Negative Datasets obtained using the proposed hierarchical clustering approach*: The negative datasets obtained in the last step of the hierarchical clustering process were used as testing negative datasets. In total there are 9 datasets of 5000 examples (see Results section).

For all the datasets a feature extraction process was applied and the data obtained through this process were normalised in the range $[0, 1]$ to apply the proposed method. Furthermore, in a previous step to the evaluation of the model, those interactions from every testing dataset were filtered out to remove overlapping with the training set. In this way, the possible overestimated classification accuracy is avoided through a clustering process selecting a representative negative dataset and filtering step.

7.3 Feature Extraction applied in this chapter

In this chapter, the same feature extraction process is applied as described in section 5.2. However, in this case, 25 features were extracted from the proposed databases. Hence, feature extraction process for the proposed datasets (see previous section 7.2) was applied using well-known databases in proteomics, specially for yeast model organism. The calculated features cover different proteomic information integrating diverse databases already explained in 5.2: Gene Ontology Annotation (GOA) Database [37], MIPS Comprehensive Yeast Genome Database [84], Homologous Interactions database (HINTdb) [163], 3D Interacting Domains database [201] and Swisspfam [24].

Please note that co-expression values from Jansen et al. [106] is not included in this feature extraction approach because there was no enough information for each pair of proteins considered in the utilised datasets. This problem happens above all with the 4 millions negative set from Saeed et al. [185]. Therefore, the full list of considered features in this approach is shown in table 7.1.

7.4 Feature selection approach based on Mutual Information and mRMR criterion

In this chapter, through a pattern recognition process, the irrelevant or redundant features are filtered out improving greatly the learning performance of classifiers [23][146]. This reduction in the number of features, also known as *feature selection*. In the work of this chapter, the PPI prediction is considered as a classification problem, so each sample point represents a pair of proteins which must be classified into one out of two possible classes: non-interacting or interacting pair.

The features selection algorithm can be classified in two groups: filter and wrapper [62] [111] as previously explained in chapter 2. Other authors have utilised a combination of filter and wrapper algorithms [166], in fact in this feature selection approach, a combination between filter and wrapper is used. First, a filter method is

applied in order to obtain the relevance of features and subsequently a wrapper method is performed using support vector machine models from the obtained relevance order.

Different criteria have been applied to evaluate the goodness of a feature[62] [22]. In this case, the proposed filter features selection method is based on mutual information as relevance measure and redundancy between the features through minimal-redundancy-maximal-relevance criterion (mRMR) proposed by Peng et al. [166]. This filter mRMR method is a fast and efficient method because its incremental nature, showing better feature selection and accuracy in classifier including wrapper approach [166] [62]. The mutual information and the description of minimal-redundancy-maximal-relevance criterion (mRMR) method is already explained in chapter 2.

In this chapter, mRMR criterion method was used as filter algorithm with the purpose to obtain the relevance of proposed features. Subsequently a SVM model is trained for each incremental combination of features in ascendant relevance order. Such combination of features is applied adding a feature in a time according on the relevance, starting from the most relevant one, adding the next most relevant one until the feature number 25. In total, 25 SVM models are trained using grid search to estimate the hyper-parameters. A parallel approach was implemented for this filter-wrapper proposal because of memory and computational requirements, reducing the time to obtain the best combination of features that minimise the error classification.

7.5 Support Vector Machine applied in this chapter

The description of support vector machines is already explained in section 2.2.1 of chapter 2. Nonetheless, in this chapter, RBF kernel is applied because of this kernel offers generally better performance as previously shown (see section 5.5). In addition, by the means of a modification of SVM model, the model is able to return a confidence score as it is already explain in section 5.5 of the chapter 5.

The implementation for SVM was taken from the library LIBSVM [40] for Matlab TM. (in this case R2010a). Specifically, C-SVM and RBF kernel was used in the development of the presented work.

7.6 Clustering Methodology

A clustering approach was applied to negative dataset proposed by Saeed et al. [185] in order to obtain a relevant, representative and significant negative subset for training reliable SVM models. Saeed et al. provide more than 4 million of high quality negative pairs, however it is not practicable such amount of data to train a model and there is also an over-representation of negative data that hides the positive samples effect.

In practice, this 4 million set was divided in subset of 50000 pairs approximately (49665 samples) creating 84 subsets of negative samples. This division was carried out due to memory requirement of the available computing system, using the maximum

allowed limit. A classical k-means clustering algorithm [132] was applied to each subset obtaining the 5000 most representative , i.e., reducing 10% of data. Then, new subsets of 50000 negative samples were created adding the 5000 respective samples in order. And again the k-means algorithm is applied to the new subsets obtaining the 5000 most representative samples. This process is repeated until to obtain the last 5000 most representative samples which have a similar size to the proposed positive set (see figure 7.1). This approach is an "hierarchical" and iterative k-means based clustering algorithm which can be run in a parallel computing platform (see section Parallelism Approach) considering the k-means clustering independently in every iteration.

More formally, observing the figure 7.1. In the iteration 1, given an initial group of subsets of 50000 pairs approximately $\mathbf{C} = C_1^1, C_2^1, \ldots, C_1^{84}$. The "hierarchical" clustering approach is a iterative k-means process applied each C_i^j where i is the iteration and j is the subset order. The set resulted for the k-means method is called R_i^j using the same indices i and j from the input subset C_i^j. Thus, R_i^j is formed for the most 5000 representative negative samples set from C_i^j selected by k-means. In the next iterations, C_{i+1}^j is the subset formed by summation of the 10 most 5000 representative negative set R_i^j . When it is not possible applied the summation of every 10 subset R_i^j because there is an inferior number of subsets, then the summation is composed by the maximum number of subsets until complete all considered data. In general, $C_i^j = \sum_{m=(j-1)*10+1}^{j*10} R_{i-1}^m$ given the iteration i and the subset j. In the work developed in this chapter, 3 iterations were executed until obtaining the most 5000 representative negative set from the whole set of more than 4 millions of negative samples. The iteration 2, there were 9 subsets $C_2^1, C_2^2, \ldots, C_2^9$ where C_2^9 contains 20000 pairs. The subsets resulted by k-means $R_2^1, R_2^2, \ldots R_2^9$ create a new C_3^1 of 45000 elements. In the final step, in iteration 3 it is obtained R_3^1 that will be used as part of training set as representation of the negative space from the whole negative set. The $R_2^1, R_2^2, \ldots R_2^9$ will be used as testing set in Results section, after a filtering process from training set, they are called $R_3^{test\_1}$, $R_3^{test\_2}$, $R_2^{test\_3}$, $R_3^{test\_4}$, $R_3^{test\_5}$, $R_3^{test\_6}$, $R_3^{test\_7}$, $R_3^{test\_8}$, $R_3^{test\_9}$.

With this process, the main goal of obtaining a representative negative dataset and not biased from a high-quality negative set is fulfilled.

7.7 Parallelism approach applied for this problem

The filter/wrapper feature selection proposed in this chapter demands a high computational resources. The classical and simple master-slave approach was adopted [83], a master process sends tasks and data to slave process, master process receives results from slaves and controls the finalization of the tasks. In this case, the tasks are to train SVM model including grid search for hyper-parameters and data are the selected features, training and testing datasets for slave processes. In addition, the "hierarchical" k-means clustering algorithm from previous section could be implemented in a parallel computing platform using this approach.

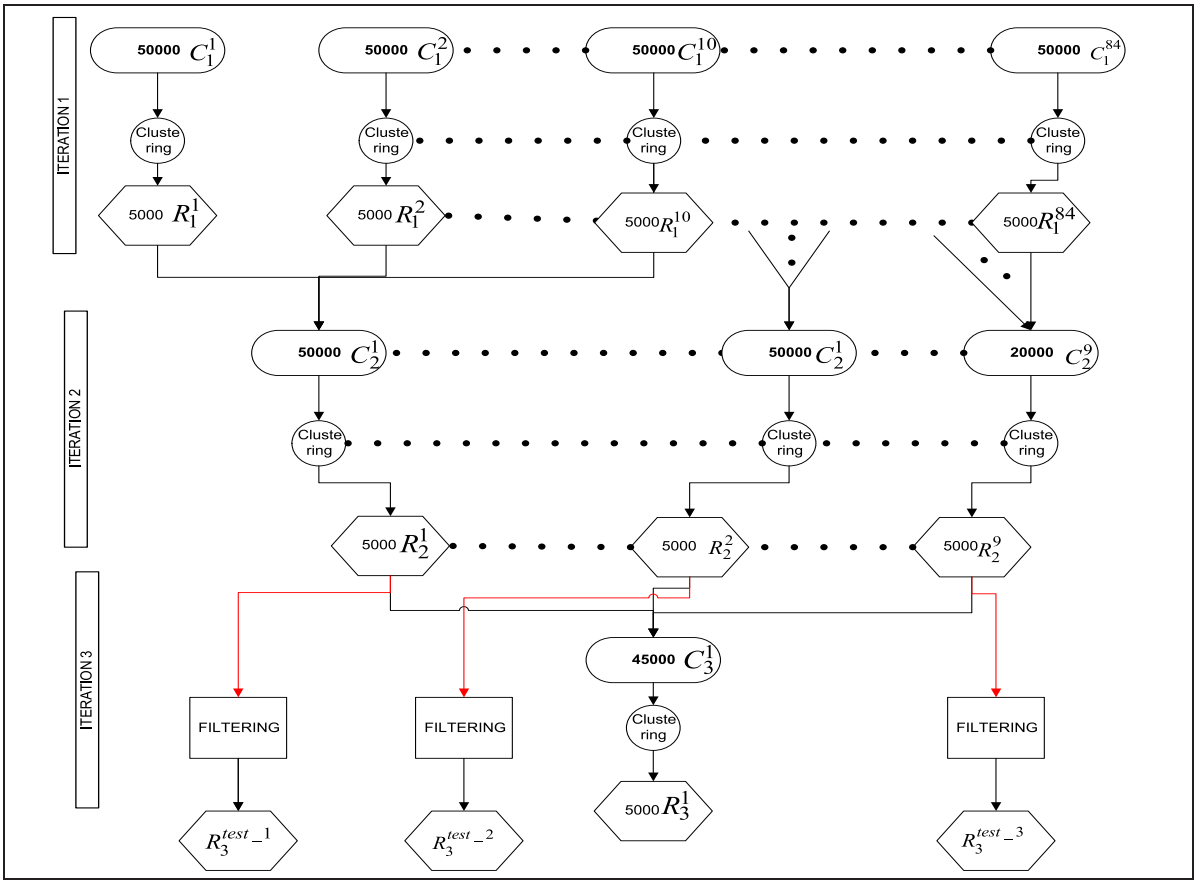


Fig. 7.1: Diagram for the proposed "hierarchical" k-means based clustering algorithm applied. It is a iterative k-means process. The application to this problem, selecting the most 5000 representative negative samples of the whole set.

The implementation of this approach was realized using MPIMEX [82], a new interface that allows MATLAB standalone applications to call MPI (Message Passing Interface) standard routines. This interface was developed in ATC research group where this thesis is presented. MPI is a library specification for message-passing, proposed as a standard by a broadly based committee of vendors, implementers, and users as it is defined in http://www-unix.mcs.anl.gov/mpi/.

This parallel approach was running in a cluster of computer. This cluster was formed by 13 nodes with dual processors Intel Xeon E5320 2.86GHz, 4GB RAM memory and 250GB HDD. All nodes are connected using Gigabit Ethernet. The operating system installed is Linux CentOS 4.6 (rocks) TM. . This cluster was purchased using public funds from Spanish Ministry of Education project TIN 2007-60587. The time of execution was reduced from 16 days in a single computer to 32 hours to train all the SVM models.

7.8 Results and Discussion

The results are composed of two parts. In the fist part a "sub-optimal" set of features are selected though the filter/wrapper feature selection process using the parallel approach. The training data for RBF-SVM model is composed by a GSP set and for a GSN set which is the set resulted of applying iterative clustering approach as explained in section Material and Methods. In the second part, taking this suboptimal set of features, three RBF-SVM classifiers are constructed using three training sets respectively. All training sets have the same GSP set for positive examples. In one case the GSN set is the negative set obtained using the hierarchical clustering method from the first part and in a second case the GSN set is a randomly selected negative set as commented. The third case, the GSN set was created using the approach proposed by Yu el al. [233], it is a "balanced" set to GSP. Subsequently, a comparison of the results obtained of three RBF-SVM classifiers trained with the all proposed negative datasets is discussed.

Previously to the filter/wrapper feature selection process, the feature extraction process is applied to all available datasets. The 25 features were also extracted for the 4 million negative set from Saeed et al. [185], but due to computational requirements the whole set where divided into 84 subsets of 50000 samples approximately. In order to obtain a representative negative dataset of the whole negative space, the iterative k-means clustering approach was applied to this 84 subsets as explained in previous section of Clustering. In total, three iterations selecting 5000 negative representative samples were realized using the clustering approach. In the first iteration, euclidean k-means method was applied to the 84 subsets creating 5000 centroids, and 9 new subsets (8 subset of 50000 and the last one of 20000 negative examples) were obtained adding the selected 5000 negative representative samples of each previous subsets. In the second iteration, the k-means was applied again to the 9 new subsets taking 5000 new negative representative samples of each subsets and creating another new subset of 45000 samples (the representatives of 9 subsets summation). In the third and last iteration, the last 5000 most representative negative samples taking as GSN set for training data was obtained of clustering the previous subset. The taken negative pairs were selected using the minimum Euclidean distance to the centroid of the each cluster. A diagram of this process is represented in figure 7.1.

In this way, the considered data (GSP and clustered GSN sets) was used to apply the presented paralleled filter/wrapper feature selection process. Because of memory requirements in the construction of the 25 SVM models, this data were randomly divided in 70% for training SVM and 30% for testing the performance of obtained models. Hence, four randomly division of data as 4 training/test datasets were used in this feature selection approach in a cluster of computers as commented in section Parallelism Approach. In order to obtain the best hyper-parameters for SVM models, grid-search and 10-fold cross-validation were implemented. In figure 7.2 is shown for

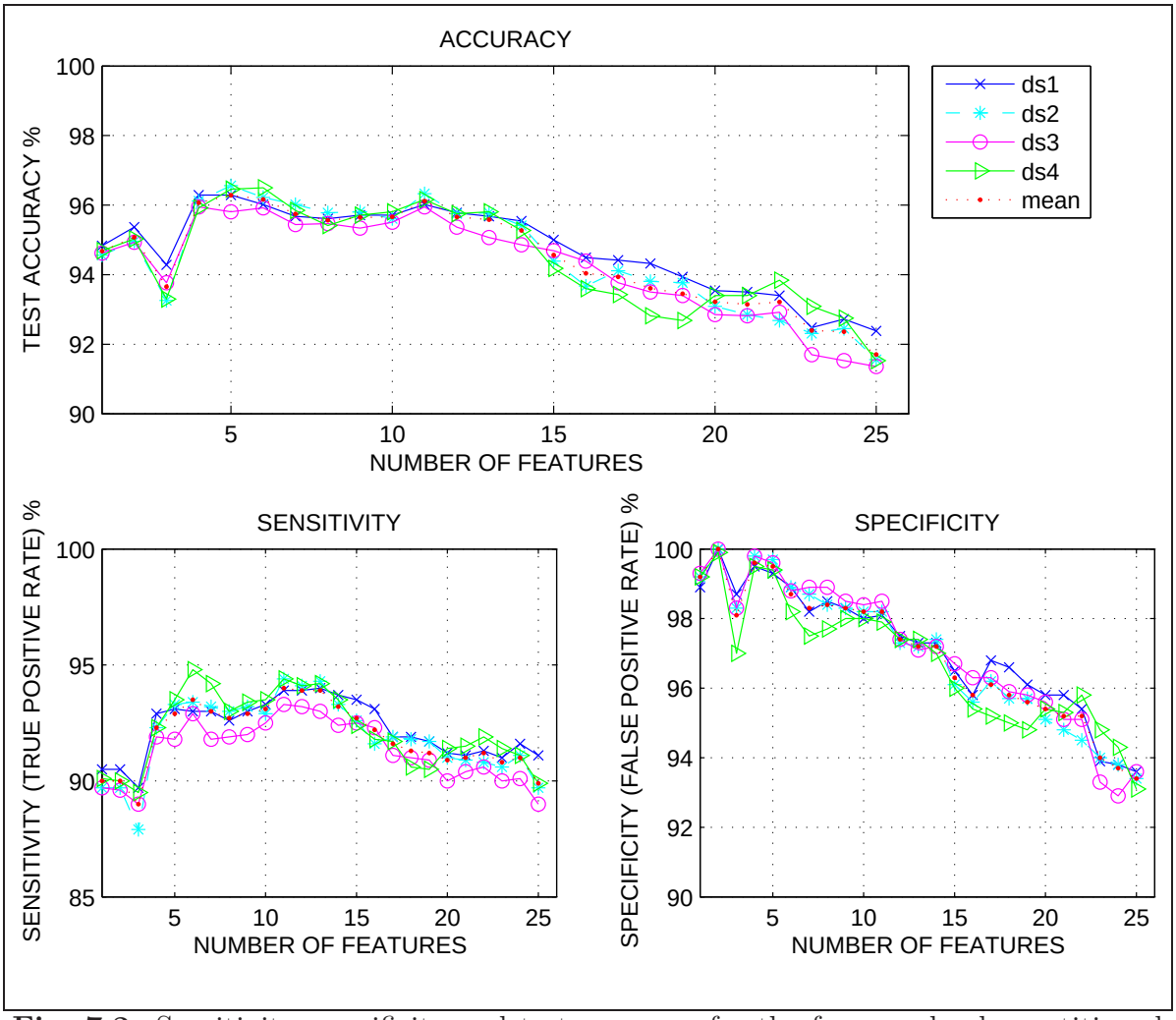


Fig. 7.2: Sensitivity, specificity and test accuracy for the four randomly partitioned datasets and their average values.

all 25 SVM models the accuracy, sensitivity and specificity obtained using the order of feature relevance reported by mRMR filter method. It can be observed that an excess of information may lead to overfitting, i.e, the interaction information decreases when adding more features to the models, specially for testing case. The last added features were considered for mRMR method as more irrelevant or redundant than the features in the first positions. In figure 7.2, it can be observed that the performance stalled or not significantly improvement is reached beyond 6 features, even it gets worse due to an excess of information, so the sub-optimum selected set is composed for that 6 features: 13^{th} referring to *global similarity measure for* 1^{st} *features: common GO terms using all ontologies*, 3^{rd} referring to *number of Swisspfam domains for a pair in 3did*, 10^{th} referring to *common terms for the two proteins in MIPS phenotype catalogue*,

8^{th} referring to *common terms for the two proteins in MIPS functional catalogue*, 7^{th} referring to*common terms for the two proteins in MIPS complexes catalogue* and 2^{nd} referring to *number of shared homologous proteins between a pair of proteins*.

In the selected sub-optimum set, the features concerning protein complex, phenotypes and functional data from MIPS CYGD catalogues have already been used successfully and proved themselves to be reliable in interacting prediction analysis [120] [106] [230] [34] [240] [35]. Note that a global similarity measures were also included to this sub-optimum set of features with the purpose to improve the performance of the classifier in PPI prediction. At the same time, domain information (3^{rd} feature) have provided a suitable framework in PPI prediction works [54] [96]. Moreover, the second feature refers to homology which the relevance of data has been shown in previous publications [52] [162] [163] [185].

In order to check if the SVM models trained with 6 features are significant, a ROC (Receiver Operating Characteristic) was plotted using the confidence score presented in this work, previously explained in section Methods. The ROC curve shows the sensitivity values with respect to 1-specificity values. The used statistic to measure the goodness of the classification was the widely extended AUC (Area Under Curve) [66] [86]. This statistic represents the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. In figure 7.3 and table 7.2 are shown the results for 6 features SVM model and 25 features SVM model showing better performance the SVM trained with a sub-optimum set. As mentioned, this reduction in the number of features implies a significant saving in memory, calculation and other computational requirements, obtaining a overfitting utilising the whole set.

In the second part, the behaviour of this approach is tested using the selected subset of the six most relevant characteristics. Three RBF-SVM models are built with three training sets, they share the same GSP but GSN is different. In one case the GSN is the negative set from the first part created using the proposed hierarchical clustering approach presented in this paper (it is also called clustered training dataset). In the second case the GSN is a randomly selected negative set (called random training dataset) and the last case, the GSN is a negative set "balanced" to GSP set obtained using the approach by Yu et al. [233]. This third GSN is created using a selection of unbiased negative examples based on the frequency of the proteins in the positive set. The testing datasets, detailed in Material Section, cover from positive and negative sets and they were in different ways: experimentally, from the literature and computationally. Additionally, in order to make a reliable comparison, previous to the evaluation of the models, the interactions for each testing dataset were filtered out to avoid overlapping with the respective training set. The new sizes of the testing datasets are shown in table 7.3.

Therefore, the results of these models are shown in table 7.4 and figure 7.4 for positive datasets and figure 7.5 for negative datasets. In general, the SVM model

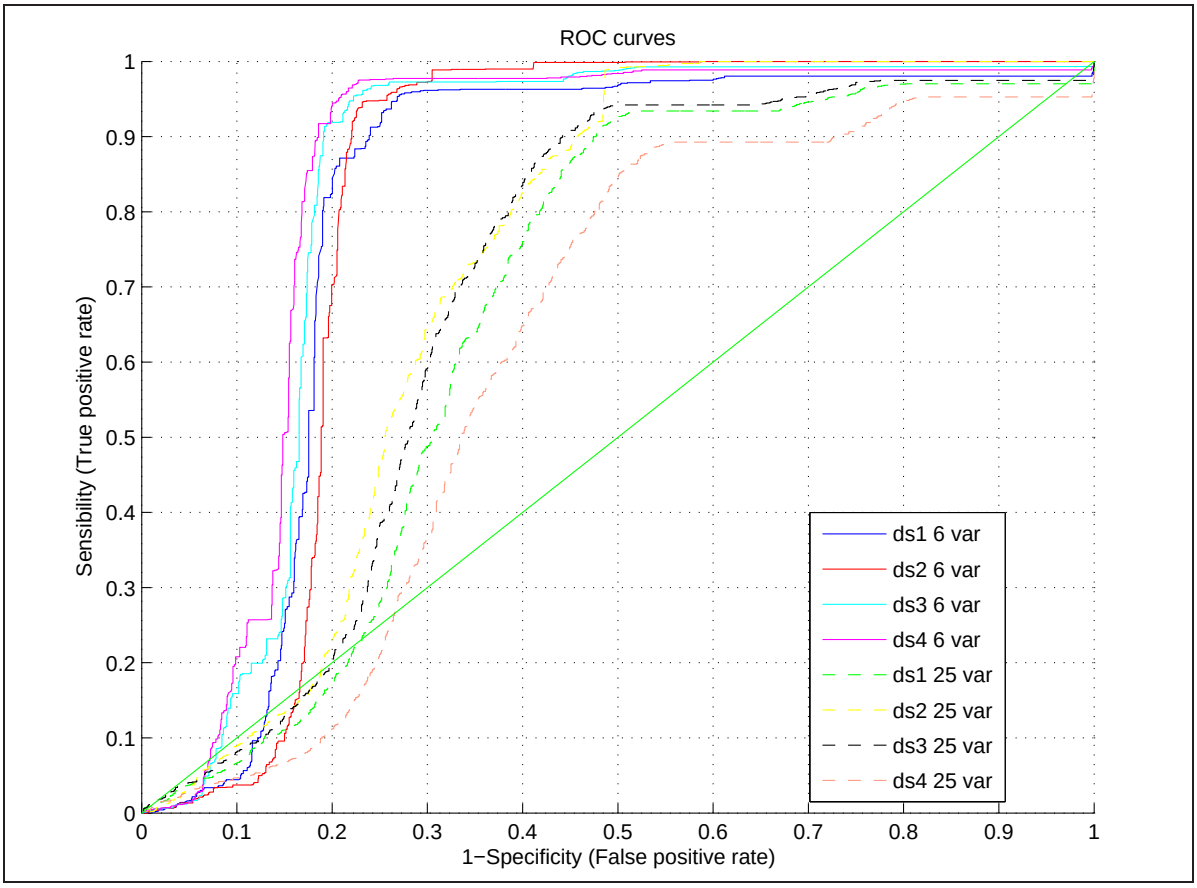


Fig. 7.3: ROC curve for the four randomly partitioned groups (6 and 26 features).

trained using the negative generated by the proposed hierarchical clustering approach presented in this paper has a better performance in comparison with the rest of models, i.e, the models that used the randomly selected negative set and the balanced negative set. Globally, the obtained results were slightly worse in the experimental datasets than in the computational and literature datasets. The models classify the literature-extracted dataset "LC-multiple" with a range between 93-95% of accuracy. For the computationally obtained "Binary-GS" dataset, the classifiers obtain range of accuracy between 92-95%. Between the experimental datasets "Uetz screen" [212] and "Ito core" [101], the reported accuracies are sightly lower than the previous datasets with ranges 72-81% between and 76-80% respectively for the case of the models trained the negative set from clustering approach and the negative set from randomly selection. Nevertheless in the case of the model trained using the "balanced" negative set, the accuracies for both datasets are about 50%. However considering the nature and complexity of the filtering in experimental data, the obtained accuracy is still satisfactory at least the case of the model trained using the negative set from clustering approach. Referring to the

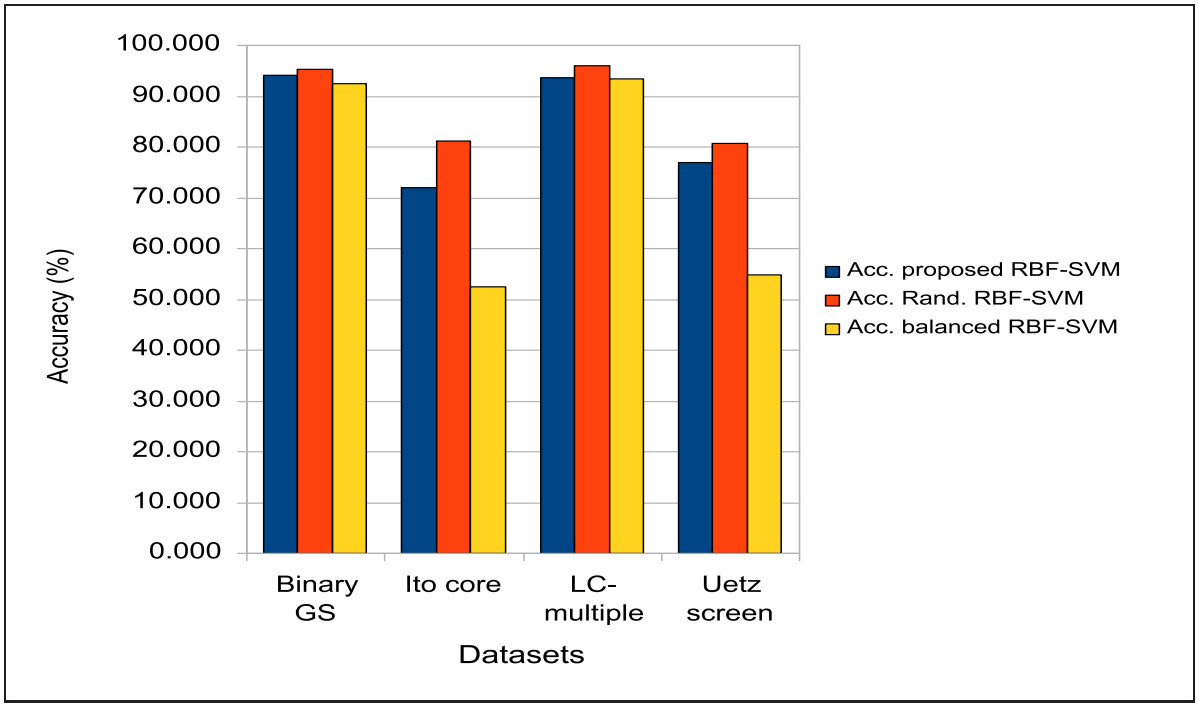
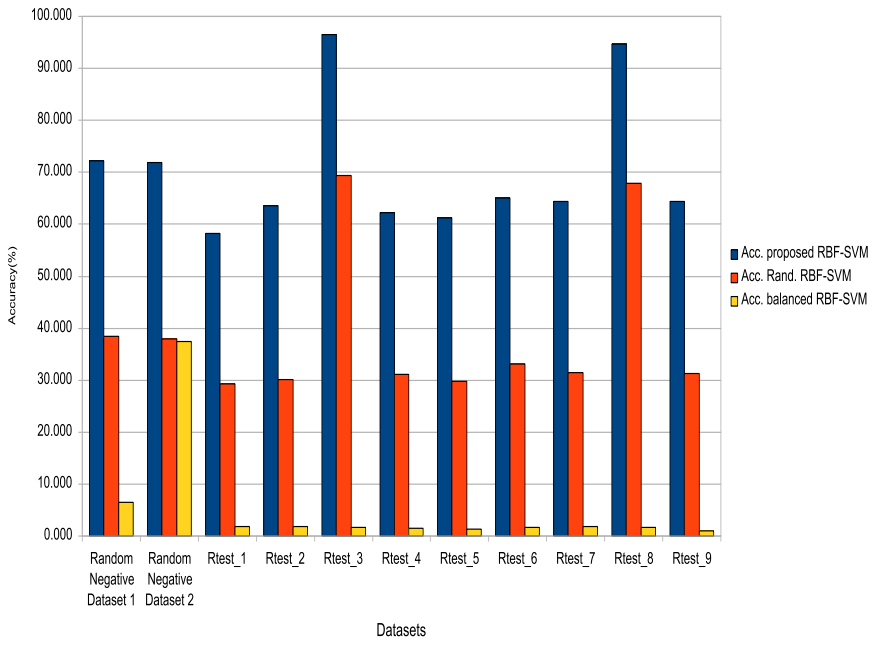


Fig. 7.4: Comparison of accuracy obtained in positive datasets for the three trained models: the SVM model trained using the training set formed by GSP set and the GSN set obtained using the proposed hierarchical clustering method (clustered), the SVM model trained using the training set where the GSN set was randomly selected (Rand. RBF-SVM) and the balanced RBF-SVM is the SVM model trained using the training set formed by GSP set and the GSN set obtained using the approach to create "balanced" negative set by Yu et al.

negative datasets, the model trained using the GSN examples extracted by clustering method in the training set obtained as expected better results than the model trained using randomly negative set, with a minimum relative difference about 28% comparing to randomly selected negative set and a general maximum about 90% in the case of the model trained using "balanced" negative set. The set of the clustering approach have a relevant representation of the negative search space from a high-reliability negative set from Saeed et al. [185]. But with the "balanced" negative set is not happening this, the negative set is "balanced" to positive, but due to it is based on the frequencies of proteins in the positive part is not enough to recognise any negative case. Although the results of negative datasets respecting to positive datasets are worse, the difficult and complexity to predict negatives make the results still acceptable. In can be observed that the relative difference in positive datasets are better for the model trained with the randomly-selected negative set but that difference is not so strong, even can be a slightly overestimation. The accuracy could be artificially inflated by a bias towards



Please note that Rtest_1, Rtest_2, Rtest_3, Rtest_4, Rtest_5, Rtest_6, Rtest_7, Rtest_8 and Rtest_9 correspond to: $R_3^{test\_1}$, $R_3^{test\_2}$, $R_2^{test\_3}$, $R_3^{test\_4}$, $R_3^{test\_5}$, $R_3^{test\_6}$, $R_3^{test\_7}$, $R_3^{test\_8}$, $R_3^{test\_9}$. And the "balanced" negative set is created using the approach by Yu et al.[233]

Fig. 7.5: Comparison of accuracy obtained in negative datasets for the two trained models: the SVM model trained using the training set formed by GSP set and the GSN set obtained using the proposed hierarchical clustering method (clustered), the SVM model trained using the training set where the GSN set was randomly selected (Rand. RBF-SVM) and the balanced RBF-SVM is the SVM model trained using the training set formed by GSP set and the GSN set obtained using the approach to create "balanced" negative.

dominant samples in the positive data as Yu et al. showed [233]. With a such sub-optimum set of features, a SVM model is able to classify PPIs with relative notorious accuracy any positive and negative datasets.

First, in Patil and Nakamura [162], the authors used a Bayesian approach, previously proposed by Jansen et al. [106] with only three features for the filtering out of high-throughput datasets of the organisms *Saccharomyces cerevisiae* (Yeast), *Caenorhabditis elegans, Drosophila melanogaster* and *Homo sapiens*. Their model was able to obtain a sensibility of 89.7% and a specificity of 62.9%, being only capable to attain a prediction accuracy of 56.3% for true interactions for the datasets Y2H,

external to the model. For two datasets called "Ito" and "Uetz" (see table 7.4), the presented model trained with the negative set from clustering method reported a classification rates between 76-93%. In Jiang et al. [109], a mixed framework is proposed combining high-quality data filtering with decision trees in PPI prediction, taking as base the notation of all GO ontologies, aiming an accuracy in a range of 65-78%. Thus, that information was incorporated in combination with other features to improve the generalisation of the proposed approach. Other similarity measures have been proposed, mainly based in the GO annotations, e.g. the works by Wu et al. [230] that was able to detect the 35% of the cellular complexes from the MIPS CYGD catalogues, or in the work by Wang et al. [222] for the validation of gene expression analysis. Nevertheless, the authors did not take into account the cellular component ontology because it was considered that this ontology includes ambiguous annotations that may lead to error. In this approach, it was opted for proposing a set of similarity measures that permit their generalisation to a wide range of databases in the obtaining of the presented prediction model.

7.9 Conclusions of this chapter

In this chapter, solving the previous issues commented in 5.6, a new approach to build a SVM classifier in PPI prediction is presented. The approach has several notorious processes: a feature extraction using well-known databases, a new filter-wrapper feature selection implemented in a master-slave parallel approach, a reliable and representative negative dataset for training by the means of "hierarchical" k-means clustering. The filter method is based on the statistical concept mutual information using mRMR criterion which is reliable and quick method. In addition, a confidence score is presented through a modification of SVM model implementation. A comparison between randomly selected negative dataset, a "balanced" negative set obtained using Yu et al. approach [233] and negative dataset obtained using the 'hierarchical" k-means clustering method presented in this paper is done where the model training using the set resulted by the clustering approach has better performance. This comparison also allowed to check the generalization capacity of the presented approach for the sake of evaluation of external datasets previously filtered. Hence a fair negative selection method is presented avoiding the overestimation in classification of PPI.

For further work, a hierarchical parallel clustering could improve the performance of classifier with the purpose of obtained a balanced negative dataset using a more complex clustering algorithm. It is considered to apply this approach to other model organisms as Homo sapiens. A parallel approach was applied making a better load balancing would be suitable to reduce time computation in the filter/wrapper feature selection approach.

In summary, it is concluded that by combining data from several databases, using reliable positive and clustered negative samples for training, supporting a set of

widely applicable similarity measures to the feature extraction process, using mutual information methods for feature selection, and RBF-SVM models capable to return a confidence score, a reliable approach to the validation of protein-protein interaction datasets is presented.

In the next chapter 8, with the purpose to tackle a more specific problem, a bioinformatics approach has been outlined which aims to characterize complex molecular associations such as the E3-machinery – protein substrate relationship.

Table 7.1: Description of the 25 extracted features

Number	Description	Type
1^{st}	$\#(A_{GOA} \cap B_{GOA})$ from GOA DB taking 3 ontologies together (P,F,C)	integer
2^{nd}	Number of homologous for $(protA, ProtB)$ from HintDB	integer
3^{rd}	$\#[(A_{SPFAM} \cap 3DID) + (B_{SPFAM} \cap 3DID)]$, A and B are domains extracted form SwissPfam, 3DID is 3did database	integer
4^{th}	$\#(A_{GOA-P} \cap B_{GOA-P})$ from GOA DB taking Biological Process ontology	integer
5^{th}	$\#(A_{GOA-C} \cap B_{GOA-C})$ from GOA DB taking Cellular Compartment ontology	integer
6^{th}	$\#(A_{GOA-F} \cap B_{GOA-F})$ from GOA DB taking Molecular Function ontology	integer
7^{th}	$\#(A_{MIPS-F} \cap B_{MIPS-F})$ from functional MIPS catalogue identifiers	integer
8^{th}	$\#(A_{MIPS-C} \cap B_{MPIS-C})$ from complexes MIPS catalogue identifiers	integer
9^{th}	$\#(A_{MIPS-P} \cap B_{MIPS-P})$ from proteins MIPS catalogue identifiers	integer
10^{th}	$\#(A_{MPIS-FE} \cap B_{MPIS-FE})$ from phenotypes MIPS catalogue identifiers	integer
11^{th}	$\#(A_{MPIS-FCC} \cap B_{MIPS-FCC})$ from sub-cellular compartments MIPS catalogue identifiers	integer
12^{th}	Local similarity of 1^{st} feature	real
13^{th}	Global similarity of 1^{st} feature	real
14^{th}	$\#[((A_{SPFAM} \cap 3DID) + (B_{SPFAM} \cap 3DID))]/\#(A_{SPFAM} \cup B_{SPFAM})$	real
15^{th}	Local similarity of 4^{th} feature	real
16^{th}	Local similarity of 5^{th} feature	real
17^{th}	Local similarity of 6^{th} feature	real
18^{th}	Global similarity of 4^{th} feature	real
19^{th}	Global similarity of 5^{th} feature	real
20^{th}	Global similarity of 6^{th} feature	real
21^{th}	Local similarity of 7^{th} feature	real
22^{th}	Local similarity of 8^{th} feature	real
23^{th}	Local similarity of 9^{th} feature	real
24^{th}	Local similarity of $10th$ feature	real
25^{th}	Local similarity of 11^{th} feature	real

The symbol # indicates the number of elements in a set. See equations 5.1 and 5.2.

Table 7.2: Results for R.O.C.: Area Under Curve (AUC)

Training & Test Group	6 features SVM	25 features SVM
1^{st}	0.808	0.672
2^{nd}	0.812	0.725
3^{rd}	0.836	0.698
4^{th}	0.846	0.619
Mean	0.826	0.678
Std. Deviation	0.016	0.039

The ROC curve was constructed using the proposed confidence score for the four randomised sets (70% training, 30% test). The RBF Kernel SVM were trained using 6 features and 25 features .
Std Dev: Standard Deviation.

Table 7.3: New sizes of datasets after filtering process

Datasets	Size of filtering training set with GSN set obtained using the presented hierarchical clustering	Size of filtering training set with randomly selected GSN set	Size of filtering training set with "balanced" GSN set obtained from the approach by Yu et al. [233]
Binary GS	933	937	987
Ito core	680	686	700
LCmultiple	2362	2380	2468
Uetz screen	574	584	594
Random Negative Dataset 1	4893	4894	4894
Random Negative Dataset 2	4895	4894	4898
$R_3^{test\_1}$	4735	4995	4992
$R_3^{test\_2}$	4788	4995	4994
$R_3^{test\_3}$	4814	4991	4991
$R_3^{test\_4}$	4844	4987	4992
$R_3^{test\_5}$	4854	4983	4986
$R_3^{test\_6}$	4816	4991	4994
$R_3^{test\_7}$	4837	4985	4990
$R_3^{test\_8}$	4797	4994	4994
$R_3^{test\_9}$	4873	4994	4996

Table 7.4: Accuracy using the most 6 relevant features for three RBF-SVM models

Datasets	Acc. proposed RBF-SVM	Acc. Rand. RBF-SVM	Acc. "balanced" RBF-SVM	% relative difference for this proposal vs "Rand" model	% relative difference for this proposal vs "balanced" model
Binary GS	94,111	95,411	92,401	-1,381	1,817
Ito core	72,059	81,195	52,571	-12,678	27,045
LC-multiple	93,750	95,924	93,517	-2,319	0,249
Uetz screen	76,857	80,822	54,882	-5,159	28,592
Random Negative Dataset 1	72,211	38,353	6,537	46,888	90,947
Random Negative Dataset 2	71,951	37,937	37,444	47,274	47,959
$R_3^{test\_1}$	58,184	29,349	1,883	49,558	96,764
$R_3^{test\_2}$	63,596	30,150	1,882	52,591	97,041
$R_3^{test\_3}$	96,469	69,365	1,683	28,096	98,255
$R_3^{test\_4}$	62,221	31,061	1,522	50,080	97,554
$R_3^{test\_5}$	61,248	29,862	1,364	51,244	97,773
$R_3^{test\_6}$	64,992	33,120	1,702	49,040	97,381
$R_3^{test\_7}$	64,441	31,454	1,824	51,189	97,170
$R_3^{test\_8}$	94,705	67,821	1,702	28,387	98,203
$R_3^{test\_9}$	64,334	31,237	1,061	51,446	98,351

Acc. is the accuracy of RBF SVM model. *proposed RBF-SVM* is the SVM model trained using the training set formed by GSP set and the GSN set obtained using the proposed hierarchical clustering method. *Rand. RBF-SVM* is the SVM model trained using the training set where the GSN set was randomly selected. *"balanced" RBF-SVM* is the SVM model trained using the training set formed by GSP set and the GSN set obtained using the approach to create "balanced" negative set by Yu et al. [233] % relative difference is the percentage of relative difference using the "our proposal RBF-SVM" as basis.

8

CHARACTERIZATION OF THE PROTEIN INTERACTION NEIGHBOURHOOD OF E3 ENZYMES IN THE UBIQUITIN-PROTEASOME SYSTEM

8.1 Introduction

As previously commented, protein-protein interactions (PPIs) are of great importance in almost any biological function carried out within the cell. Given the enormous size and complexity of the interactomic data that is been gathered nowadays, with an estimated size over 130000 interactions potentially taking place within the human cell [216], a comprehensive view of the collection of all PPIs that take place in the cell (also known as interactome) has been proven imperative in order to get a proper understanding on how biological systems work. The use of bioinformatic tools is mandatory in order to extract information from interaction networks of this size. Of particular interest is the study of the topology of the network and its relevance in specific biological functions. In this chapter, an bioinformatic approach is applied in order to characterize a complex biological function such as controlled protein degradation.

Protein-protein interaction prediction using bioinformatic approaches based on topological studies have successfully been used [78, 3, 85] for several model organisms, even applied for human diseased [177, 125]. In this way, domain structure, considered as fundamental unit of interaction [163, 96], provides a suitable framework for PPI prediction [54, 96, 163, 129, 80]. Other approaches have inferred protein complexes from protein–protein interaction networks [240], even applied to Ubiquitin-Proteasome System network [215], or calculated biological relevance of PPIs [10] by the means of topological clustering analysis.

In any living cell, damaged and obsolete proteins are digested into their constitutive

aminoacid residues through controlled protein degradation, a tightly regulated, ATP-consuming and highly complex process that is mostly performed by the Ubiquitin-Proteasome System (UPS) [190]. Besides its obvious role in recycling protein construction material, this system has an important regulatory mechanism in many cellular events such as DNA repair [218, 219], cell cycle control [13, 209] or transcription activation [1, 241]. Moreover, alterations of the UPS have been implicated in the pathogenesis of many diseases, such as inflammatory and immunological disorders [190] or, more prominently, neurodegenerative diseases such as Parkinson's [126, 113], Huntington's [17] or Alzheimer's diseases [205].

The UPS performs its degradation function in two distinctive phases. As a first step, the ubiquitylation cascade is the responsible for marking degradation substrates with the small peptide ubiquitin so they can be digested into their constitutive parts and formed by ubiquitin-activating enzymes (E1), ubiquitin-conjugating enzymes (E2) and ubiquitin ligases (E3). In a second place, proteins marked with poly-ubiquitin chains are degraded in the proteasome, a multi-protein barrel with protease activity that digests the proteins into aminoacid residues that can be re-used in the cell.

In this chapter, it was aimed to characterize the relationship that ubiquitin ligases (E3s) have with their protein interaction partners in a systematic and unbiased approach. Using protein-protein interaction information found in publicly available databases such as Biogrid [199] and Bind [9], the specific interactome for each known E3 has been characterized and produced connectivity profiles for each E3 family (as classified by the E3-related domains they contain). On top of that, domain information of the interacting partners of the E3s has been extracted and researched whether families with comparable connectivity profiles interact with functionally close domains, thus trying to find out if similar functions are connected to the E3s in an analogous fashion. Ultimately, it is hoped that the information produced in this chapter will help identifying potential substrates for specific ubiquitin ligases.

The rest of the chapter is organized as follows. The section 8.2 makes a brief description of the Ubiquitin-Proteasome System (UPS). The section 8.3 describes the data used to develop the presented bioinformatic approach. The section 8.4 explains a new methodology for E3 characterization using neighbourhood analysis. The section 8.5 describes how characterizing E3 families using hierarchical clustering (clustergram) proposed in this chapter. The section 8.6 describes the results obtained using the proposed bioinformatic approach and a discussion about the results is also included. Finally, the conclusions are drawn from this chapter in the section 8.7.

8.2 Biological Background: Ubiquitin-Proteasome System

As stated above, there are two distinctive phases in the UPS-driven degradation process: (a) The ubiquitylation cascade, in which the conjugation of the small peptide ubiquitin (76 residues) to the target protein takes place in order to mark it for

degradation with a poly-ubiquitin chain; and (b) proteasomal degradation, in which the digestion of the ubiquitin-marked protein in the protease-complex known as the 26S-proteasome takes place, with the subsequent release of aminoacids and poly-ubiquitin chains that can be turned into free ubiquitin by de-ubiquitylating enzymes.

Due to its complexity and importance, the scope of the work of this chapter is focused on the first phase of the process. The enzyme cascade formed by E1s (ubiquitin activating enzymes), E2s (ubiquitin conjugating enzymes) and E3s (ubiquitin ligases) marks the proteins with poly-ubiquitin chains, as is depicted in figure 8.1. First, an E1 enzyme "activates" ubiquitin generating a high-energy thiol ester intermediate where ubiquitin is bound to an internal E1 Cys residue at the expense of one molecule of ATP. Then the activated ubiquitin is transferred to an E2 enzyme, from where it can be conjugated to a selected target with the assistance of an E3 ligase. There are roughly two ways in which such transfer can be made, depending on the type of E3 that recognizes and brings a particular substrate to the reaction: (a) the ubiquitin can be transferred directly from the E2 to the substrate, specifically bound to the E3, as happens in the RING type of E3s, or (b) the ubiquitin is transferred first to the E3 and then from there to the substrate, like in the HECT-type (Homologous to the E6-AP C-Terminus) ligases. Following addition of a single ubiquitin peptide to a protein substrate (mono-ubiquitination), further ubiquitin molecules can be added to the first, yielding a poly-ubiquitin chain. The poly-ubiquitin chains that target proteins for destruction are usually expanded through the Lys-48 (K-48) residue and at least four ubiquitin molecules must be attached to lysine residues on the condemned protein in order for it to be recognised by the 26S-proteasome. However, a variety of linkages involving all possible lysine residues have also been described, such as K-63 chains, not recognized by the proteasome and involved in processes such as DNA repair [125]. An important antagonistic mechanism are de-ubiquitylating enzymes (DUBs), that have the double role of counteracting the activity of the ubiquitylating cascade, removing poly-ubiquitin chains from already marked substrates before they enter the proteasome, and renew the free-ubiquitin pool by cutting down used poly-ubiquitin chains that come out of the proteasome [190, 198].

Given that the UPS accounts for most of the protein degradation that occurs in the cell and that specific degradation conditions vary from substrate to substrate, great specificity is required for proper function. This is accomplished by an increasing complexity of the system down the ubiquitylation cascade. There are a few dozens of E2 enzymes and in contrast hundreds of E3s that specifically recognize proteins when they are misfolded, phosphorylated or marked in some way. Although the signals that are used by E3s to recognize their substrates are not fully understood, it is known that phosphorylation can create what is called a "phosphodegron", a protein substrate specifically marked for E3 recognition and subsequent ubiquitylation [69, 157, 131]. This way, a degradation event can be controlled via a kinase cascade and modulated by its corresponding phosphatase. The specific set of regulatory elements and protein

interactions that control such type of events is known only in a few, well-researched examples. Moreover, it must be considered that a few hundreds of E3s cannot account for the enormous complexity that the regulation of specific degradation events requires.

Being by far the largest group of UPS-related proteins and the ones responsible of substrate specificity, the work developed in this chapter has been focused on ubiquitin ligases. As stated above, E3s provide specificity to the ubiquitin-conjugating reaction by recognizing the substrate and helping in the transference of ubiquitin from E2s. There are several families of E3s, classified by domain-specific information (see families and references in table 8.4. In the work developed in this chapter, any protein containing an E3-related domain is automatically considered a putative E3, even if there is no experimental evidence supporting its role as ubiquitin ligase. As some of the RING ligases work as multi-protein complexes (e. g., SCF-cullin complexes) that need the concurrence of several distinctive subunits to act as ubiquitin-ligases, some of the domains classified as E3s do not actually have ubiquitin ligase activity, but are linkers or structural components exclusive to multi-protein E3s [242, 58]. To have a brief overview on the different E3 families used in this work, see table 8.4.

Differences between E3 families may allow us to characterize the ubiquitylation reaction. This post-translational modification can only happen if a specific pattern of regulatory conditions (and thus, a specific set of protein-protein interactions) is given; so the neighbourhood of an E3-interacting partner could have a determinant influence in the interaction. The neighbourhood of a specific E3 interaction consists of all interacting proteins directly connected to the interacting pair. Some indirect interactions in the neighbourhood or a specific configuration of interactions between proteins could "define" the ubiquitylation reaction.

8.3 Material utilised in this chapter

In this chapter, as mentioned in the introduction, an bioinformatics approach is proposed for characterizing complex molecule associations. Specifically, the selected case in this work is the E3-machinery protein substrate relationship. Thus so, the material taken is composed by a protein-protein interacting network called master network, and a PPI dataset related to Ubiquitin-Proteasome System (UPS). Here the material listed in detail:

- It was created an integrated human protein-protein interaction network (MasterNet), by combining the PPI networks derived from major PPI databases, high-throughput studies and literature datasets (see table 8.1). This MasterNet is a comprehensive network of human protein-protein interactions, with 12510 proteins and 125958 interactions between them. All the edges in the MasterNet are equally weighted.

- Further, a subset of MasterNet focused around E3-ligases, E3 PPI dataset, was

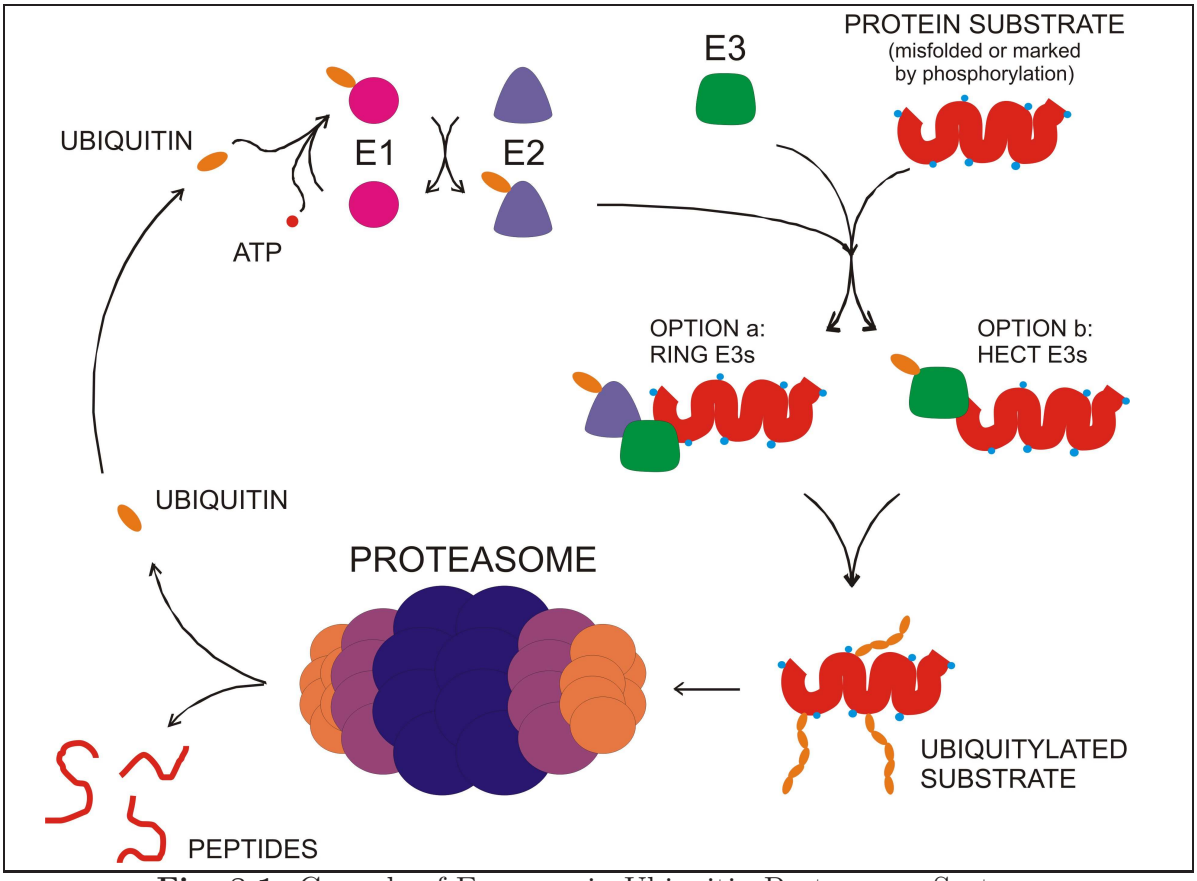


Fig. 8.1: Cascade of Enzymes in Ubiquitin-Proteasome System.

constructed for neighbourhood analysis.

In table 8.2, it can be observed the sizes of the data sets used in the work. For the PPI dataset, specifically 11918 E3-interacting partner interactions were considered, given the presented aim in the characterization of E3 machinery-protein substrate relationship.

The goal of this work proposed in this chapter is to develop a new methodology to perform a topological analysis of E3-interacting partners neighbourhood integrating the information about domains which are participating in these interactions. Structural protein investigators suggest the domain structure as fundamental unit of interaction [163, 96], supplying a suitable framework for protein-protein interactions prediction approaches. Thus, some approaches use domain-domain interactions to predict PPIs [54, 96, 116, 80]; or, in other cases, this data is combined with other kind of interaction information [163, 129]. This domain analysis is developed for the sake of providing more detailed data about E3 interactions in Ubiquitin Proteome System.

Table 8.1: List of datasets used to construct MasterNet

PPI datasets	Reference
Genome wide Y2H screen I	Stelzl et al., 2005 [202]
Genome wide Y2H screen II	Rual et al., 2005 [181]
Huntingtin's disease PPI network - Y2H screen	Goehler et al., 2004 [77]
Huntingtin's disease PPI network - Y2H screen and TAP	Kaltenbach et al., 2007 [112]
Y2H screen for inherited Ataxia	Lim et al., 2006 [127]
Repeated Y2H screening dataset	Venkatesan et al., 2009 [216]
Aging Network - Y2H screen	Bell et al., 2009 [14]
UPS network I	Sjoerd et al., 2009 [213]
UPS network II	Markson et al., 2009 [135]
Mouse signaling PPI data from AfCS (mapped to Human genes)	http://www.signaling-gateway.org
Network for Smad signaling	Colland et al., 2004 [41]
PPIs of proteins in MHC class III region and mRNA decay	Lehner et al., 2004 [121]
Network of Nuclear receptors	Albers et al., 2005 [2]
PPIs between KIAA proteins	Nakayama et al., 2002 [150]
Integrin adhesome network - literature curated	Zaidel-Bar et al., 2007 [236]
Large scale immunoprecipitation /mass spectrometry network	Ewing et al., 2007 [63]
TransMac affinity purification/mass spectrometry	Jeronimo et al., 2007 [108]
HPRD database	Keshava Prasad et al., 2009 [115]
Reactome database	Matthews et al., 2009 [139]
BioGrid database	Stark et al., 2006 [199]
IntAct database	Aranda et al, 2010 [5]
Mint database	Ceol et al., 2010 [38]
DIP database	Salwinski et al., 2004 [186]
MIPS database	Pagel et al., 2005 [160]
BIND database	Bader et al., 2001 [9]
PDZBase database	Beuming et al., 2005 [21]
I2D database	Brown et al., 2005 [33]
Database of Cell Signaling	http://stke.sciencemag.org/cm/

Table 8.2: Datasets

E3 PPI dataset	11918
Master PPI Net	125958

The domain information has been extracted from the well-known database Swisspfam [24]. Swisspfam is a compendium of domain structures through SWISSPROT and TrEMBL [24] databases and according to domains from the Pfam database [64]. Pfam stores protein domain families including annotations and alignments of multiple sequence generated using Hidden Markov Models [107].

8.4 A new methodology for E3 Characterization: Neighbourhood Analysis

Topological investigations have been used successfully in protein-protein interaction prediction [78, 3, 85] including even detection of human genes causing diseases [177, 125]. In this chapter, a new approach is proposed analysing the neighbourhood created by E3 enzyme interactions, and providing new measures in order to quantifying and give meaning to the connectivity rates in such interacting networks.

In this way, given an E3-interacting partner interaction, its neighbourhood is composed by those direct interactions of E3 enzyme and the interactions of the interacting partner (see figure 8.2). Thus, some interactions can be shared between the E3 enzyme and the interacting partner. The larger the number of shared connections, the more "connected" the interaction is.

Under these assumptions, it was performed the connectivity quantification of every E3-interaction partner pair, classifying interactions in three neighbourhoods according to which neighbour is interacting directly with a given E3 enzyme, its interacting partner or both. Therefore, given a pair $(P_i, P_j) \in PPI_{E3}$ being the pair of interacting proteins in the set PPI_{E3} from the E3 PPI dataset (see section material) composed by 11819 E3 enzyme-interacting partner interactions, the used sets were defined including the considered neighbourhoods as follows:

- The set of common neighbours is denoted by $S^2_{P_{ij}} = \{P^1_{ij}, ..., P^h_{ij}, ..., P^{n_2}_{ij}\}$ being n^2_{ij} is the number of common neighbours for the giver pair (P_i, P_j).

- $S^1_{P_i} = \{P^1_i, ..., P^m_i, ..., P^{n^1_i}_i\}$ is the set composed by the neighbours (interacting proteins) of P_i (the E3 enzyme) which are not included in the Common Neighbourhood $S^2_{P_{ij}}$, being n^1_i is the number of elements in $S^1_{P_i}$.

- The set of neighbours of P_j (the interacting partner) could be expressed as $S^3_{P_j} = \{P^1_j, ..., P^p_j, ..., P^{n^3_j}_j\}$ which are not included in the $S^2_{P_{ij}}$, being n^3_j is the number of elements in $S^3_{P_j}$.

- Neighbourhood I is denoted by $\Phi^1_{P_i} = S^1_{P_i} \cup S^2_{P_{ij}}$. This set is composed by the direct neighbours to the E3 enzyme P_i.
- Neighbourhood-III is denoted by $\Phi^3_{P_j} = S^3_{P_j} \cup S^2_{P_{ij}}$. This set is formed by the direct neighbours to the interacting partner P_j.

These terms are represented graphically in figure 8.2. Note that neighbourhoods are not exclusive, and so proteins in Common Neighbourhood can be part of Neighbourhood I or Neighbourhood III.

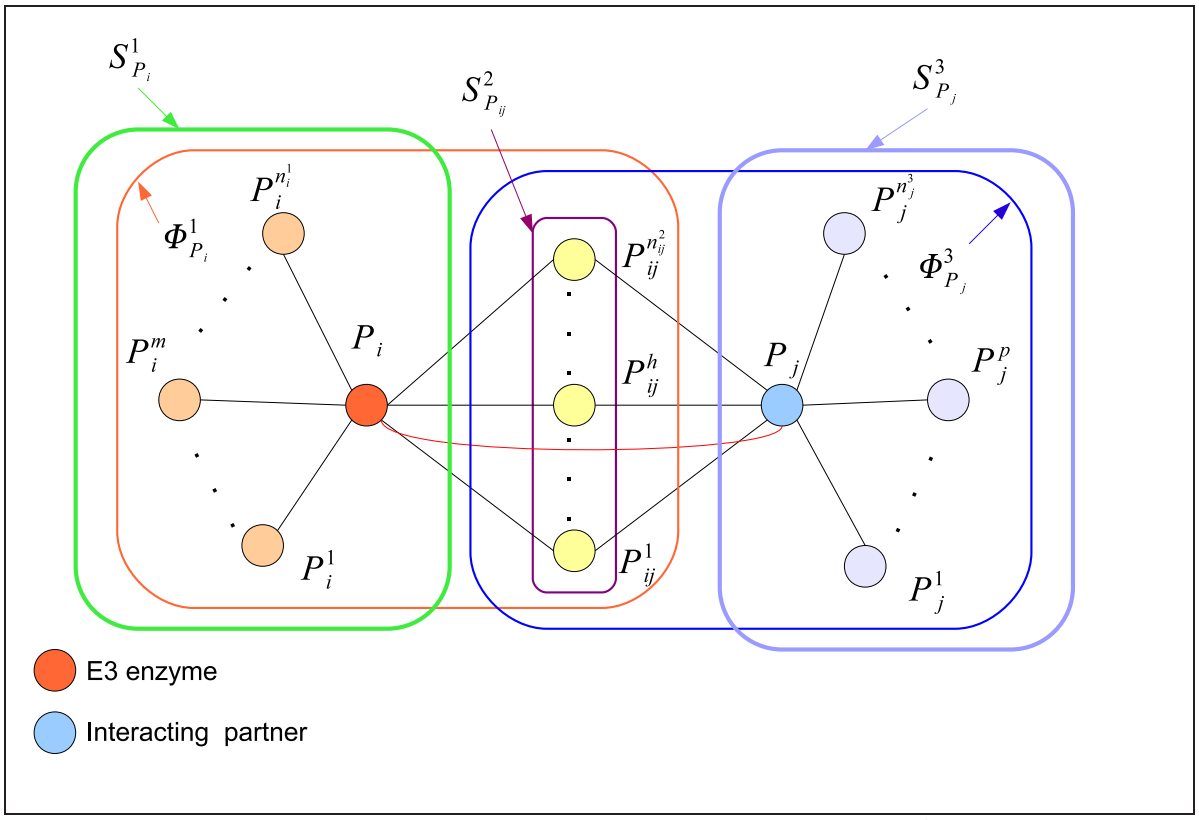


Fig. 8.2: The considered pair is (P_i, P_j). Neighbourhood I ($\Phi^1_{P_i}$) for E3 enzyme, Neighbourhood III ($\Phi^3_{P_j}$) for Interacting Partner and Common Neighbourhood $S^2_{P_{ij}}$. Red line is the direct interaction in E3-Interacting Partner (P_i, P_j). $S^{n_i^1}_{P_i}$ is set composed by the n_i^1 neighbours of P_i (E3 enzyme) which are not included in $S^2_{P_{ij}}$ (the set of common neighbours). The set of neighbours for P_j is $S^{n_j^3}_{P_j}$.

In order to quantify the connectivity of an E3-interacting partner pair (P_i, P_j), a measure called $neighb_{P_iP_j}$ is proposed. It can be observed that connectivity rate of a pair is fundamentally determined by $S^2_{P_{ij}}$ (Common Neighbourhood), i.e., according to

the number of common elements in that pair. Consequently, $neighb_{P_iP_j}$ is calculated as the number of proteins in $S^1_{P_i}$ (n^2_j) divided by the number of E3 enzyme neighbours ($n^1_i + n^2_j$), where the lowest value is 0 and the highest 1. Thus, $neighb_{P_iP_j}$ could be expressed as:

$$neighb_{P_iP_j} = \frac{n^2_{ij}}{n^1_i + n^2_{ij}} \tag{8.1}$$

8.5 Characterizing E3 families using hierarchical clustering

The characterization of E3 enzyme families is developed through topological analysis of the E3 enzyme-interacting partner neighbourhoods. At the same time, the domains of interacting proteins were incorporated into this analysis, designing connectivity profiles of E3 enzyme domain families, in an effort to identify potential ubiquitylation substrates.

First, and in order to establish connectivity profiles of each E3 domain family, it is designed a classification algorithm of E3-interacting partner pairs into E3 families connectivity groups using the obtained $neighb_{P_iP_j}$ values. The number of connectivity groups is fixed and equal for all E3 families, and it has to be selected previously. This algorithm classifies E3 PPI according to E3 domain family, and then obtains the neighbours of each pair through the master network. Subsequently, the $neighb_{P_iP_j}$ values of each pair are calculated for each E3 family and each pair is classified according to its connectivity within each family. These steps are detailed in mnemotechnic description in algorithm 8.1.

Nevertheless, this first step is not enough to obtain connectivity profiles that can identify which type of substrates are involved with each E3 family. Since families are defined by the type of E3-related domain that the proteins within them bear, the domains of their interacting partners were incorporated to the analysis. Such domains are extracted from both the E3 enzymes and the interacting partner using SwissPfam databases [24] as previously commented (see Materials section).

With the intention of filtering out those non-significant and non-relevant domains extracted from all the PPIs involved, a biological enrichment process is applied. Such process is fundamentally based on the construction a statistical distribution. This distribution is calculated using the domain frequencies obtained in randomly interaction neighbourhood for each E3 enzyme within each family. Subsequently, a p-value is assessed [188]. In this case, this value means the probability of randomly appearance for each domain in the distribution. Domains with p-value < 0.05 are considered significant from the biological point of view. The process of calculating the distribution is described in algorithm 8.2, and the complete process of domain extraction and biological enrichment process is detailed in algorithm 8.3.

One of the aims of this approach is to find coincidental interaction patterns between E3 families. Using this domain analysis, a new measure is proposed to estimate the

Read M E3-interacting partner pairs from dataset file
Read all PPIs from Master Network
for $r = 1$ **to** M **do**
  Extract the neighbours from Master Net for the E3 enzyme $P_i := \Phi^1_{P_i}$.
  Extract the neighbours from Master Net for the interacting partner $P_j := \Phi^3_{P_j}$.
end for
for $r = 1$ **to** M **do**
  Calculate n^2_{ij} common neighbours between $\Phi^1_{P_i}$ and $\Phi^3_{P_j} = S^2_{P_{ij}}$.
  Calculate $neighb_{P_iP_j}$ for each pair (P_i, P_j).
end for
for $r = 1$ **to** M **do**
  Classify PPI according to its E3 enzyme P_i into E3 family.
end for
{ for each E3 family and for each PPI in that E3 family}
for $s = 1$ **to** O **do**
  for $t = 1$ **to** Q **do**
    Clustering PPI using $neighb_{P_iP_j}$
    {Establish a number of connectivity groups from lowest connectivity to highest, e.g., for 4 groups, classify pairs in $G = \{g_1, g_2, g_3, g_4\}$.}
  end for
end for

Note: M is the number of E3-interacting partner interactions in dataset (in this case 11918 interactions). O is the number of E3 families (in this case 15 families). Q is the number of E3-interacting partner interactions in a specific E3 family.

Algoritmo 8.1: Grouping E3 enzyme-interacting partner interactions using $neighb_{P_iP_j}$

relative abundance of a specific domain d_i in a connectivity group (g_j) within E3 families. Such measure called g_{value}, it is calculated for each domain d_i within E3 family for a specific connectivity group (g_j). Therefore $g_{value}(d_k, g_l)$ is calculated as the division of such domain(d_i) frequency in a specific connectivity group (g_j) divided by sum of all specific domain frequencies from all connectivity groups. The $g_{value}(d_k, g_l)$ is expressed as equation 8.2:

$$g_{value}(d_k, g_l) = \frac{f_{d_k} \in g_l}{\sum_{k=1}^{F} f_{d_k} \in G} \tag{8.2}$$

where the f is the frequency. The d_k is the specific domain in a specific connectivity group l. G is the set of all connectivity groups (g_l). F is the number of connectivity groups.

This way, $g_{value}(d_k, g_l)$ can be used in the construction of clustergrams, one for

SET $K := 10000$
for $r = 1$ **to** K **do**
  for $s = 1$ **to** Q **do**
    for $t = 1$ **to** D **do**
      Create a random interacting network for E3 enzyme P_i. {remove all the interactions keeping only E3 enzyme P_i from selected s PPI (P_i, P_j)}
      Calculate the new frequencies for the found domains in the new interacting pair.
    end for
  end for
end for
for $r = 1$ **to** O **do**
  for $d = 1$ **to** D **do**
    Calculate the p-value for domain d using previous distribution.
  end for
end for

Note: O is the number of E3 families. Q is the number of E3-interacting partner interactions in a specific E3 family. D is the number of all domains in E3 families. K is set to 1000 interactions to calculate a distribution.

Algoritmo 8.2: Biological Enrichment Process

each connectivity group. Clustergrams are quite common tools in bioinformatics and represent a hierarchical clustering using a dendrogram and heat map simultaneously [12, 79, 60, 55]. By the means of these clustergrams, It can be checked graphically the E3 domain families clusters according to connectivity and extract information searching for interaction patterns. All algorithms were implemented in Matlab ®(2010a, The MathWorks, Natick, MA).

8.6 Results and discussion of the proposed approach

The experimental results are composed of two different parts: grouping E3 families according to $neighb_{P_iP_j}$ and characterization of E3 families including domains through clustergrams. The E3-interacting partner PPI dataset was used as the data source for all experiments.

First, a connectivity profiling for E3 enzyme families was carried out through neighbourhood analysis, as previously explained in the section the Methods. The $neighb_{P_iP_j}$ values are calculated in order to group E3 families through the algorithm 8.1. The number of connectivity groups was established to 4 in order to simplify the interpretation of the results. Using the values $neighb_{P_iP_j}$ in range [0,1], the connectivity intervals for each group are:

Read PPI dataset
Reading SwissPfam DB
Classifying E3 families using $neighb_{P_iP_j}$ in connectivity groups (see previous section)

for $r = 1$ **to** O **do**
  for $s = 1$ **to** F **do**
    for $t = 1$ **to** H **do**
      Extraction of domains for E3 enzyme
      Extraction of domains for interacting partner
      Calculate the frequencies for the found domains in the pair.
    end for
  end for
end for{see previous process}
Enrichment process for each connectivity group in E3 family (randomly distribution of domain frequencies)
Calculate p-values for found domains for each group in E3 family

Note: O is the number of E3 families. H is the number of E3-interacting partner interactions in a specific E3 family in a specific $neighb_{P_iP_j}$ group. F is the number of connectivity groups using $neighb_{P_iP_j}$.

Algoritmo 8.3: Full Procedure to extract domain features

- Low connectivity (G_1 or group 1): $[0.0 - 0.25)$.
- Fair connectivity (G_2 or group 2): $[0.25 - 0.5)$.
- Medium connectivity (G_3 or group 3): $[0.5 - 0.75)$.
- High connectivity (G_4 or group 4): $[0.75 - 1.0]$.

For a better understanding of the 4 connectivity groups, some conceptual examples were represented graphically from lowest connected to highest connectivity in figures 8.4(a), 8.4(b), 8.4(c), 8.4(d). The results of E3 family connectivity grouping, showing graphically using percentages of E3 enzyme-interacting partner interactions into each connectivity group using $neighb_{P_iP_j}$ are represented in figure 8.3 and table 8.3.

As shown on figure 4 and table 8.4, the majority of the families characterized here show predominance of low connectivity relationships with their interacting partners (11 out of 15 families). Most of these families, with the exception of the cullin, F-box and skp1-like domain families, are believed to have a substrate-recognition function. This suggests that a complete absence or a low number of common interacting partners probably characterizes the substrate-E3 relationship. Those families lacking dominance of the low connectivity group are either present in well-known multiprotein complexes

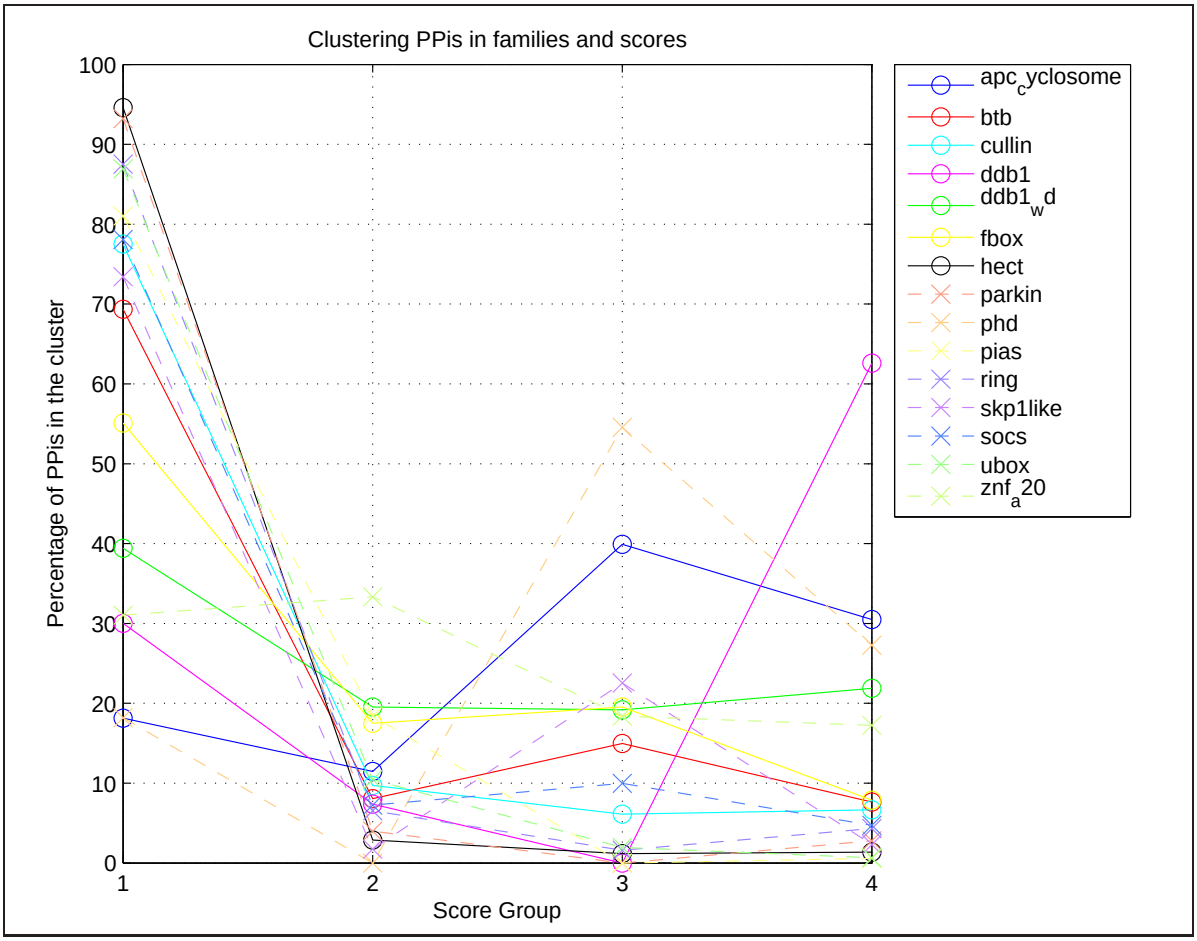


Fig. 8.3: Percentage of PPIs (E3 enzyme-interacting partner) grouping using $neighb_{P_iP_j}$ per E3 family.

with a relatively large proportion of common interacting partners, such as the APC-cyclosome domain [151, 189], domain families whose affiliation to the E3-ligase function is still under dispute (like the PHD domain family, [102, 211]) or domains known to be present in tightly-connected proteins with double functionality (A20-zinc-finger domain, 5). Nevertheless, it is obviously not possible to predict the potential function of an E3 domain family just using these rudimentary connectivity profiles, as there are families known to be structural components of multi-protein complexes that show low connectivity values (cullin, F-box, skp1-like domains).

In the second part, domains of used interacting proteins were incorporated, and calculate for all domains the $g_{value}(d_k, g_l)$ within each connectivity group of each family as explained in previous sections. The $g_{value}(d_k, g_l)$ is normalised in range [-1,1] for clustergram analysis. In clustergram, Y axis represent a domain and its $g_{value}(d_k, g_l)$, and X axis represent E3 families grouped through hierarchical clustering. For the sake

Table 8.3: Percentage of E3 enzyme-interacting partner interactions into connectivity groups

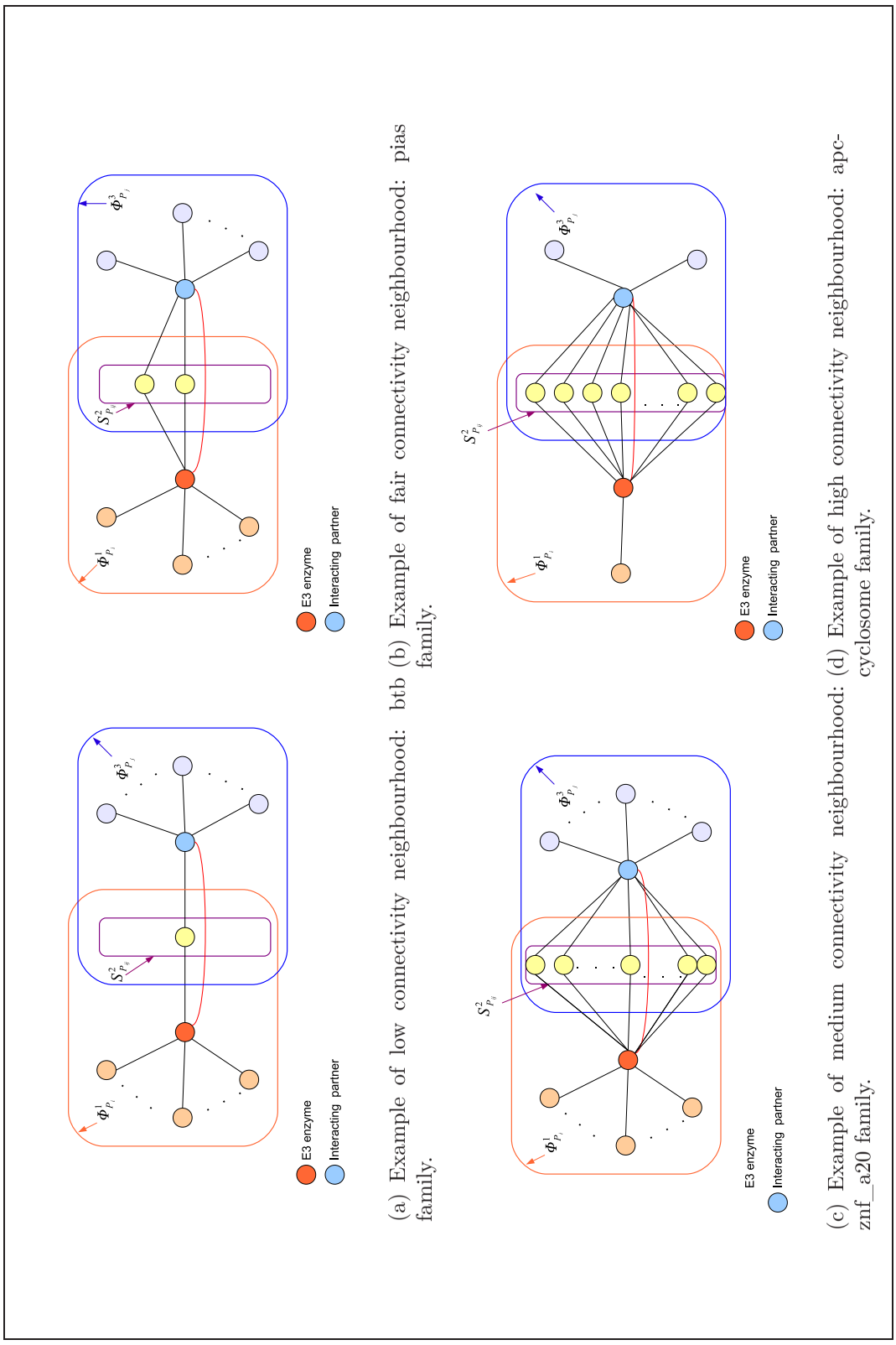
E3 enzyme families	Group 1	Group 2	Group 3	Group 4
apc_cyclosome	18.142	11.466	39.913	30.479
btb	69.350	8.049	14.959	7.642
cullin	77.500	9.722	6.111	6.667
ddb1	30.028	7.365	0.000	62.606
ddb1_wd	39.421	19.530	19.168	21.881
fbox	55.102	17.493	19.534	7.872
hect	94.595	2.872	1.182	1.351
parkin	93.227	3.984	0.000	2.789
phd	18.182	0.000	54.545	27.273
pias	81.034	18.391	0.000	0.575
ring	87.524	6.538	1.609	4.330
skp1like	73.410	1.734	22.543	2.312
socs	78.080	7.246	9.964	4.710
ubox	86.928	10.458	1.961	0.654
znf_a20	31.034	33.333	18.391	17.241

Note: Percentage of PPIs (E3 enzyme-interacting partner) grouping into connectivity groups (group 1 for g_1, group 2 for g_2, group 3 for g_3, group 4 for g_4) using $neighb_{P_iP_j}$ per E3 family.

of clarity, the domains names were removed for X axis labels. The colour represent the intensity of $g_{value}(d_k, g_l)$, it changes from blue for lowest value to dark red for highest value.

Therefore, in order to further enrich the proposed characterization, information about which domain families were more frequently present in the interacting partners of each family domain was used to build specific clustergrams for each one of the connectivity groups. This way, it is aimed to find out whether families with similar connectivity profiles share similar target domains or domains with similar functions. A summary on which domains are forming most prominent clusters in each connectivity group can be found in tables 8.5, 8.6 and 8.7. A brief description on each domain can be found in the Pfam website. As can be seen in tables 8.5, 8.6 and 8.7, and in the interacting domain network 8.9, clusters clearly connecting E3 families are found in connectivity groups 2, 3 and 4. Remarkably, E3 domain families associated in those clusters are all known to participate in multi-protein E3 ligase complexes (with the exception of the u-box – btb connection, which has not been previously reported in the literature). This way, the connection between E3 families derived from the clustering results could be describing the functional link existing between those families, known to be interacting partners in several multi-protein E3 complexes.

Additionally, the domain families present in the interacting partners found in the clusters tend to have similar cellular functions; a tendency that seems more evident in those families grouped under the highest connectivity groups (specially 3 and 4, see families socs + skp1-like or ring + ddb1_wd + ddb1; but also in 2, see both ddb1 + ring and ubox + ddb1_wd). For example, in the cluster relating the ring, ddb1_wd and the ddb1 E3 families in the high connectivity group (4), it could be found a clear predominance of domains involved in processes such as transcription, splicing regulation or RNA metabolism. All these functions are known to be heavily regulated in tightly connected interaction motifs, involving large and complex protein machineries such as RNA polymerases or the spliceosome. Thus, is not surprising finding here a large proportion of common interacting domain families connecting two or more potential pieces of such a fundamental regulatory element as an E3 ligase.



(a) Example of low connectivity neighbourhood: btb family.

(b) Example of fair connectivity neighbourhood: pias family.

(c) Example of medium connectivity neighbourhood: znf_a20 family.

(d) Example of high connectivity neighbourhood: apc-cyclosome family.

Fig. 8.4: several grades of connectivity examples, from lowest to highest (a,b,c,d)

Table 8.4: UPS Families Connectivity

E3 domain families	Connectivity profile predominance	References
hect	1	Pickart, 2001 [168]; Scheffner et al., 1995 [187] ; Pickart et al., 2004 [169]; Metzger & Weissman, 2010 [144]
Binds to E2 and acts as ub-ligase, direct transfer of poly-ub chain to substrate.		
ring	1	Metzger & Weissman, 2010 [144]
Can bind to e2s and act as ub-ligases. Subtypes too.		
pias	1-2	Rytinki et al, 2009 [183]
ring subtype		
parkin	1	Xiong et al. 2009 [231]
ring subtype		
ubox	1-2	Aravind, 2000 [6]; Pringa, 2001 [172]; Ohi et al., 2003 [155]; Hibbert et al, 2009 [90]
Extremely degenerated ring domain, still can adopt a ring-like structure.		
znf_a20	2-1-3-4	Wertz et al., 2004 [225], Shembade e al., 2010 [193]
Ub-ligase. Found in protein with OTUB domain (DUB function), double functionality and tightly connected.		
phd	3-4-1	Uchida, 2004 [211]; Ivanov, 2007 [102]
Dubious E3, due to difficulties to identify domain. Involved in one case with SUMOylation (see Ivanov reference [102]).		
E3 complexes		
SCF complexes		Willems, 2004 [226]; Petroski & Deshaies, 2005 [167]; Jackson & Xiong, 2009 [103]
btb	1-3	
Substrate-binding, binds directly to cullin.		
ddb1	4-1	
Adaptor for WD40-containing proteins, binds to cullin directly.		
ddb1_wd	1-4-2-3	
WD40-containing proteins, aka DWDs, substrate-binding, binds to cullin through ddb1.		
fbox	1-3-2	
Present in substrate-binding proteins, binds to skp1-like.		
socs	1	
ate-binding, binds to elongin adaptors (that bear skp1-like or ub-like domains).		
skp1-like	1-3	
Adaptor for F-Box proteins.		
cullin	1	
Scaffold. Binds to RING-domain protein and to E2.		
apc_cyclosome	3-4-1-2	Zachariae, 1998 [235]; Schreiber e al, 2011 [189]
Specific subtype of SCF complex.		

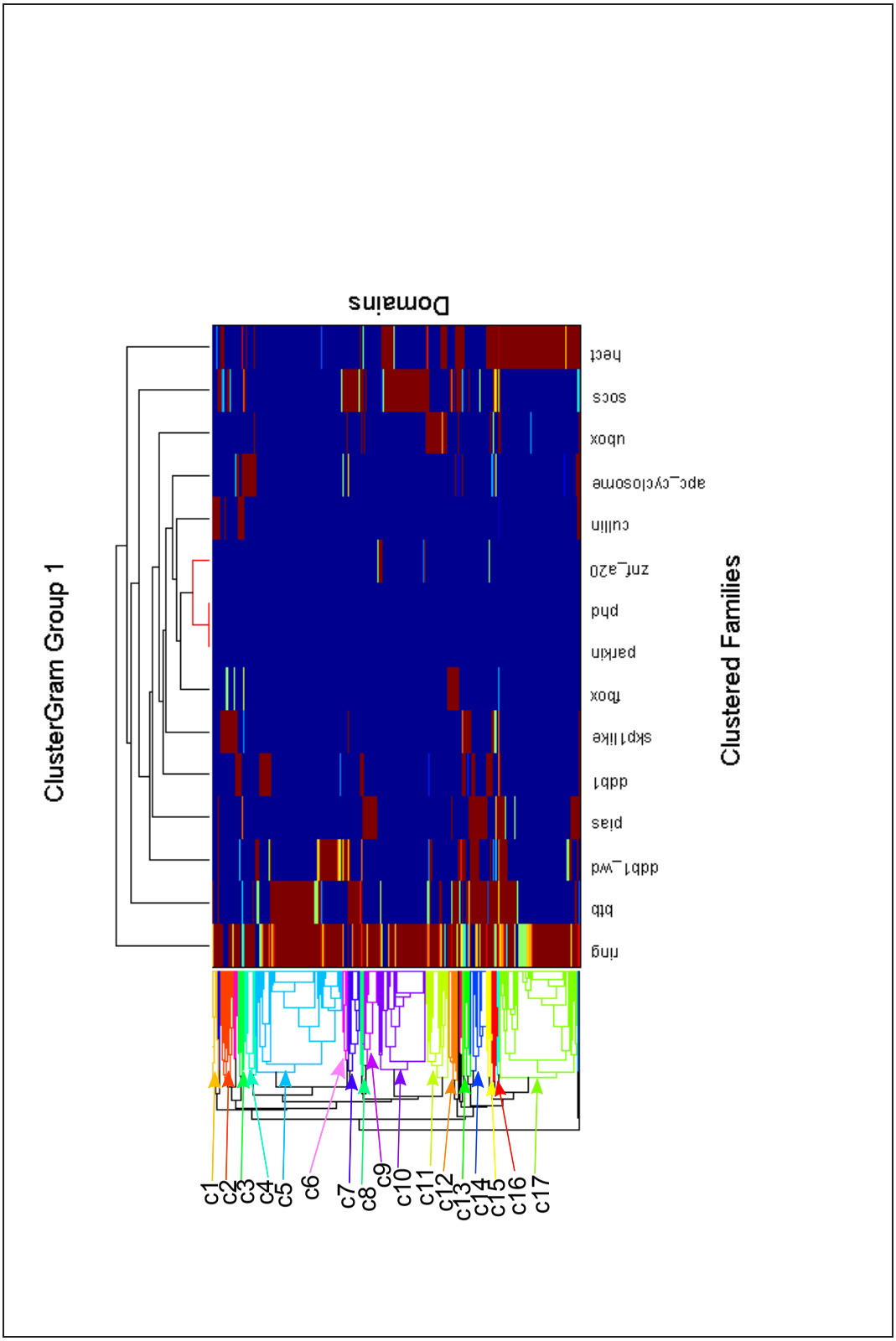


Fig. 8.5: Clustergram for G1 (Group 1) lowest connectivity group. Note that clusters are named as c+number

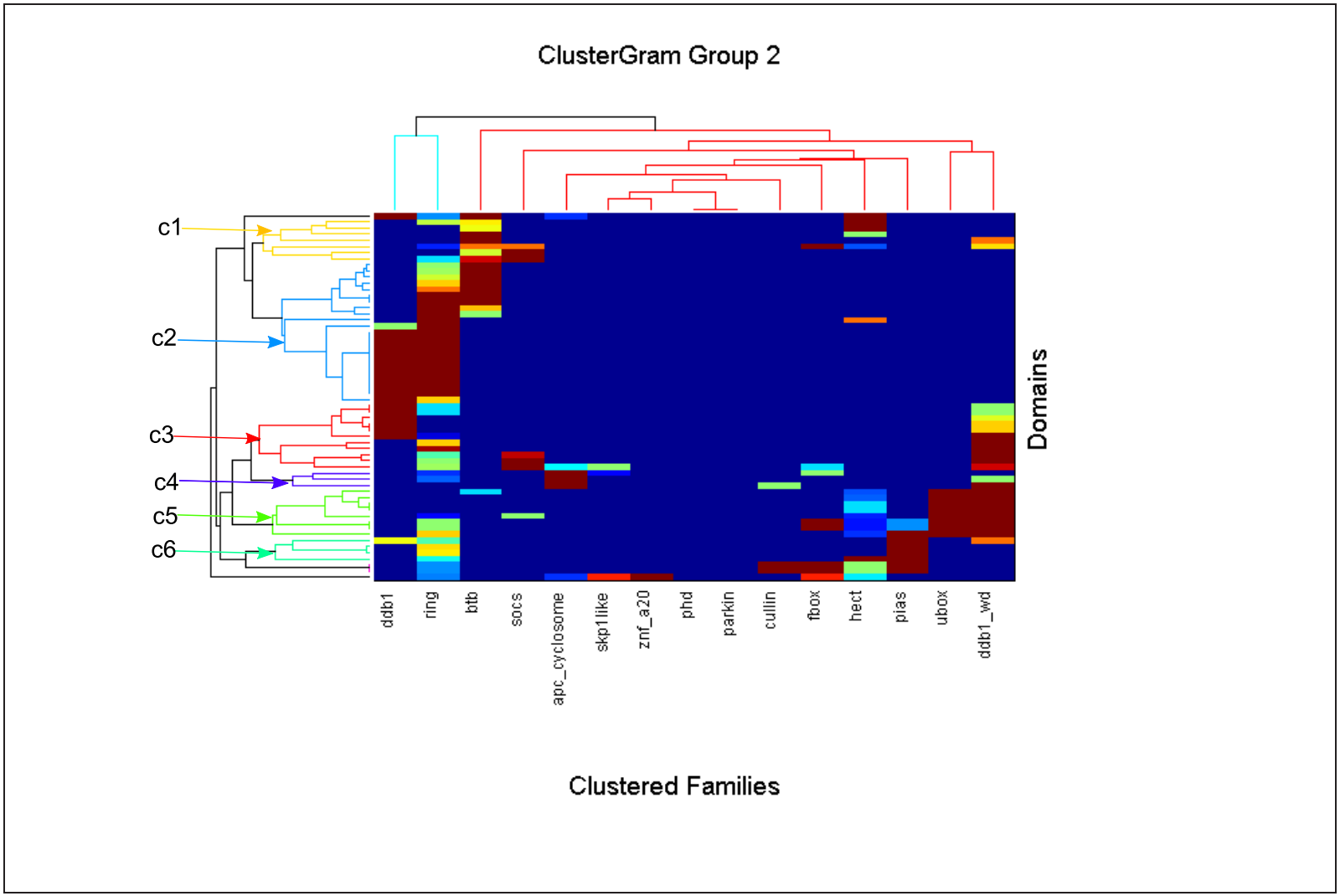


Fig. 8.6: Clustergram for G2 (Group 2), fair connectivity group. Note that clusters are named as c+number

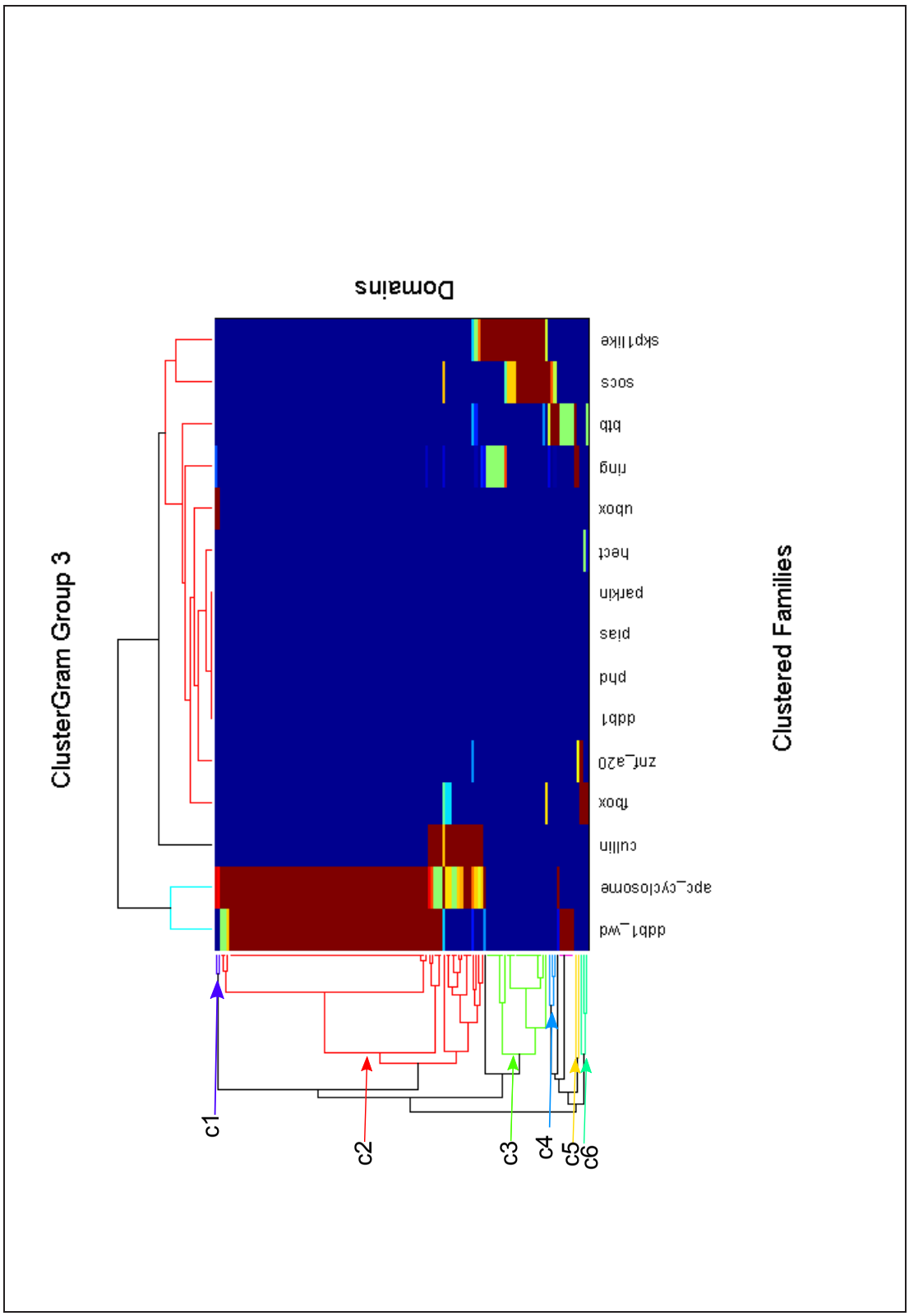


Fig. 8.7: Clustergram for G3 (Group 3), medium connectivity group. Note that clusters are named as c+number

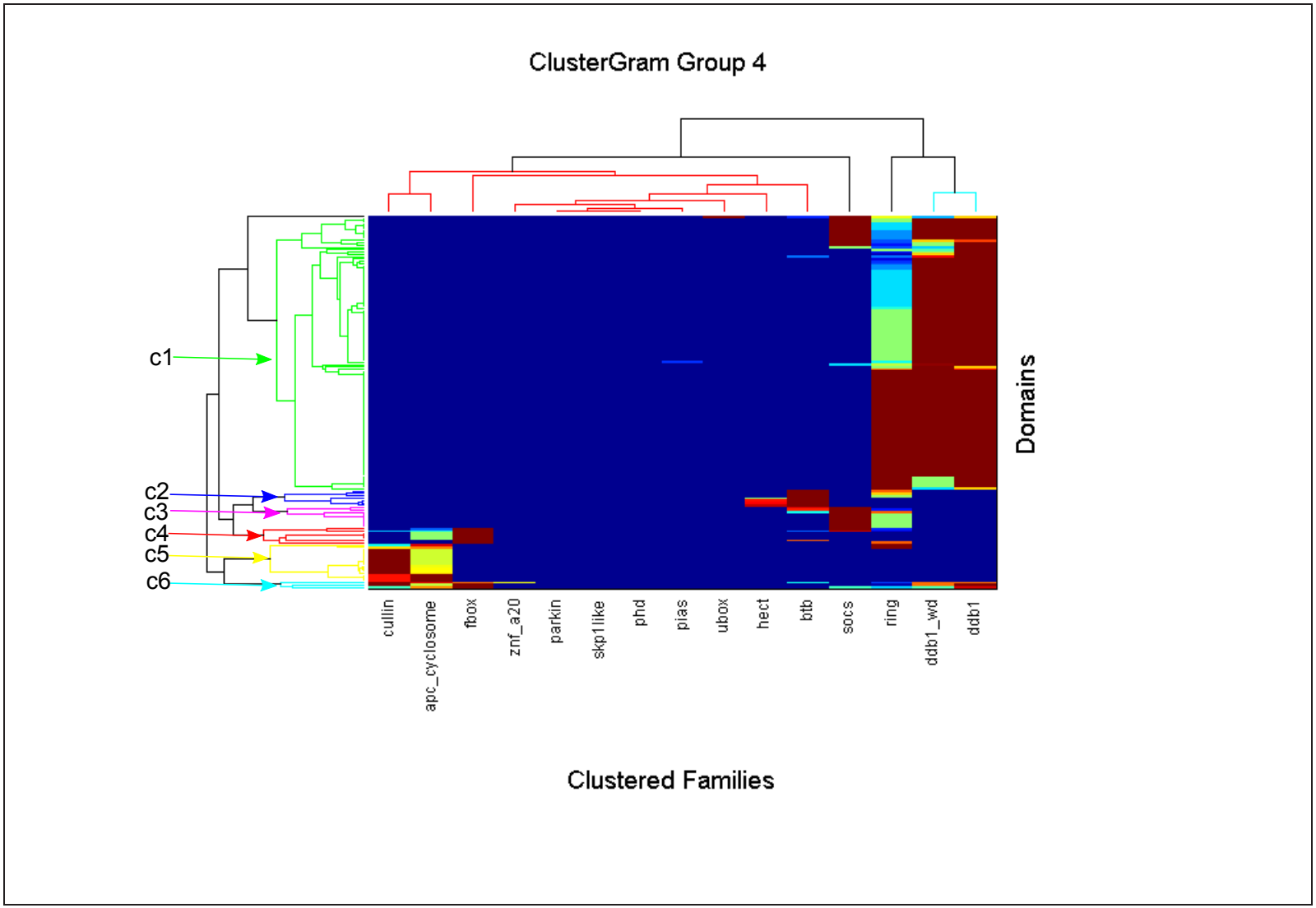


Fig. 8.8: Clustergram for G4 (Group 4), highest connectivity group. Note that cluster are named as c+number

Table 8.5: Domain Function Grouping (Clustergram 2)

CG	Connected E3 domain families	cluster	Interacting domain families functional affiliation	# CIDF	Relevant Domains (Pfam Name)
2	ddb1 + ring	2	DNA metabolism	3	ERCC4, Rep_fac-A_3, RPA_C
2	ddb1 + ring	2	DNA repairment	3	Rad10, XPG_I, XPG_N
2	ddb1 + ring	2	DNA/RNA binding, transcription control	1	Homeobox
2	ddb1 + ring	2	Nucleotide metabolism	2	APOBEC_C, APOBEC_N
2	ddb1 + ring	2	PPI, RNA binding	1	SAM_1
2	ddb1 + ring	2	PPI, RNA binding, DNA binding	1	zf-C2H2
2	ddb1 + ring	2	PPI, UPS	1	WWE
2	ddb1 + ring	2	Signal transduction, transport	1	efhand
2	ddb1 + ring	2	Structural, motor	1	Myosin_head
2	ddb1 + ring	2	Structural, phosphorylation, DNA metabolism	1	Exo_endo_phos
2	ddb1 + ring	2	Transcription	2	RNA_pol_L, Tap-RNA_bind
2	ddb1 + ring	2	Unknown function	5	Agenet, DGCR6, Pfam-B_16688, Pfam-B_17343, Pfam-B_83527
2	ubox + ddb1_wd	5	PPI	1	MH2
2	ubox + ddb1_wd	5	PPI, transcription control, nuclear import	1	MH1
2	ubox + ddb1_wd	5	Protein binding	2	WW
2	ubox + ddb1_wd	5	Signal transduction	2	C2
2	ubox + ddb1_wd	5	Unknown function	3	Pfam-B_220427, Pfam-B_61665, Pfam-B_842
2	ubox + ddb1_wd	5	UPS	1	HECT

Note. CG: clustergram. CIDF: common interacting domain families. For cluster see clustergram figure 8.6, cluster is named as c+number.

Table 8.6: Domain Function Grouping (Clustergram 3)

CG	Connected E3 domain families	cluster	Interacting domain families functional affiliation	# CIDF	Relevant Domains (Pfam Name)
3	ddb1_wd + apc_cyclosome	2	Apoptosis regulation	1	BIR
3	ddb1_wd + apc_cyclosome	2	Cell cycle control	6	Cyclin_C, Cyclin_N, Mad3_BUB1_I, MAT1, M-inducer_phosp, Mis12_component, BRCT
3	ddb1_wd + apc_cyclosome	2	Cell cycle control, DNA repairment	1	Securin
3	ddb1_wd + apc_cyclosome	2	Cell cycle control, UPS	1	APC2
3	ddb1_wd + apc_cyclosome	2	Chaperoning	1	CS
3	ddb1_wd + apc_cyclosome	2	DNA metabolism	1	HORMA, ResIII
3	ddb1_wd + apc_cyclosome	2	DNA metabolism, transcription	1	SNF2_N
3	ddb1_wd + apc_cyclosome	2	Gene expression regulation	2	NRIF3, RCC1
3	ddb1_wd + apc_cyclosome	2	Mitosis	11	Borealin, CHL4, INCENP_ARK-bind, Mis12, Mis14, Mis6, Nnf1, Nuf2, Rod_C, Spc24, Zw10
3	ddb1_wd + apc_cyclosome	2	Mitosis, meiosis	1	Shugoshin_N
3	ddb1_wd + apc_cyclosome	2	Nuclear migration	1	NUDE_C
3	ddb1_wd + apc_cyclosome	2	Nuclear trafficking	1	RanGAP1_C
3	ddb1_wd + apc_cyclosome	2	Organelle transport	1	Kinesin
3	ddb1_wd + apc_cyclosome	2	Phosphatase	1	Metallophos
3	ddb1_wd + apc_cyclosome	2	Phosphatase, UPS	1	Rhodanese
3	ddb1_wd + apc_cyclosome	2	Phosphorylation	1	Pkinase
3	ddb1_wd + apc_cyclosome	2	PPI	2	HEAT, POLO_box
3	ddb1_wd + apc_cyclosome	2	PPI, other	1	WD40
3	ddb1_wd + apc_cyclosome	2	Ribosome	1	Ribosomal_S27e
3	ddb1_wd + apc_cyclosome	2	Structural	1	CAP_GLY
3	ddb1_wd + apc_cyclosome	2	Unknown function	52	B9, DUF1395, DUF2352, Pfam-B_1040, Pfam-B_104975, Pfam-B_113070, Pfam-B_119407, Pfam-B_127828, Pfam-B_128, Pfam-B_12908, Pfam-B_14122, Pfam-B_14677, Pfam-B_150035, Pfam-B_16386, Pfam-B_18193, Pfam-B_185392, Pfam-B_19006, Pfam-B_19083, Pfam-B_194134, Pfam-B_1951, Pfam-B_19539, Pfam-B_20531, Pfam-B_205532, Pfam-B_2324, Pfam-B_26096, Pfam-B_30963, Pfam-B_36445, Pfam-B_36494, Pfam-B_38082, Pfam-B_38981, Pfam-B_4027, Pfam-B_40393, Pfam-B_40536, Pfam-B_4173, Pfam-B_45622, Pfam-B_4710, Pfam-B_4780, Pfam-B_48956, Pfam-B_51890, Pfam-B_54808, Pfam-B_63159, Pfam-B_63718, Pfam-B_64986, Pfam-B_70135, Pfam-B_70389, Pfam-B_7125, Pfam-B_71486, Pfam-B_74745, Pfam-B_8002, Pfam-B_84039, Pfam-B_89131, Pfam-B_91842
3	ddb1 + apc_cyclosome	2	UPS, RING domain	1	zf-C3HC4
3	socs + skp1-like	3	Cell cycle control	1	BRCT
3	socs + skp1-like	3	DNA metabolism	1	ResIII
3	socs + skp1-like	3	DNA repairment	1	TFIIH_BTF_p62_N
3	socs + skp1-like	3	Nucleic acids packing	1	Helicase_C
3	socs + skp1-like	3	Protease	1	Peptidase_M24
3	socs + skp1-like	3	Transcription	5	FCP1_C, KOW, NIF, SPT16, Supt5
3	socs + skp1-like	3	Transcription regulation	2	BSD, TFIIS
3	socs + skp1-like	3	Transcription regulation, cell cycle control	1	COBRA1
3	socs + skp1-like	3	Unknown function	7	DUF1227, Pfam-B_1252, Pfam-B_15002, Pfam-B_2486, Pfam-B_7807, Pfam-B_9184, Rtt106
3	socs + skp1-like	3	Unknown function, transcription regulation?	1	TH1
3	socs + skp1-like	3	UPS	1	Skp1_POZ

Note. CG: clustergram. CIDF: common interacting domain families. For cluster see clustergram figure 8.7, cluster is named as c+number.

Table 8.7: Domain Function Grouping (Clustergram 4)

CG	Connected E3 domain families	cluster	Interacting domain families functional affiliation	# CIDF	Relevant Domains (Pfam Name)
4	cullin + apc_cyclosome	5	Cell cycle control, UPS	4	APC_CDC26, APC10, APC2, APC8
4	cullin + apc_cyclosome	5	Unknown function	9	Pfam-B_11656, Pfam-B_19539, Pfam-B_2003, Pfam-B_20961, Pfam-B_22580, Pfam-B_2556, Pfam-B_65518, Pfam-B_70135, Pfam-B_7971
4	cullin + apc_cyclosome	5	UPS	3	Cullin, PC_rep, UQ_con
4	cullin + apc_cyclosome	6	PPI	1	TPR_1
4	cullin + apc_cyclosome	6	PPI, other	1	WD40
4	cullin + apc_cyclosome	6	UPS	1	ubiquitin
4	ring + ddb1_wd + ddb1	1	Cell cycle control	1	zf-CCCH
4	ring + ddb1_wd + ddb1	1	DNA binding	1	CSD
4	ring + ddb1_wd + ddb1	1	DNA binding, RNA polyadenylation	1	CPSF_A
4	ring + ddb1_wd + ddb1	1	DNA metabolism	2	Mago_nashi, R3H
4	ring + ddb1_wd + ddb1	1	DNA packing	1	PHF5
4	ring + ddb1_wd + ddb1	1	DNA repairment	1	NUDIX
4	ring + ddb1_wd + ddb1	1	Endoplasmic reticulum transport	1	Sec63
4	ring + ddb1_wd + ddb1	1	Mitosis	2	SMC_hinge, SMC_N
4	ring + ddb1_wd + ddb1	1	NNAA binding	2	SAP, Zf-CCHC
4	ring + ddb1_wd + ddb1	1	Nuclear trafficking	1	zf-RanBP
4	ring + ddb1_wd + ddb1	1	Nucleic acids packing	1	Helicase_C
4	ring + ddb1_wd + ddb1	1	Nucleolus	2	ROKNT, Zf-RNPHF
4	ring + ddb1_wd + ddb1	1	Other, antioxidant	1	Lactamase_B
4	ring + ddb1_wd + ddb1	1	Protein binding, NNAA binding	1	MIF4G
4	ring + ddb1_wd + ddb1	1	Ribosome	1	Ribosomal_L7Ae
4	ring + ddb1_wd + ddb1	1	RNA binding	3	dsrm, KH_1, RRM_1
4	ring + ddb1_wd + ddb1	1	RNA metabolism	6	LSM, NTP_transf_2, PAP_central, PAP_RNA-bind, RMMBL, Smg4_UPF3
4	ring + ddb1_wd + ddb1	1	snRNA related	2	MIF4G_like, MIF4G_like_2
4	ring + ddb1_wd + ddb1	1	Splicing	9	PRO8NT, PROCN, PROCT, PRP1_N, PSP, RRM_4, SF3b1, U5_2-snRNA_bdg, U6-snRNA_bdg
4	ring + ddb1_wd + ddb1	1	Transcription	12	DNA_RNApol_7kD, PRP4, PWI, RBM1CTR, RNA_pol_A_bac, RNA_pol_L, RNA_POL_M_15KD, RNA_pol_N, RNA_pol_Rpb4, RNA_pol_Rpb6, RNA_pol_Rpb8, TFIIF_beta
4	ring + ddb1_wd + ddb1	1	Transcription control	1	CBFNT
4	ring + ddb1_wd + ddb1	1	Unknown function	58	DUF1605, DUF618, Pfam-B_101290, Pfam-B_1093, Pfam-B_111803, Pfam-B_11621, Pfam-B_12405, Pfam-B_1258, Pfam-B_13959, Pfam-B_140379, Pfam-B_14911, Pfam-B_15526, Pfam-B_15994, Pfam-B_165324, Pfam-B_16945, Pfam-B_170767, Pfam-B_170869, Pfam-B_172256, Pfam-B_18142, Pfam-B_19120, Pfam-B_19750, Pfam-B_20301, Pfam-B_20467, Pfam-B_22944, Pfam-B_23505, Pfam-B_23725, Pfam-B_2388, Pfam-B_25495, Pfam-B_2642, Pfam-B_29039, Pfam-B_30151, Pfam-B_31333, Pfam-B_32240, Pfam-B_32301, Pfam-B_37563, Pfam-B_38155, Pfam-B_39039, Pfam-B_4383, Pfam-B_44509, Pfam-B_44653, Pfam-B_48116, Pfam-B_49499, Pfam-B_51557, Pfam-B_5847, Pfam-B_6173, Pfam-B_6189, Pfam-B_6253, Pfam-B_64, Pfam-B_6570, Pfam-B_7405, Pfam-B_7695, Pfam-B_87015, Pfam-B_9107, Pfam-B_92116, Pfam-B_92140, Pfam-B_97594, Pfam-B_9884, SPRY
4	ring + ddb1_wd + ddb1	1	Unknown function, DNA binding?	1	CPSF_A
4	ring + ddb1_wd + ddb1	1	Unknown function, DNA metabolism?	1	DUF382
4	ring + ddb1_wd + ddb1	1	Unknown function, RNA binding?	2	G-patch, Surp
4	ring + ddb1_wd + ddb1	1	Unknown function, RNA metabolism?	1	HA2
4	ring + ddb1_wd + ddb1	1	Unknown function, transcription regulation?	1	zf-NF-X1
4	ring + ddb1_wd + ddb1	1	UPS, RING domain	1	zf-C3HC4
4	ring + ddb1_wd + ddb1	1	UPS, transcription	1	Mov34

Note. CG: clustergram. CIDF: common interacting domain families. For cluster see clustergram figure 8.8, cluster is named as c+number.

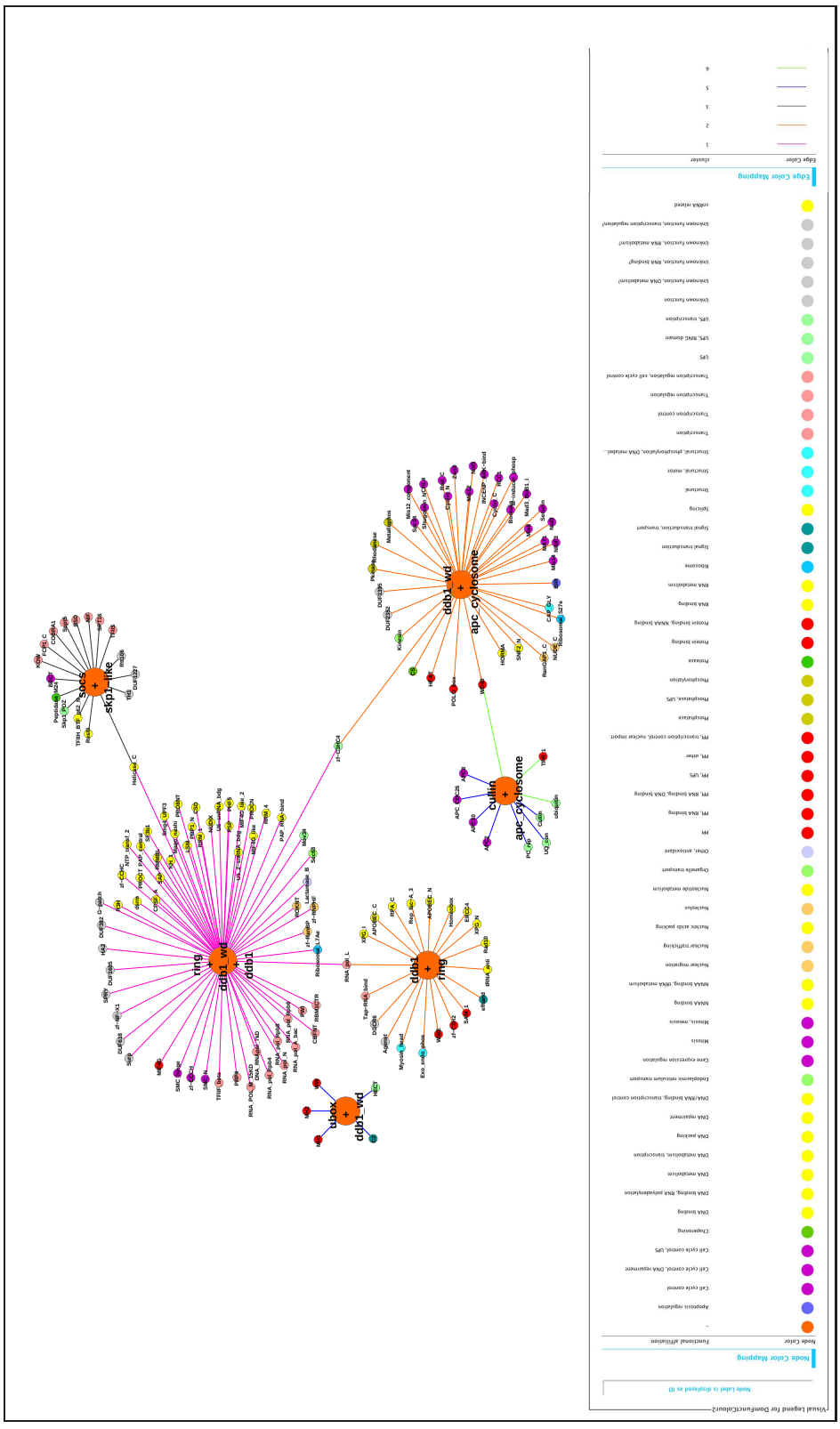


Fig. 8.9: The following figure shows the interaction network for specific domains associated to E3 families clusters. The domains were extracted from tables 8.5, 8.6 and 8.7. Figure created using Cytoscape [195].

8.7 Conclusions

In the work developed in this chapter, a new bioinformatic approach has been presented in order to characterize complex molecular associations such as the E3-machinery – protein substrate relationship. By the means of a network neighbourhood analysis between an E3 and a potential substrate, a connectivity profile of each E3 domain family and the domains of their interacting partners was drawn in an effort to identify potential ubiquitylation substrates. Thus, with the purpose to establish the connectivity profiles, a classification algorithm that uses E3-interacting partner interaction information to draw connectivity groups between different E3 families was implemented. In the second part, the information of domains from interacting proteins was used for the sake of finding coincidental interacting patterns between E3 families.

It was proved that it is not suitable to predict potential structural function of an E3 domain family just using a basic connectivity profiles; given that most of the families just show low connectivity values and that those that do show higher connectivity do so because they are formed by proteins that perform their ubiquitylating activity as a multiprotein complex. It seems that a simple connectivity profiling is good enough to identify some E3s that work in multi-protein machineries, but then insufficient to characterize the rest of the ubiquitin ligases. Nevertheless, an enrichment process and clustergram analysis was applied with the intention of making the most of the information in our hands and characterizing the interacting partners (or potential substrates) of each family, in an effort to find commonalities between those preferred by different families. Thus, interacting partners found in the clusters tend to have similar cellular functions and this tendency is clear in those families grouped under the highest connectivity groups. In fact, a large proportion of common interacting domain families were found connecting two or more potential pieces of multi-protein E3 ligases, thus enforcing the speculation that specific E3 families could tend to target proteins with common and specific functions.

Lack of available biological knowledge about many of the proteins classified as E3s in the presented dataset (and of their substrates) has been an important drawback in this analysis, depriving us of a sufficiently large "positive" set in which to test E3-substrate predictions. But even with all its limitations, it could be successfully correlated E3 domain families known to work in a collaborative fashion and to detect a common pattern in the functionality of the proteins they might be targeting. In addition, a reliable “positive set” could lead to build a robust predictor and it is one goal that is considered as future work. In this way, a supervised learning classifier as a support vector machine (SVM) model has already been used successfully as PPI predictor [238, 16, 48, 129, 237]. Although this classifier requires of positive and negative examples, a negative set is easier to obtain under some usual assumptions, for example by pairing up randomly selected proteins [238, 173, 232] or pairing proteins which are not sharing the same cellular compartment [106].

Thus, for further work, the application of this approach to better known post-translational modifications, such as phosphorylation, could avoid the found problems and show its real potential, bringing out its full predictive capabilities and helping to identify kinase targets, for example.

Because of the size of the set of interactions considered (126000 aprox.) and the possibility of integration of additional information of diverse kind (e.g., sequence, structure or functional annotations), the presented approach also faces some limitations due to memory and computational requirements. To solve this issue, parallelism-based computational strategies could be applied, i.e, using algorithms that divide the main problem into sub-problems. The classical approaches for this problem are 1) dividing used data (memory division), 2) tasks (function division) or 3) both; and then, these sub-problems are scattered into computational nodes in a cluster or grid. This way, calculation time and memory requirements would be reduced and would yield a much more powerful tool.

9 CONCLUSIONS, CONTRIBUTIONS AND FURTHER WORK

This chapter gathers the conclusions from this PhD thesis, along with the main contributions obtained from a depth researching process during the thesis period. Furthermore, in the last part of this chapter, further work is included as factional extension of the lines presented in this thesis.

9.1 Conclusions and contributions

There are several main conclusions that can be extracted the work developed in this thesis. From the conclusions extracted on a depth analysis of literature, it is known that protein-protein interactions (PPIs) are of great importance in almost any biological function carried out within the cell. Actually, a great part of the current biological researching is focused on protein networks in order to understand the main cell mechanisms [80]. Although the experimental techniques have improved in recent years, the computational approaches for predicting PPIs are of fundamental importance because of cost and time requirements associated with the experimental techniques [80]. Therefore, the main aim of the present research work is to provide a reliable, accurate and general methodology for predicting protein-protein interactions.

Different computational methods can be found in literature for PPI prediction from bayesian approaches [106, 163] to algebraic topological approaches [96] applied to diverse model organism from yeast (*Saccharomyces cerevisiae*) to humans (*Homo Sapiens*). In this thesis, the different approaches to implement the proposed methodology for predicting protein-protein interactions utilised support vector machine (SVM) models which has demonstrated its reliability and generalisation capability, good performance returning high sensibility and high specificity for classifications of any datasets along this thesis. The chosen model organism was yeast, although, in

the last of the thesis, a specific approach to characterize E3 enzymes was oriented to human.

Supervised learning methods such as support vector machine (SVM) models requires complete, meaningful and reliable positive and negative datasets (Gold Standard Positive and Negative sets, called GSP and GSN). Positive datasets are reported by authors indicating also several level of reliability. Nevertheless, negative datasets are difficult to find because data are not reported by experimentalists [185]. It is a common practise to use datasets with randomly paired proteins [238, 173, 232] or by selecting pairs of proteins that are not sharing the same sub-cellular compartment [106]. Hence, in this thesis, the different approaches used high-reliability GSP and GSN in order to obtain a reliable and accurate PPI predictors showing high sensibility and high specificity levels in the results. GSP was extracted from Saeed el at. [185]. Several GSNs were used in the different presented approached: randomly selected from the negative set proposed by Saeed et al. [185], "balanced" negative set obtained through applying approach proposed by Yu et al. [233] and a representative negative set obtained using the hierarchical clustering method presented in this thesis. High accuracy rates were obtained using these Gold Standard Set for training the SVM models, though the proposed hierarchical clustering method aims better performance considered as other important reached aim of this research work.

However, the good performance of the approaches proposed for the presented methodology is also attained by the sake of two important goals of this thesis: feature extraction and feature selection. A fundamental feature extraction through well-known datasets in proteomics was implemented, including two important similarity measures that showed their influence in order to improve the performance of trained SVM models. Three new feature selection methods to select the relevant features were presented in this thesis showed also its effectiveness: an ensemble approach using the margin based feature selection criterion methods (Simba, Relief and G-flip proposed by R. Gilad-Bachrach el at. [175]), a filter/wrapper feature selection method using Relief algorithm as filter and SVM models for wrapper approach, and a parallel filter-wrapper feature selection implemented in a master-slave parallel approach using the minimal-redundancy-maximal-relevance criterion of the mutual information as filter method and SVM models for wrapper approach.

Although the main aim has been a general methodology for predicting protein-protein interactions applied to Yeast, a secondary important goal has been reached in this thesis. In the last part, with the purpose to face an specific problem, an characterization of the protein interaction neighbourhood of E3 enzymes in the Ubiquitin-Proteasome System (UPS) has been carried out. The ubiquitin-proteasome system is the primary intracellular machinery responsible for elimination of unfolded proteins and for the selective destruction of regulatory proteins involved in a wide range of cellular processes [98]. Then, this approach is motivated because alterations of the UPS have been implicated in the pathogenesis of many diseases, such as inflammatory

and immunological disorders [190] or, more prominently, neurodegenerative diseases such as Parkinson's [126, 113], Huntington's [17] or Alzheimer's diseases [205].

Thus so, in this thesis is also outlined a bioinformatics approach that aims to characterize complex molecular associations such as the E3-machinery – protein substrate relationship. The implementation for this approach is realized through two main process: an E3-family connectivity profile analysis and an clustergram analysis using an previous biologically enrichment process. Analysing the network neighbourhood between an E3 and a potential substrate, a connectivity profile has been drawn for each E3 domain family and the domains of their interacting partners, and thus this possible procedure is proposed with the purpose of sorting out potential ubiquitylation substrates and help to define the structural constituents of the E3-substrate complexes. After, this approach is extended by the means of an biologically enrichment process and clustergram analysis. Thanks to this last step, interacting partners with similar cellular functions were found in clusters for those domain families associated to multi-protein E3 complexes and found grouped under the highest connectivity groups, providing a potential framework for further research.

The main contributions developed during the of this PhD thesis can be summarized schematically below:

1. A reliable, accurate, general methodology for predicting protein-protein interactions is provided. Several approaches of this methodology are implemented in chapters 5, 6 and 7, sharing the proposed common procedure: feature extraction from well-known databases in proteomics, feature selection, creation of the SVM model and validation of the model comparing to other approach or testing external datasets obtained.

2. A feature extraction process from well-known databases is proposed, extracting also a number of features using the two similarity measures presented in this research work which have a fundamental effect in the good performance of the trained SVM models (see 5.2).

3. Three feature extraction were implemented in order to filtering out the relevant or redundant features obtaining a sub-optimal set of characteristics for training more accuracy, less complex and reliable SVM models. The feature selection approaches were:

   (a) an ensemble approach using the margin based feature selection criterion methods (Simba, Relief and G-flip proposed by R. Gilad-Bachrach el at. [175]), see chapter 5.

   (b) a filter/wrapper feature selection method using Relief algorithm as filter and SVM models for wrapper approach, see chapter 6.

(c) a parallel filter-wrapper feature selection implemented in a master-slave parallel approach using the minimal-redundancy-maximal-relevance criterion based on mutual information as filter method and SVM models for wrapper approach. MPI-MEX [82], a parallel interface for MPI in MATLAB implemented in the Department of Computer Architecture and Computer Technology in the University of Granada (Spain) by Dr. Alberto Guillén, was used in the developing of this approach. This approach was motivated for the computational requirements such as memory and calculation. See chapter 7.

4. The predictor models have a good performance and have been validated using different nature datasets, i.e, experimental, computational and literature-collected datasets showing high sensibility and sensitivity values in the classification of PPI.

5. An confidence score for each predicted pair of proteins is also presented. This score may help to validate PPI and it was used to generate ROC curves in the respective analysis of each chapter (5, 6 and 7). Through a slight modification of support vector machine model implementation, they were able to return this confidence score (see 5.4). This score is motived because a more universal confidence scoring approach is more preferable because it is free of biases due to the influence of particular experimental setup context [30].

6. Three reliable negative datasets were used in order to build more accuracy SVM models. In every chapter 5, 6 and 7 were incorporated as part of training sets respectively. Three types of negative datasets were:

   (a) randomly selected from the negative set created by Saeed et al. [185].

   (b) "balanced" negative set obtained of the approach by Yu et al. [233].

   (c) a representative negative set obtained using the hierarchical clustering method presented in this thesis. High accuracy rates were obtained including this negative set as part of training set in order to build the SVM models. This hierarchical clustering method attains better performance than the rest of utilised negative sets.

7. An bioinformatics approach that aims to characterize complex molecular associations such as the E3-machinery – protein substrate relationship has been presented in this thesis. Studying the network neighbourhood between an E3 and a potential substrate, a connectivity profile has been drawn for each E3 domain family and the domains of their interacting partners, in an effort to sort out potential ubiquitylation substrates and help to define the structural constituents of the E3-substrate complexes. An E3-family connectivity profile analysis was applied in order to find interacting partners between families. Extending this approach through an

biologically enrichment process and clustergram analysis, interacting partners with similar cellular functions were found in clusters for those domain families associated to multi-protein E3 complexes and found grouped under the highest connectivity groups, providing a potential framework for further research.

9.2 Further work

For further work, the general methodology for predicting protein-protein interactions could be applied to other model organisms included human organism. However, new features should be proposed and implemented in other to improve the accuracy and reliability of the predictor models. It is considering to use other variable selection methods as well as using other machine learning methodologies such as fuzzy systems that provide interpretable solutions [89] although SVM model shows good performance.

The proposed hierarchical parallel clustering could improve the performance of classifier with the purpose of obtained a balanced negative dataset using a more complex clustering algorithm. A parallel approach was applied making a better load balancing would be suitable to reduce time computation in the filter/wrapper feature selection approach.

The algorithm MSSC [96] explained in chapter 3 could be modified, instead of applying a greedy approach, a genetic algorithm using the confidence score as fitness presented in this thesis.

In the case of UPS analysis, for further work could be the application of the proposed approach to better known post-translational modifications, such as phosphorylation, could avoid the found problems and show its real potential, bringing out its full predictive capabilities and helping to identify kinase targets, for example. On the other hand, because of the size of the set of interactions considered (126000 aprox.) and the possibility of integration of additional information of diverse kind (e.g., sequence, structure or functional annotations), the presented approach also faces some limitations due to memory and computational requirements. To solve this issue, parallelism-based computational strategies could be applied: 1) memory division), 2) function division or 3) both. It is also considering applied the general PPI prediction methodology proposed in this paper to this problem despite the lack of available biological knowledge.

9.3 Publications

This sections gathers all publications produced directly or collaborative contributions during the course of this PhD thesis. The publications are sorted chronologically within each group:

Journals

1. **J.Urquiza**, H. Pomares, L.J.Herrera, I.Rojas, J.Ortega, A.Prieto. *Method for Prediction of Protein-Protein Interactions in Yeast using Genomics/Proteomics*

Information and Feature Selection. Neurocomputing 74 (2011).pp. 2683-2690. ISSN 0925-2312. DOI: 10.1016/j.neucom.2011.03.025. (Impact Factor: 1.429).

2. L.J. Herrera, H. Pomares, I. Rojas, A. Guillen, G. Rubio, **J. Urquiza**. *Global and Local Modelling in RBF Networks*. Neurocomputing. 2011. ISSN 0925-2312. DOI: 10.1016/j.neucom.2011.03.027. In Press. (Impact Factor: 1.429).

3. J. P. Florido, H. Pomares, I. Rojas, **J. M. Urquiza** and M. A. Lopez-Gordo *A deterministic model selection scheme for incremental RBFNN construction in time series forecasting*. Neural Computing & Applications. Springer London. pages 1-16, 2010. ISSN: 0941-0643. DOI: 10.1007/s00521-010-0466-5 (Impact 0.563)

4. M. A. Lopez, H. Pomares, F. Pelayo, **J.Urquiza**, J. Perez, *Evidences of cognitive effects over auditory steady-state responses by means of artificial neural networks and its use in brain-computer interfaces*, Neurocomputing, Volume 72, Issues 16-18, Financial Engineering; Computational and Ambient Intelligence, October 2009, Pages 3617-3623, ISSN 0925-2312. DOI: 10.1016/j.neucom.2009.04.021 (Impact Factor: 1.429).

International conferences

1. **Urquiza J.**, Rojas I., Pomares H., Herrera LJ, Pérez-Florido J. *Using Machine Learning Techniques And Techniques And Genomic/Proteomic Information For Protein-Protein Classification*. Histology and Histopathology (Cellular and Molecular Biology). Termis EU 2011 Annual Meeting (V). Tissue Engineering and Regenerative Medicine International Society. 7-10 June 2011. Granada (Spain). Volume 26 (supplement 1) 2011. ISSN 0213-3911. PAGE 300 (33.02) **Impact Factor 2.404 (2009)** IN PRESS.

2. Francisco M. Ortuño Guzman, I. Rojas, H. Pomares, **J. M. Urquiza**, J. P. Florido. *Emerging Methodologies in Multiple Sequence Alignment Using High Throughput Data*. Proceedings of 5th International Conference on Practical Applications of Computational Biology & Bioinformatics PACBB 2011. April. Series: Advances in Intelligent and Soft Computing, Vol. 93. ISBN: 978-3-642-19913-4

3. **J. M. Urquiza**, I. Rojas, H. Pomares, L. J. Herrera, J. P. Florido, F. Ortuño. *Using Machine Learning Techniques and Genomic/Proteomic Information from Known Databases for PPI Prediction*. Proceedings of 5th International Conference on Practical Applications of Computational Biology & Bioinformatics PACBB 2011. April. Series: Advances in Intelligent and Soft Computing, Vol. 93. ISBN: 978-3-642-19913-4

4. Javier P.Florido, Hector Pomares, Ignacio Rojas, **Jose Miguel Urquiza**, Luis Javier Herrera and Manuel Gonzalo. *Effect of Preprocessing Methods on Microarray*

based SVM classifiers in Affymetrix Genechips Proceedings of 2010 International Joint Conference on Neural Networks. Barcelona (Spain) 19-23 July 2010 [Paper #392]

5. Antonio Mora, Luis Javier Herrera, **Jose Urquiza**, Rojas Ignacio and J.J. Merelo. *Applying Support Vector Machines and Mutual Information to Book Losses Prediction.* Proceedings of 2010 International Joint Conference on Neural Networks. Barcelona (Spain) 19-23 July 2010 [Paper #837]

6. J. P. Florido, H. Pomares, L. J. Herrera, I. Rojas, **J. M. Urquiza**, *Time Series Prediction using Mutual Information and RBFNNs.* 1st International Workshop on mining of non-conventional data (MINCODA09) within the 13th Conference of the Spanish Association for Artificial Intelligence (CAEPIA TTIA2009), ISBN: 978-84-608-0978-4, pp.25-32, Sevilla (Spain), 4-9 Nov. 2009.

7. **J.M. Urquiza**, I. Rojas, H. Pomares, L.J. Herrera, G. Rubio, J.P. Florido, *Prediction of protein-protein interactions in yeast using SVMs with genomics/proteomics information and feature selection.* Proceedings of 24th International Symposium on Computer and Information Sciences, 2009. ISCIS 2009, IEEE CONFERENCE. pp.: 645 - 650. ISBN: 978-1-4244-5021-3, Northern Cyprus, 14-16 Sept. 2009.

8. **J.M. Urquiza**, I. Rojas, H. Pomares, J.P.Florido, G. Rubio, L. J. Herrera, J.C.Calvo, J. Ortega. *Method for Prediction of Protein-Protein Interactions in Yeast using Genomics/Proteomics Information and Feature Selection.* Proceedings of IWANN (International Work-Conference on Artificial Neural Networks) 2009, LNCS (Lecture Notes on Computer Science). Bio-Inspired Systems: Computational and Ambient Intelligence. pp.: 853-860 ISBN: 978-1-4244-5023-7, Salamanca (Spain), 10-12 June 2009. Selected for special issue.

9. J. C. Calvo, J. Ortega, M. Anguita, **J. M. Urquiza**, J. P. Florido. *Protein Structure Prediction by Evolutionary Multi-objective Optimization: Search Space Reduction by Using Rotamers* Proceedings of IWANN (International Work-Conference on Artificial Neural Networks) 2009, LNCS. Bio-Inspired Systems: Computational and Ambient Intelligence. pp.: 861-868 ISBN: 978-3-642-02477-1. Salamanca (Spain), 10-12 June 2009.

10. J. P. Florido, H. Pomares, I. Rojas, J. C. Calvo, **J. M. Urquiza**, M. Gonzalo Claros. *On Selecting the Best Pre-processing Method for Affymetrix Genechips* ". Proceedings of IWANN (International Work-Conference on Artificial Neural Networks) 2009, LNCS. Bio-Inspired Systems: Computational and Ambient Intelligence. pp.: 845-852, ISBN: 978-3-642-02477-1. Salamanca (Spain), 10 - 12 June 2009.

11. L. J. Herrera, H. Pomares, I. Rojas, A. Guillén, G. Rubio and **J. Urquiza**. *Global and Local Modelling in Radial Basis Functions Networks*, Proceedings of IWANN (International Work-Conference on Artificial Neural Networks) 2009 LNCS. Bio-Inspired Systems: Computational and Ambient Intelligence. pp.:49-56, ISBN 978-3-642-02477-1. Salamanca (Spain), 10 - 12 June 2009.

National conferences

1. J.P.Florido, H.Pomares, L.J.Herrera, I.Rojas, **J.M. Urquiza**.*Addressing RBFNNs for Time Series Prediction: inputs and network model selection*. University of Granada (Spain). V SIMPOSIO DE TEORÍA Y APLICACIONES DE MINERÍA DE DATOS, TAMIDA 2010. CEDI2010 Valencia 7-10 Sept. 2010 [paper #356]

2. Javier Perez, Hector Pomares, Ignacio Rojas, **Jose Miguel Urquiza.***Applications of computational intelligence to the integration of heterogeneous biological data sources.* Sesión 9-3 : Modelos, métodos y algoritmos en inteligencia computacional. CEDI2010 Valencia 7-10 Sept. 2010 [paper #333]

3. Héctor Pomares, Ignacio Rojas, Alberto Guillen, Luis Javier Herrera, Jesús Gonzalez, Miguel Damas, Javier Perez Florido, **José Urquiza**, Olga Valenzuela, Cornelio García. *Use of Evolutionary Algorithms for Two-level Minimization of Boolean Functions* (Paper #57) Sesión 9-3: Modelos, métodos y algoritmos en inteligencia computacional. Posters CEDI 2010. Valencia 7-10 Sept. 2010

4. **J. M. Urquiza**, H. Pomares, J. P. Florido, M. A. López-Gordo. *Prediction of water consumption through a GA-based variable selection method*. II SIMPOSIO DE INTELIGENCIA COMPUTACIONAL (SICO 2007), pp.: 389-396, ISBN: 978-84-9732-606-3. ZARAGOZA (Spain), 11 sept. 2007.

5. **J. M. Urquiza**, José M. Ríos, Mancia Anguita, Eduardo Ros. *Real-Time Motion Processing Algorithm in a PC*. II SIMPOSIO DE INTELIGENCIA COMPUTACIONAL (SICO 2007), pp.: 65-72, ISBN: 978-84-9732-606-3. ZARAGOZA (Spain), 11 sept. 2007.

6. J. P. Florido, H. Pomares, **J. M. Urquiza**, M.A. López-Gordo, M. Damas. *Data Distribution for Cross Validation using a Genetic Algorithm for Time Series Prediction*. II SIMPOSIO DE INTELIGENCIA COMPUTACIONAL (SICO 2007), pp.: 373-380, ISBN: 978-84-9732-606-3. ZARAGOZA (Spain), 11 sept. 2007.

Chapter of Book

1. **J.M. Urquiza**, I. Rojas, H. Pomares, L. J. Herrera. *Method for Prediction of Protein-Protein Interactions in Yeast using Genomics/Proteomics Information and Feature Selection.* : Proteomics: Methods, Applications and Limitations (Chapter

5). Editors: Giselle C. Rancourt. Pub. Date: 2010 - 3rd Quarter. ISBN: 978-1-61668-800-4

BIBLIOGRAFÍA

[1] Wassim M. Abida, Anatoly Nikolaev, Wenhui Zhao, Wenzhu Zhang, and Wei Gu. FBXO11 promotes the neddylation of p53 and inhibits its transcriptional activity. *Journal of Biological Chemistry*, 282(3):1797 –1804, January 2007.

[2] Michael Albers, Harald Kranz, Ingo Kober, Carmen Kaiser, Martin Klink, Jörg Suckow, Rainer Kern, and Manfred Koegl. Automated yeast two-hybrid screening for nuclear receptor-interacting proteins. *Molecular & Cellular Proteomics*, 4(2):205 –213, February 2005.

[3] Istvan Albert and Reka Albert. Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics*, 20(18):3346–3352, December 2004.

[4] R Apweiler, T K Attwood, A Bairoch, A Bateman, E Birney, M Biswas, P Bucher, L Cerutti, F Corpet, M D Croning, R Durbin, L Falquet, W Fleischmann, J Gouzy, H Hermjakob, N Hulo, I Jonassen, D Kahn, A Kanapin, Y Karavidopoulou, R Lopez, B Marx, N J Mulder, T M Oinn, M Pagni, F Servant, C J Sigrist, and E M Zdobnov. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research*, 29(1):37–40, January 2001. PMID: 11125043.

[5] B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, A. T. Ghanbarian, S. Kerrien, J. Khadake, J. Kerssemakers, C. Leroy, M. Menden, M. Michaut, L. Montecchi-Palazzi, S. N. Neuhauser, S. Orchard, V. Perreau, B. Roechert, K. van Eijk, and H. Hermjakob. The IntAct molecular interaction database in 2010. *Nucleic Acids Research*, 38(Database):D525–D531, October 2009.

[6] L. Aravind and Eugene V. Koonin. The u box is a modified RING finger – a

common domain in ubiquitination. *Current Biology*, 10(4):R132–R134, February 2000.

[7] Sunil Arya, Theocharis Malamatos, and David M Mount. Space-efficient approximate voronoi diagrams. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, STOC '02, page 721–730, Montreal, Quebec, Canada, 2002. ACM. ACM ID: 510011.

[8] Francisco Azuaje, Haiying Wang, Huiru Zheng, Olivier Bodenreider, and Alban Chesneau. Predictive integration of gene Ontology-Driven similarity and functional interactions. In *Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on*, pages 114–119, 2006.

[9] Gary D. Bader, Ian Donaldson, Cheryl Wolting, B. F. Francis Ouellette, Tony Pawson, and Christopher W. V. Hogue. BIND—The biomolecular interaction network database. *Nucleic Acids Research*, 29(1):242–245, January 2001. PMID: 11125103 PMCID: 29820.

[10] Joel S Bader, Amitabha Chaudhuri, Jonathan M Rothberg, and John Chant. Gaining confidence in high-throughput protein interaction networks. *Nat Biotech*, 22(1):78–85, January 2004.

[11] Pierre Baldi and Søren Brunak. *Bioinformatics: The Machine Learning Approach, Second Edition*. The MIT Press, second edition edition, August 2001.

[12] Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17(Suppl 1):S22–S29, June 2001.

[13] David Barford. Structure, function and mechanism of the anaphase promoting complex (APC/C). *Quarterly Reviews of Biophysics*, 44(02):153–190, 2011.

[14] Russell Bell, Alan Hubbard, Rakesh Chettier, Di Chen, John P. Miller, Pankaj Kapahi, Mark Tarnopolsky, Sudhir Sahasrabuhde, Simon Melov, and Robert E. Hughes. A human protein interaction network shows conservation of aging processes between human and invertebrate species. *PLoS Genet*, 5(3):e1000414, March 2009.

[15] Richard Ernest Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, first edition ~ 1st printing edition, 1961.

[16] Asa Ben-Hur and William Stafford Noble. Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21(suppl_1):i38–46, June 2005.

[17] Eric J. Bennett, Thomas A. Shaler, Ben Woodman, Kwon-Yul Ryu, Tatiana S. Zaitseva, Christopher H. Becker, Gillian P. Bates, Howard Schulman, and Ron R. Kopito. Global changes to the ubiquitin system in huntington's disease. *Nature*, 448(7154):704–708, 2007.

[18] Tord Berggard, Sara Linse, and Peter James. Methods for the detection and analysis of protein-protein interactions. *PROTEOMICS*, 7(16):2833–2842, 2007.

[19] Ann-Charlotte Berglund, Erik Sjölund, Gabriel Östlund, and Erik L. L. Sonnhammer. InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Research*, 36(Database issue):D263–D266, January 2008. PMID: 18055500 PMCID: 2238924.

[20] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucl. Acids Res.*, 28(1):235–242, January 2000.

[21] Thijs Beuming, Lucy Skrabanek, Masha Y. Niv, Piali Mukherjee, and Harel Weinstein. PDZBase: a protein–protein interaction database for PDZ-domains. *Bioinformatics*, 21(6):827 –828, March 2005.

[22] J. Bins and B. A Draper. Feature selection from huge feature sets. In *Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001. Proceedings*, volume 2, pages 159–165 vol.2. IEEE, 2001.

[23] Peter Block, Juri Paern, Eyke Huellermeier, Paul Sanschagrin, Christoph A. Sotriffer, and Gerhard Klebe. Physicochemical descriptors to discriminate protein-protein interactions in permanent and transient complexes selected by means of machine learning algorithms. *Proteins: Structure, Function, and Bioinformatics*, 65(3):607–622, 2006.

[24] Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J. Martin, Karine Michoud, Claire O'Donovan, Isabelle Phan, Sandrine Pilbout, and Michel Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.*, 31(1):365–370, January 2003.

[25] Y.C. Bor, J. Swartz, Y. Li, J. Coyle, D. Rekosh, and Marie-Louise Hammarskjold. Northern blot analysis of mRNA from mammalian polyribosomes. *Protocol Exchange*, September 2006.

[26] Peer Bork, Lars J Jensen, Christian von Mering, Arun K Ramani, Insuk Lee, and Edward M Marcotte. Protein interaction networks from yeast to human. *Current Opinion in Structural Biology*, 14(3):292–299, June 2004.

[27] Tewis Bouwmeester, Angela Bauch, Heinz Ruffner, Pierre-Olivier Angrand, Giovanna Bergamini, Karen Croughton, Cristina Cruciat, Dirk Eberhard, Julien Gagneur, Sonja Ghidelli, Carsten Hopf, Bettina Huhse, Raffaella Mangano, Anne-Marie Michon, Markus Schirle, Judith Schlegl, Markus Schwab, Martin A. Stein, Andreas Bauer, Georg Casari, Gerard Drewes, Anne-Claude Gavin, David B. Jackson, Gerard Joberty, Gitte Neubauer, Jens Rick, Bernhard Kuster, and Giulio Superti-Furga. A physical and functional map of the human TNF-[alpha]/NF-[kappa]B signal transduction pathway. *Nat Cell Biol*, 6(2):97–105, February 2004.

[28] Brian Bowen, Jay Steinberg, U.K. Laemmli, and Harold Weintraub. The detection of DNA-binding proteins by protein blotting. *Nucleic Acids Research*, 8(1):1 –20, January 1980.

[29] Alan Boyden. Genetics and homology. *The Quarterly Review of Biology*, 10(4):448–451, December 1935. ArticleType: research-article / Full publication date: Dec., 1935 / Copyright © 1935 The University of Chicago Press.

[30] Pascal Braun, Murat Tasan, Matija Dreze, Miriam Barrios-Rodiles, Irma Lemmens, Haiyuan Yu, Julie M Sahalie, Ryan R Murray, Luba Roncari, Anne-Sophie de Smet, Kavitha Venkatesan, Jean-Francois Rual, Jean Vandenhaute, Michael E Cusick, Tony Pawson, David E Hill, Jan Tavernier, Jeffrey L Wrana, Frederick P Roth, and Marc Vidal. An experimentally derived confidence score for binary protein-protein interactions. *Nat Meth*, 6(1):91–97, January 2009.

[31] Dale E. Bredesen, Rammohan V. Rao, and Patrick Mehlen. Cell death in the nervous system. *Nature*, 443(7113):796–802, October 2006.

[32] Ingo Brigandt. A theory of conceptual advance: Explaining conceptual change in evolutionary, molecular, and evolutionary developmental biology. http://etd.library.pitt.edu/ETD/available/etd-08032006-145211/, September 2006. The theory of concepts advanced in the dissertation aims at accounting for a) how a concept makes successful practice possible, and b) how a scientific concept can be subject to rational change in the course of history. Traditional accounts in the philosophy of science have usually studied concepts in terms only of their reference; their concern is to establish a stability of reference in order to address the incommensurability problem. My discussion, in contrast, suggests that each scientific concept consists of three components of content: 1) reference, 2) inferential role, and 3) the epistemic goal pursued with the concept's use. I argue that in the course of history a concept can change in any of these three components, and that change in one component—including change of reference—can be accounted for as being rational relative to other

components, in particular a concept's epistemic goal. This semantic framework is applied to two cases from the history of biology: the homology concept as used in 19th and 20th century biology, and the gene concept as used in different parts of the 20th century. The homology case study argues that the advent of Darwinian evolutionary theory, despite introducing a new definition of homology, did not bring about a new homology concept (distinct from the pre-Darwinian concept) in the 19th century. Nowadays, however, distinct homology concepts are used in systematics/evolutionary biology, in evolutionary developmental biology, and in molecular biology. The emergence of these different homology concepts is explained as occurring in a rational fashion. The gene case study argues that conceptual progress occurred with the transition from the classical to the molecular gene concept, despite a change in reference. In the last two decades, change occurred internal to the molecular gene concept, so that nowadays this concept's usage and reference varies from context to context. I argue that this situation emerged rationally and that the current variation in usage and reference is conducive to biological practice. The dissertation uses ideas and methodological tools from the philosophy of mind and language, the philosophy of science, the history of science, and the psychology of concepts.

[33] Kevin R Brown and Igor Jurisica. Online predicted human interaction database. *Bioinformatics (Oxford, England)*, 21(9):2076–2082, May 2005. PMID: 15657099.

[34] F. Browne, Haiying Wang, Huiru Zheng, and F. Azuaje. Supervised statistical and machine learning approaches to inferring pairwise and Module-Based protein interaction networks. In *Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on*, pages 1365–1369, 2007.

[35] Fiona Browne, Haiying Wang, Huiru Zheng, and Francisco Azuaje. A knowledge-driven probabilistic framework for the prediction of protein-protein interaction networks. *Computers in Biology and Medicine*, 40(3):306–317, March 2010.

[36] Carol J. Bult, Janan T. Eppig, James A. Kadin, Joel E. Richardson, and Judith A. Blake. The mouse genome database (MGD): mouse biology and model systems. *Nucleic Acids Research*, 36(Database issue):D724–D728, January 2008. PMID: 18158299 PMCID: 2238849.

[37] Evelyn Camon, Michele Magrane, Daniel Barrell, Vivian Lee, Emily Dimmer, John Maslen, David Binns, Nicola Harte, Rodrigo Lopez, and Rolf Apweiler. The gene ontology annotation (GOA) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Research*, 32(Database issue):D262–D266, January 2004. PMC308756.

[38] A. Ceol, A. Chatr Aryamontri, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli, and G. Cesareni. MINT, the molecular interaction database: 2009 update. *Nucleic Acids Research*, 38(Database):D532–D539, November 2009.

[39] J. Cerezo and V. Madrid. Técnicas, estrategias y usos de biología molecular en medicina. *Revista de Investigación Clínica*, 126:603–11, 1995.

[40] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a Library for Support Vector Machines*. Technical Report, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[41] Frédéric Colland, Xavier Jacq, Virginie Trouplin, Christelle Mougin, Caroline Groizeleau, Alexandre Hamburger, Alain Meil, Jérôme Wojcik, Pierre Legrain, and Jean-Michel Gauthier. Functional proteomics mapping of a human signaling pathway. *Genome Research*, 14(7):1324 –1332, July 2004.

[42] Gene Ontology Consortium. The gene ontology (GO) database and informatics resource. *Nucl. Acids Res.*, 32(suppl_1):D258–261, January 2004.

[43] The UniProt Consortium. The universal protein resource (UniProt). *Nucl. Acids Res.*, 35(suppl_1):D193–197, January 2007.

[44] Florence Corpet, Florence Servant, Jérôme Gouzy, and Daniel Kahn. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Research*, 28(1):267–269, January 2000. PMC102458.

[45] R. Correa-Rotter and G. Gamba. Biología molecular en medicina. viii. análisis de la expresión génica. *Revista de Investigación Clínica*, 49:163–6, 1997.

[46] Corinna Cortes and Vladimir Vapnik. Support vector network. *Mach. Learn.*, 1995.

[47] T. M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley and Sons, June 2006.

[48] Roger A. Craig and Li Liao. Improving Protein-Protein interaction prediction based on phylogenetic information using a Least-Squares support vector machine. *Annals of the New York Academy of Sciences*, 1115(Reverse Engineering Biological Networks: Opportunities and Challenges in Computational Methods for Pathway Inference):154–167, 2007.

[49] Koby Crammer, Ran Gilad-bachrach, Amir Navot, and Naftali Tishby. Margin analysis of the LVQ algorithm. *IN: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 2002*, pages 462—469, 2002.

[50] Universidad Internacional de Andalucía. Documentación del máster en bioinformática. Technical report, Universidad Internacional de Andalucía, 2008.

[51] Kurt J. De Vos, Andrew J. Grierson, Steven Ackerley, and Christopher C.J. Miller. Role of axonal transport in neurodegenerative diseases*. *Annual Review of Neuroscience*, 31(1):151–173, July 2008.

[52] Charlotte M. Deane, Lukasz Salwinski, Ioannis Xenarios, and David Eisenberg. Protein interactions: Two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, 1(5):349–356, May 2002.

[53] Edward F. DeLong and Norman R. Pace. Environmental diversity of bacteria and archaea. *Systematic Biology*, 50(4):470–478, August 2001. ArticleType: primary_article / Full publication date: Aug., 2001 / Copyright © 2001 Society of Systematic Biologists.

[54] Minghua Deng, Shipra Mehta, Fengzhu Sun, and Ting Chen. Inferring Domain–Domain interactions from Protein–Protein interactions. *Genome Research*, 12(10):1540–1548, October 2002. PMC187530.

[55] Joseph L. DeRisi, Vishwanath R. Iyer, and Patrick O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680 –686, October 1997.

[56] Vikrant Deshmukh, Chris Cannings, and Alun Thomas. Estimating the parameters of a model for protein–protein interaction graphs. *Mathematical Medicine and Biology*, 23(4):279 –295, December 2006.

[57] A Drawid and M Gerstein. A bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *Journal of Molecular Biology*, 301(4):1059–75, August 2000. PMID: 10966805.

[58] David M Duda, Daniel C Scott, Matthew F Calabrese, Erik S Zimmerman, Ning Zheng, and Brenda A Schulman. Structural regulation of cullin-RING ubiquitin ligase complexes. *Current Opinion in Structural Biology*, 21(2):257–264, April 2011.

[59] Selina S Dwight, Rama Balakrishnan, Karen R Christie, Maria C Costanzo, Kara Dolinski, Stacia R Engel, Becket Feierbach, Dianna G Fisk, Jodi Hirschman, Eurie L Hong, Laurie Issel-Tarver, Robert S Nash, Anand Sethuraman, Barry Starr, Chandra L Theesfeld, Rey Andrada, Gail Binkley, Qing Dong, Christopher

Lane, Mark Schroeder, Shuai Weng, David Botstein, and J Michael Cherry. Saccharomyces genome database: underlying principles and organisation. *Briefings in Bioinformatics*, 5(1):9–22, March 2004. PMID: 15153302.

[60] M B Eisen, P T Spellman, P O Brown, and D Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863–14868, December 1998. PMID: 9843981.

[61] Hanna Engelberg-Kulka, Shahar Amitai, Ilana Kolodkin-Gal, and Ronen Hazan. Bacterial programmed cell death and multicellular behavior in bacteria. *PLoS Genet*, 2(10):e135, October 2006.

[62] P. A Estevez, M. Tesmer, C. A Perez, and J. M Zurada. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 20(2):189–201, February 2009.

[63] Rob M Ewing, Peter Chu, Fred Elisma, Hongyan Li, Paul Taylor, Shane Climie, Linda McBroom-Cerajewski, Mark D Robinson, Liam O'Connor, Michael Li, Rod Taylor, Moyez Dharsee, Yuen Ho, Adrian Heilbut, Lynda Moore, Shudong Zhang, Olga Ornatsky, Yury V Bukhman, Martin Ethier, Yinglun Sheng, Julian Vasilescu, Mohamed Abu-Farha, Jean-Philippe Lambert, Henry S Duewel, Ian I Stewart, Bonnie Kuehl, Kelly Hogue, Karen Colwill, Katharine Gladwish, Brenda Muskat, Robert Kinach, Sally-Lin Adams, Michael F Moran, Gregg B Morin, Thodoros Topaloglou, and Daniel Figeys. Large-scale mapping of human protein–protein interactions by mass spectrometry. *Molecular Systems Biology*, 3:89–89, 2007. PMID: 17353931 PMCID: 1847948.

[64] Robert D. Finn, John Tate, Jaina Mistry, Penny C. Coggill, Stephen John Sammut, Hans-Rudolf Hotz, Goran Ceric, Kristoffer Forslund, Sean R. Eddy, Erik L. L. Sonnhammer, and Alex Bateman. The pfam protein families database. *Nucl. Acids Res.*, 36(suppl_1):D281–288, January 2008.

[65] E Fix and j Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties (Technical report 4). Technical report, USAF school of Aviation Medicine, 1951.

[66] James Fogarty, Ryan S Baker, and Scott E Hudson. Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. In *Proceedings of Graphics Interface 2005*, GI '05, page 129–136, Victoria, British Columbia, 2005. Canadian Human-Computer Communications Society. ACM ID: 1089530.

[67] Etienne Formstecher, Sandra Aresta, Vincent Collura, Alexandre Hamburger, Alain Meil, Alexandra Trehin, Céline Reverdy, Virginie Betin, Sophie Maire, Christine Brun, Bernard Jacq, Monique Arpin, Yohanns Bellaiche, Saverio Bellusci, Philippe Benaroch, Michel Bornens, Roland Chanet, Philippe Chavrier, Olivier Delattre, Valérie Doye, Richard Fehon, Gérard Faye, Thierry Galli, Jean-Antoine Girault, Bruno Goud, Jean de Gunzburg, Ludger Johannes, Marie-Pierre Junier, Vincent Mirouse, Ashim Mukherjee, Dora Papadopoulo, Franck Perez, Anne Plessis, Carine Rossé, Simon Saule, Dominique Stoppa-Lyonnet, Alain Vincent, Michael White, Pierre Legrain, Jérôme Wojcik, Jacques Camonis, and Laurent Daviet. Protein interaction mapping: A drosophila case study. *Genome Research*, 15(3):376–384, March 2005. PMC551564.

[68] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition, Second Edition*. Academic Press, 2 edition, October 1990.

[69] Francesco Galli, Mariangela Rossi, Yuri D'Alessandra, Marco De Simone, Teresa Lopardo, Ygal Haupt, Osnat Alsheich-Bartok, Shira Anzi, Eitan Shaulian, Viola Calabrò, Girolama La Mantia, and Luisa Guerrini. MDM2 and fbw7 cooperate to induce p63 protein degradation following DNA damage and cell differentiation. *Journal of Cell Science*, 123(14):2423 –2433, July 2010.

[70] Anne-Claude Gavin, Markus Bosche, Roland Krause, Paola Grandi, Martina Marzioch, Andreas Bauer, Jorg Schultz, Jens M. Rick, Anne-Marie Michon, Cristina-Maria Cruciat, Marita Remor, Christian Hofert, Malgorzata Schelder, Miro Brajenovic, Heinz Ruffner, Alejandro Merino, Karin Klein, Manuela Hudak, David Dickson, Tatjana Rudi, Volker Gnau, Angela Bauch, Sonja Bastuck, Bettina Huhse, Christina Leutwein, Marie-Anne Heurtier, Richard R. Copley, Angela Edelmann, Erich Querfurth, Vladimir Rybin, Gerard Drewes, Manfred Raida, Tewis Bouwmeester, Peer Bork, Bertrand Seraphin, Bernhard Kuster, Gitte Neubauer, and Giulio Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, January 2002.

[71] Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. Large margin principals for feature selection - a tutorial, December 2004.

[72] Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. Large margin principles for feature selection. In Isabelle Guyon, Masoud Nikravesh, Steve Gunn, and Lotfi A. Zadeh, editors, *Feature Extraction*, volume 207, pages 585–606. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[73] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni,

M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg. A protein interaction map of drosophila melanogaster. *Science*, 302(5651):1727–1736, December 2003.

[74] A. Glucksmann. CELL DEATHS IN NORMAL VERTEBRATE ONTOGENY. *Biological Reviews*, 26(1):59–86, February 1951.

[75] CHEMIE.DE Information Service GmbH. Chemie.de, portal especializado en inglés, clon de quimica.es. chemie.de gestiona portales científicos especializados., 2011.

[76] CHEMIE.DE Information Service GmbH. Quimica.es, portal especializado. chemie.de gestiona portales científicos especializados., 2011.

[77] Heike Goehler, Maciej Lalowski, Ulrich Stelzl, Stephanie Waelter, Martin Stroedicke, Uwe Worm, Anja Droege, Katrin S. Lindenberg, Maria Knoblich, Christian Haenig, Martin Herbst, Jaana Suopanki, Eberhard Scherzinger, Claudia Abraham, Bianca Bauer, Renate Hasenbank, Anja Fritzsche, Andreas H. Ludewig, Konrad Buessow, Sarah H. Coleman, Claire-Anne Gutekunst, Bernhard G. Landwehrmeyer, Hans Lehrach, and Erich E. Wanker. A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to huntington's disease. *Molecular Cell*, 15(6):853–865, September 2004.

[78] Debra S. Goldberg and Frederick P. Roth. Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences of the United States of America*, 100(8):4372 –4376, April 2003.

[79] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531 –537, October 1999.

[80] Alvaro J González and Li Liao. Predicting domain-domain interaction based on domain profiles with feature selection and support vector machines. *BMC Bioinformatics*, 11:537–537, 2010. PMID: 21034480 PMCID: 2989984.

[81] Griffiths. *Genética Moderna*. McGraw-Hill Interamericana de España, 2000.

[82] Alberto Guillen, Dusan Sovilj, Amaury Lendasse, Fernando Mateo, and Ignacio Rojas. Minimising the delta test for variable selection in regression problems.

International Journal of High Performance Systems Architecture, 1(4):269 – 281, 2008.

[83] A. Guillén, H. Pomares, J. González, I. Rojas, O. Valenzuela, and B. Prieto. Parallel multiobjective memetic RBFNNs design and feature selection for function approximation problems. *Neurocomput.*, 72(16-18):3541–3555, October 2009. ACM ID: 1609258.

[84] U. Guldener, M. Munsterkotter, G. Kastenmuller, N. Strack, J. van Helden, C. Lemer, J. Richelles, S. J. Wodak, J. Garcia-Martinez, J. E. Perez-Ortin, H. Michael, A. Kaps, E. Talla, B. Dujon, B. Andre, J. L. Souciet, J. De Montigny, E. Bon, C. Gaillardin, and H. W. Mewes. CYGD: the comprehensive yeast genome database. *Nucl. Acids Res.*, 33(suppl_1):D364–368, January 2005.

[85] Luke Hakes, John W Pinney, David L Robertson, and Simon C Lovell. Protein-protein interaction networks and biology[mdash]what's the connection? *Nat Biotech*, 26(1):69–72, January 2008.

[86] J A Hanley and B J McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839 –843, 1983.

[87] Henning Hermjakob, Luisa Montecchi-Palazzi, Gary Bader, Jérôme Wojcik, Lukasz Salwinski, Arnaud Ceol, Susan Moore, Sandra Orchard, Ugis Sarkans, Christian von Mering, Bernd Roechert, Sylvain Poux, Eva Jung, Henning Mersch, Paul Kersey, Michael Lappe, Yixue Li, Rong Zeng, Debashis Rana, Macha Nikolski, Holger Husi, Christine Brun, K Shanker, Seth G N Grant, Chris Sander, Peer Bork, Weimin Zhu, Akhilesh Pandey, Alvis Brazma, Bernard Jacq, Marc Vidal, David Sherman, Pierre Legrain, Gianni Cesareni, Ioannis Xenarios, David Eisenberg, Boris Steipe, Chris Hogue, and Rolf Apweiler. The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nature Biotechnology*, 22(2):177–183, February 2004. PMID: 14755292.

[88] L.J. Herrera, H. Pomares, I. Rojas, A. Guillén, A. Prieto, and O. Valenzuela. Recursive prediction for long term time series forecasting using advanced models. *Neurocomputing*, 70(16-18):2870–2880, October 2007.

[89] L.J. Herrera, H. Pomares, I. Rojas, O. Valenzuela, and A. Prieto. TaSe, a taylor series-based fuzzy system model that combines interpretability and accuracy. *Fuzzy Sets and Systems*, 153(3):403–427, August 2005.

[90] Richard G. Hibbert, Francesca Mattiroli, and Titia K. Sixma. Structural aspects of multi-domain RING/Ubox e3 ligases in DNA repair. *DNA Repair*, 8(4):525–535, April 2009.

[91] Desmond J. Higham, Gabriela Kalna, and J. Keith Vass. Spectral analysis of two-signed microarray expression data. *Mathematical Medicine and Biology*, 24(2):131–148, June 2007.

[92] Yuen Ho, Albrecht Gruhler, Adrian Heilbut, Gary D. Bader, Lynda Moore, Sally-Lin Adams, Anna Millar, Paul Taylor, Keiryn Bennett, Kelly Boutilier, Lingyun Yang, Cheryl Wolting, Ian Donaldson, Soren Schandorff, Juanita Shewnarane, Mai Vo, Joanne Taggart, Marilyn Goudreault, Brenda Muskat, Cris Alfarano, Danielle Dewar, Zhen Lin, Katerina Michalickova, Andrew R. Willems, Holly Sassi, Peter A. Nielsen, Karina J. Rasmussen, Jens R. Andersen, Lene E. Johansen, Lykke H. Hansen, Hans Jespersen, Alexandre Podtelejnikov, Eva Nielsen, Janne Crawford, Vibeke Poulsen, Birgitte D. Sorensen, Jesper Matthiesen, Ronald C. Hendrickson, Frank Gleeson, Tony Pawson, Michael F. Moran, Daniel Durocher, Matthias Mann, Christopher W. V. Hogue, Daniel Figeys, and Mike Tyers. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180–183, January 2002.

[93] R. Hoffmann and A. Valencia. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, 21(Suppl 2):ii252–ii258, October 2005.

[94] Frank C. P. Holstege, Ezra G. Jennings, John J. Wyrick, Tong Ihn Lee, Christoph J. Hengartner, Michael R. Green, Todd R. Golub, Eric S. L, er, and Richard A. Young. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95(5):717–728, November 1998.

[95] Cheng B. Huang. *PhD dissertation: Multiscale Computational Methods for Morphogenesis and Algorithms for Protein-Protein Interaction Inference*. PhD thesis, Dept. of Computer Science and Eng., Univ. of Notre Dame, USA, 2005.

[96] Chengbang Huang, Faruck Morcos, Simon P Kanaan, Stefan Wuchty, Danny Z Chen, and Jesús A Izaguirre. Predicting protein-protein interactions from protein domains using a set cover approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM*, 4(1):78–87, 2007. PMID: 17277415.

[97] Won-Ki Huh, James V. Falvo, Luke C. Gerke, Adam S. Carroll, Russell W. Howson, Jonathan S. Weissman, and Erin K. O'Shea. Global analysis of protein localization in budding yeast. *Nature*, 425(6959):686–691, October 2003.

[98] J. Hwang, R. T. Saffert, and R. F. Kalejta. Elongin B-Mediated epigenetic alteration of viral chromatin correlates with efficient human cytomegalovirus gene expression and replication. *mBio*, 2(2):e00023–11–e00023–11, March 2011.

[99] Integromics. Documentación del seminario bioinformatica aplicada a la proteómica (curso práctico) impartido por integromics s.l. en cordoba. Technical report, Integromics S.L., 2007.

[100] Ivan Iossifov, Michael Krauthammer, Carol Friedman, Vasileios Hatzivassiloglou, Joel S. Bader, Kevin P. White, and Andrey Rzhetsky. Probabilistic inference of molecular networks from noisy data sources. *Bioinformatics*, 20(8):1205–1213, May 2004.

[101] Takashi Ito, Tomoko Chiba, Ritsuko Ozawa, Mikio Yoshida, Masahira Hattori, and Yoshiyuki Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8):4569–4574, April 2001.

[102] Alexey V. Ivanov, Hongzhuang Peng, Vyacheslav Yurchenko, Kyoko L. Yap, Dmitri G. Negorev, David C. Schultz, Elyse Psulkowski, William J. Fredericks, David E. White, Gerd G. Maul, Moshe J. Sadofsky, Ming-Ming Zhou, and Frank J. Rauscher. PHD Domain-Mediated e3 ligase activity directs intramolecular sumoylation of an adjacent bromodomain required for gene silencing. *Molecular Cell*, 28(5):823–837, December 2007.

[103] Sarah Jackson and Yue Xiong. CRL4s: the CUL4-RING e3 ubiquitin ligases. *Trends in biochemical sciences*, 34(11):562–570, November 2009. PMID: 19818632 PMCID: 2783741.

[104] M D Jacobson, M Weil, and M C Raff. Programmed cell death in animal development. *Cell*, 88(3):347–354, February 1997. PMID: 9039261.

[105] Ronald Jansen and Mark Gerstein. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Current Opinion in Microbiology*, 7(5):535–545, October 2004.

[106] Ronald Jansen, Haiyuan Yu, Dov Greenbaum, Yuval Kluger, Nevan J. Krogan, Sambath Chung, Andrew Emili, Michael Snyder, Jack F. Greenblatt, and Mark Gerstein. A bayesian networks approach for predicting Protein-Protein interactions from genomic data. *Science*, 302(5644):449–453, October 2003.

[107] F. Jelinek, L. Bahl, and R. Mercer. Design of a linguistic statistical decoder for the recognition of continuous speech. *Information Theory, IEEE Transactions on*, 21(3):250– 256, May 1975.

[108] Célia Jeronimo, Diane Forget, Annie Bouchard, Qintong Li, Gordon Chua, Christian Poitras, Cynthia Thérien, Dominique Bergeron, Sylvie Bourassa, Jack Greenblatt, Benoit Chabot, Guy G Poirier, Timothy R Hughes, Mathieu Blanchette, David H Price, and Benoit Coulombe. Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme. *Molecular Cell*, 27(2):262–274, July 2007. PMID: 17643375.

[109] Taijiao Jiang and Amy E Keating. AVID: an integrative framework for discovering functional relationships among proteins. *BMC Bioinformatics*, 6:136–136, 2005. PMID: 15929793 PMCID: 1177925.

[110] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. *null*, pages 137—142, 1998.

[111] George John, Ron Kohavi, and Karl Pfleger. Irrelevant features and the subset selection problem. In *International Conference on Machine Learning*, pages 121–129, 1994.

[112] Linda S Kaltenbach, Eliana Romero, Robert R Becklin, Rakesh Chettier, Russell Bell, Amit Phansalkar, Andrew Strand, Cameron Torcassi, Justin Savage, Anthony Hurlburt, Guang-Ho Cha, Lubna Ukani, Cindy Lou Chepanoske, Yuejun Zhen, Sudhir Sahasrabudhe, James Olson, Cornelia Kurschner, Lisa M Ellerby, John M Peltier, Juan Botas, and Robert E Hughes. Huntingtin interacting proteins are genetic modifiers of neurodegeneration. *PLoS Genetics*, 3(5), May 2007. PMID: 17500595 PMCID: 1866352.

[113] Kohichi Kawahara, Makoto Hashimoto, Pazit Bar-On, Gilbert J. Ho, Leslie Crews, Hideya Mizuno, Edward Rockenstein, Syed Z. Imam, and Eliezer Masliah. α-Synuclein aggregates interfere with parkin solubility and distribution. *Journal of Biological Chemistry*, 283(11):6979 –6987, March 2008.

[114] J F Kerr, A H Wyllie, and A R Currie. Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *British Journal of Cancer*, 26(4):239–257, August 1972. PMID: 4561027.

[115] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human protein

reference database–2009 update. *Nucleic Acids Research*, 37(Database):D767–D772, January 2009.

[116] Inyoung Kim, Yin Liu, and Hongyu Zhao. Bayesian methods for predicting interacting protein pairs using domain information. *Biometrics*, 63(3):824–833, 2007.

[117] Kenji Kira and Larry A. Rendell. A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning*, pages 249–256, Aberdeen, Scotland, United Kingdom, 1992. Morgan Kaufmann Publishers Inc.

[118] Dan E. Krane and Michael L. Raymer. *Fundamental Concepts of Bioinformatics*. Benjamin Cummings, 1 edition, September 2002.

[119] Solomon Kullback. *Information theory and statistics*. Courier Dover Publications, 1997.

[120] A. Kumar. Subcellular localization of the yeast proteome. *Genes & Development*, 16(6):707–719, March 2002.

[121] Ben Lehner and Christopher M. Sanderson. A protein interaction framework for human mRNA degradation. *Genome Research*, 14(7):1315 –1323, July 2004.

[122] Albert L. Lehninger, David L. Nelson, and Michael M. Cox. Principles of biochemistry 2nd ed. *Journal of Chemical Education*, 70(8):A223, 1993.

[123] Stanley Letovsky and Simon Kasif. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, 19(suppl_1):i197–204, July 2003.

[124] Siming Li, Christopher M. Armstrong, Nicolas Bertin, Hui Ge, Stuart Milstein, Mike Boxem, Pierre-Olivier Vidalain, Jing-Dong J. Han, Alban Chesneau, Tong Hao, Debra S. Goldberg, Ning Li, Monica Martinez, Jean-Francois Rual, Philippe Lamesch, Lai Xu, Muneesh Tewari, Sharyl L. Wong, Lan V. Zhang, Gabriel F. Berriz, Laurent Jacotot, Philippe Vaglio, Jerome Reboul, Tomoko Hirozane-Kishikawa, Qianru Li, Harrison W. Gabel, Ahmed Elewa, Bridget Baumgartner, Debra J. Rose, Haiyuan Yu, Stephanie Bosak, Reynaldo Sequerra, Andrew Fraser, Susan E. Mango, William M. Saxton, Susan Strome, Sander van den Heuvel, Fabio Piano, Jean Vandenhaute, Claude Sardet, Mark Gerstein, Lynn Doucette-Stamm, Kristin C. Gunsalus, J. Wade Harper, Michael E. Cusick, Frederick P. Roth, David E. Hill, and Marc Vidal. A map of the interactome network of the metazoan c. elegans. *Science*, 303(5657):540–543, January 2004.

[125] Wei Li and Yihong Ye. Polyubiquitin chains: functions, structures, and mechanisms. *Cellular and molecular life sciences : CMLS*, 65(15):2397–2406, August 2008. PMID: 18438605 PMCID: 2700825.

[126] Esti Liani, Allon Eyal, Eyal Avraham, Revital Shemer, Raymonde Szargel, Daniela Berg, Antje Bornemann, Olaf Riess, Christopher A. Ross, Ruth Rott, and Simone Engelender. Ubiquitylation of synphilin-1 and α-synuclein by SIAH and its presence in cellular inclusions and lewy bodies imply a role in parkinson's disease. *Proceedings of the National Academy of Sciences of the United States of America*, 101(15):5500–5505, April 2004. PMID: 15064394 PMCID: 397412.

[127] Janghoo Lim, Tong Hao, Chad Shaw, Akash J. Patel, Gábor Szabó, Jean-François Rual, C. Joseph Fisk, Ning Li, Alex Smolyar, David E. Hill, Albert-László Barabási, Marc Vidal, and Huda Y. Zoghbi. A Protein–Protein interaction network for human inherited ataxias and disorders of purkinje cell degeneration. *Cell*, 125(4):801–814, May 2006.

[128] Michael T. Lin and M. Flint Beal. Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases. *Nature*, 443(7113):787–795, October 2006.

[129] Yin Liu, Inyoung Kim, and Hongyu Zhao. Protein interaction predictions from diverse sources. *Drug Discovery Today*, 13(9-10):409–416, May 2008.

[130] Christian S Lobsiger and Don W Cleveland. Glial cells as intrinsic components of non-cell-autonomous neurodegenerative disease. *Nat Neurosci*, 10(11):1355–1360, November 2007.

[131] Xiongbin Lu, Ou Ma, Thuy-Ai Nguyen, Stephen N. Jones, Moshe Oren, and Lawrence A. Donehower. The wip1 phosphatase acts as a gatekeeper in the p53-Mdm2 autoregulatory loop. *Cancer Cell*, 12(4):342–354, October 2007.

[132] JB Macqueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[133] Marc Maillet. *Biología celular*. Masson, 2002.

[134] Edward M. Marcotte, Matteo Pellegrini, Ho-Leung Ng, Danny W. Rice, Todd O. Yeates, and David Eisenberg. Detecting protein function and Protein-Protein interactions from genome sequences. *Science*, 285(5428):751–753, July 1999.

[135] Gabriel Markson, Christina Kiel, Russell Hyde, Stephanie Brown, Panagoula Charalabous, Anja Bremm, Jennifer Semple, Jonathan Woodsmith, Simon Duley, Kourosh Salehi-Ashtiani, Marc Vidal, David Komander, Luis Serrano, Paul

Lehner, and Christopher M. Sanderson. Analysis of the human e2 ubiquitin conjugating enzyme protein interaction network. *Genome Research*, 19(10):1905 –1911, October 2009.

[136] J.P. Marques de Sá. *Pattern Recognition: Concepts, Methods and Applications*. Springer, 1 edition, October 2001.

[137] J. Lawrence Marsh, Tamas Lukacsovich, and Leslie Michels Thompson. Animal models of polyglutamine diseases and therapeutic approaches. *Journal of Biological Chemistry*, 284(12):7431 –7435, March 2009.

[138] David Martin, Christine Brun, Elisabeth Remy, Pierre Mouren, Denis Thieffry, and Bernard Jacq. GOToolBox: functional analysis of gene datasets based on gene ontology. *Genome Biology*, 5(12):R101, 2004.

[139] Lisa Matthews, Gopal Gopinath, Marc Gillespie, Michael Caudy, David Croft, Bernard de Bono, Phani Garapati, Jill Hemish, Henning Hermjakob, Bijay Jassal, Alex Kanapin, Suzanna Lewis, Shahana Mahajan, Bruce May, Esther Schmidt, Imre Vastrik, Guanming Wu, Ewan Birney, Lincoln Stein, and Peter D'Eustachio. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research*, 37(Database issue):D619–D622, January 2009. PMID: 18981052 PMCID: 2686536.

[140] Martin Meckesheimer, Andrew J Booker, Russell R Barton, and Timothy W Simpson. Computationally inexpensive metamodel assessment strategies. *AIAA JOURNAL*, 40:2053—2060, 2002.

[141] Joseph C. Mellor, Itai Yanai, Karl H. Clodfelter, Julian Mintseris, and Charles DeLisi. Predictome: a database of putative functional links between proteins. *Nucleic Acids Research*, 30(1):306 –309, January 2002.

[142] A. Mercado and G Gamba. Biología molecular en medicina. vii. hibridación molecular. *Revista de Investigación Clínica*, 49:75–8, 1997.

[143] Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, 31(1):258 –261, January 2003.

[144] Meredith B. Metzger and Allan M. Weissman. Working on a chain: E3s ganging up for ubiquitylation. *Nat Cell Biol*, 12(12):1124–1126, December 2010.

[145] H. W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Guldener, G. Mannhaupt, M. Munsterkotter, P. Pagel, N. Strack, V. Stumpflen, J. Warfsmann, and A. Ruepp. MIPS: analysis and annotation of proteins from whole genomes. *Nucl. Acids Res.*, 32(suppl_1):D41–44, January 2004.

[146] Marcin Mizianty and Lukasz Kurgan. Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. *BMC Bioinformatics*, 10(1):414, 2009.

[147] Francisco Mora Teruel and José María Segovia de Arana. *Enfermedades neurodegenerativas*. Farmaindustria, 2002.

[148] Ralf Mrowka, Andreas Patzak, and Hanspeter Herzel. Is there a bias in proteome research? *Genome Research*, 11(12):1971 –1973, December 2001.

[149] James Munkres. *Topology*. Prentice Hall, 2 edition, January 2000.

[150] Manabu Nakayama, Reiko Kikuno, and Osamu Ohara. Protein–Protein interactions between large proteins: Two-Hybrid screening using a functionally classified library composed of long cDNAs. *Genome Research*, 12(11):1773 –1784, November 2002.

[151] Kim Nasmyth. How do so few control so many? *Cell*, 120(6):739–746, March 2005.

[152] Thanh Phuong Nguyen and Tu Bao Ho. Combining domain fusions and Domain-Domain interactions to predict Protein-Protein interactions. In *BIOKDD 2007: 7th Workshop on Data Mining in Bioinformatics*, 2007.

[153] Tuan Nguyen and James A. Goodrich. Protein-Protein interaction assays: eliminating false positive interactions. *Nat Methods.*, 2006.

[154] D. J. O' Shannessy, M. Brighamburke, K. K. Soneson, P. Hensley, and I. Brooks. Determination of rate and equilibrium binding constants for macromolecular interactions using surface plasmon resonance: Use of nonlinear least squares analysis methods. *Analytical Biochemistry*, 212(2):457–468, August 1993.

[155] Melanie D. Ohi, Craig W. Vander Kooi, Joshua A. Rosenberg, Walter J. Chazin, and Kathleen L. Gould. Structural insights into the u-box, a domain associated with multi-ubiquitination. *Nat Struct Mol Biol*, 10(4):250–255, April 2003.

[156] Rafael Oliva, Francisca Ballesta, Josep Oriola, and Joan Claria. *Genética médica*. Edicions Universitat Barcelona, 2008.

[157] Birgitte B. Olsen and Barbara Guerra. Ability of CK2β to selectively regulate cellular protein kinases. *Molecular and Cellular Biochemistry*, 316(1-2):115–126, June 2008.

[158] Shao-En Ong, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen, Akhilesh Pandey, and Matthias Mann. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*, pages M200025–MCP200, May 2002.

[159] D J O'Shannessy, M Brigham-Burke, K K Soneson, P Hensley, and I Brooks. Determination of rate and equilibrium binding constants for macromolecular interactions by surface plasmon resonance. *Methods in Enzymology*, 240:323–349, 1994. PMID: 7823837.

[160] Philipp Pagel, Stefan Kovac, Matthias Oesterheld, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Pekka Mark, Volker Stümpflen, Hans-Werner Mewes, Andreas Ruepp, and Dmitrij Frishman. The MIPS mammalian protein–protein interaction database. *Bioinformatics*, 21(6):832 –834, March 2005.

[161] Ricardo Paniagua. *Citología e histología vegetal y animal*. McGraw-Hill Interamericana de España, 2007.

[162] Ashwini Patil and Haruki Nakamura. Filtering high-throughput protein-protein interaction data using a combination of genomic features. *BMC Bioinformatics*, 6(1):100, 2005.

[163] Ashwini Patil and Haruki Nakamura. HINT - a database of annotated protein-protein interactions and their homologs. *Biophysics*, 2005.

[164] Matteo Pellegrini, David Haynor, and Jason M Johnson. Protein interaction networks. *Expert Review of Proteomics*, 1(2):239–249, August 2004.

[165] Matteo Pellegrini, Edward M. Marcotte, Michael J. Thompson, David Eisenberg, and Todd O. Yeates. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 96(8):4285–4288, April 1999.

[166] Hanchuan Peng, Fuhui Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, August 2005.

[167] Matthew D. Petroski and Raymond J. Deshaies. Function and regulation of cullin-RING ubiquitin ligases. *Nat Rev Mol Cell Biol*, 6(1):9–20, January 2005.

[168] Cecile M. Pickart. Mechanisms underlying ubiquitination. *Annual Review of Biochemistry*, 70(1):503–533, June 2001.

[169] Cecile M. Pickart and Michael J. Eddins. Ubiquitin: structures, functions, mechanisms. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1695(1-3):55–72, November 2004.

[170] S. Pita Fernandez and S. Pertegas Diaz. Cad. aten. primaria. pruebas diagnósticas: Sensibilidad y especificidad. Technical Report 10, 120-124, Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Universitario de A Coruña (España), 2003.

[171] Christina Priller, Thomas Bauer, Gerda Mitteregger, Bjarne Krebs, Hans A. Kretzschmar, and Jochen Herms. Synapse formation and function is modulated by the amyloid precursor protein. *The Journal of Neuroscience*, 26(27):7212 –7221, July 2006.

[172] Ekaterini Pringa, Gustavo Martinez-Noel, Ursula Müller, and Klaus Harbers. Interaction of the RING finger-related u-box motif of a nuclear dot protein with ubiquitin-conjugating enzymes. *Journal of Biological Chemistry*, 276(22):19617 –19623, June 2001.

[173] Yanjun Qi, Judith Klein-Seetharaman, and Ziv Bar-Joseph. Random forest similarity for protein-protein interaction prediction from multiple sources. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 531–542, 2005. PMID: 15759657.

[174] Martin Raff. Cell suicide for beginners. *Nature*, 396(6707):119, November 1998.

[175] Amir Navot Ran Gilad-Bachrach and Naftali Tishby. Margin based feature selection: Theory and algorithms. *In Proc. of the 21'st ICML*, pages 43–50, 2004.

[176] Jeffrey A. Ranish, Eugene C. Yi, Deena M. Leslie, Samuel O. Purvine, David R. Goodlett, Jimmy Eng, and Ruedi Aebersold. The study of macromolecular complexes by quantitative proteomics. *Nat Genet*, 33(3):349–355, March 2003.

[177] Daniel R Rhodes, Scott A Tomlins, Sooryanarayana Varambally, Vasudeva Mahavisno, Terrence Barrette, Shanker Kalyana-Sundaram, Debashis Ghosh, Akhilesh Pandey, and Arul M Chinnaiyan. Probabilistic model of the human protein-protein interaction network. *Nat Biotech*, 23(8):951–959, 2005.

[178] I. Rojas, H. Pomares, J. Gonzáles, J. L. Bernier, E. Ros, F. J. Pelayo, and A. Prieto. Analysis of the functional block involved in the design of radial basis function networks. *Neural Processing Letters*, 12(1):1–17, 2000.

[179] Rosetta. Rosetta biosoftware and GeneGo to offer interoperability between the rosetta resolver system and MetaCor.

[180] Rosfuzah Roslan, Razib M. Othman, Zuraini A. Shah, Shahreen Kasim, Hishammuddin Asmuni, Jumail Taliba, Rohayanti Hassan, and Zalmiyah Zakaria. Utilizing shared interacting domain patterns and gene ontology information to improve protein-protein interaction prediction. *Computers in Biology and Medicine*, 40(6):555–564, June 2010.

[181] Jean-Francois Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amelie Dricot, Ning Li, Gabriel F. Berriz, Francis D. Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, Niels Klitgord, Christophe Simon, Mike Boxem, Stuart Milstein, Jennifer Rosenberg, Debra S. Goldberg, Lan V. Zhang, Sharyl L. Wong, Giovanni Franklin, Siming Li, Joanna S. Albala, Janghoo Lim, Carlene Fraughton, Estelle Llamosas, Sebiha Cevik, Camille Bex, Philippe Lamesch, Robert S. Sikorski, Jean Vandenhaute, Huda Y. Zoghbi, Alex Smolyar, Stephanie Bosak, Reynaldo Sequerra, Lynn Doucette-Stamm, Michael E. Cusick, David E. Hill, Frederick P. Roth, and Marc Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, October 2005.

[182] Vladimir Ruiz Álvarez and Rafael Menéndez Lorenzo. Métodos y aplicaciones de la biología molecular en biomedicina. Technical report, Laboratorio de Bioquímica y Fisiología y la Facultad de Ciencias Médicas Calixto García de La Habana (Cuba), 2005.

[183] Miia M. Rytinki, Sanna Kaikkonen, Petri Pehkonen, Tiina Jääskeläinen, and Jorma J. Palvimo. PIAS proteins: pleiotropic interactors associated with SUMO. *Cellular and Molecular Life Sciences*, 66(18):3029–3041, June 2009.

[184] Ramazan Saeed and Charlotte Deane. Protein protein interactions, evolutionary rate, abundance and age. *BMC Bioinformatics*, 7(1):128, 2006.

[185] Ramazan Saeed and Charlotte Deane. An assessment of the uses of homologous interactions. *Bioinformatics*, 24(5):689–695, March 2008.

[186] Lukasz Salwinski, Christopher S. Miller, Adam J. Smith, Frank K. Pettit, James U. Bowie, and David Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32(Database issue):D449–D451, January 2004. PMID: 14681454 PMCID: 308820.

[187] Martin Scheffner, Ulrike Nuber, and Jon M. Huibregtse. Protein ubiquitination involving an E1-E2-E3 enzyme ubiquitin thioester cascade. *Nature*, 373(6509):81–83, January 1995.

[188] Mark J. Schervish. P values: What they are and what they are not. *The American Statistician*, 50(3):203–206, 1996. ArticleType: research-article / Full publication date: Aug., 1996 / Copyright © 1996 American Statistical Association.

[189] Anne Schreiber, Florian Stengel, Ziguo Zhang, Radoslav I. Enchev, Eric H. Kong, Edward P. Morris, Carol V. Robinson, Paula C. A. da Fonseca, and David Barford. Structural basis for the subunit assembly of the anaphase-promoting complex. *Nature*, 470(7333):227–232, February 2011.

[190] Alan L. Schwartz and Aaron Ciechanover. Targeting proteins for destruction by the ubiquitin system: Implications for human pathobiology. *Annual Review of Pharmacology and Toxicology*, 49(1):73–96, February 2009.

[191] Benno Schwikowski, Peter Uetz, and Stanley Fields. A network of protein-protein interactions in yeast. *Nat Biotech*, 18(12):1257–1261, December 2000.

[192] Florence Servant, Catherine Bru, Sébastien Carrère, Emmanuel Courcelle, Jerjme Gouzy, David Peyruc, and Daniel Kahn. ProDom: automated clustering of homologous domains. *Briefings in Bioinformatics*, 3(3):246–251, September 2002. PMID: 12230033.

[193] Noula Shembade, Averil Ma, and Edward W. Harhaj. Inhibition of NF-κB signaling by a20 through disruption of ubiquitin enzyme complexes. *Science*, 327(5969):1135 –1139, February 2010.

[194] Christian J A Sigrist, Lorenzo Cerutti, Nicolas Hulo, Alexandre Gattiker, Laurent Falquet, Marco Pagni, Amos Bairoch, and Philipp Bucher. PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatics*, 3(3):265–274, September 2002. PMID: 12230035.

[195] Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, and Trey Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)*, 27(3):431–432, February 2011. PMID: 21149340.

[196] Antti Sorjamaa, Jin Hao, Nima Reyhani, Yongnan Ji, and Amaury Lendasse. Methodology for long-term prediction of time series. *Neurocomput.*, 70(16-18):2861–2869, October 2007. ACM ID: 1316131.

[197] E M Southern. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*, 98(3):503–517, November 1975. PMID: 1195397.

[198] Mathew E. Sowa, Eric J. Bennett, Steven P. Gygi, and J. Wade Harper. Defining the human deubiquitinating enzyme interaction landscape. *Cell*, 138(2):389–403, July 2009.

[199] C. Stark. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(90001):D535–D539, January 2006.

[200] Alexander Statnikov, Lily Wang, and Constantin F Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, 9:319–319, 2008. PMID: 18647401 PMCID: 2492881.

[201] Amelie Stein, Robert B. Russell, and Patrick Aloy. 3did: interacting protein domains of known three-dimensional structure. *Nucl. Acids Res.*, 33(suppl_1):D413–417, January 2005.

[202] Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H. Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, and Susanne Koeppen. A human Protein-Protein interaction network: A resource for annotating the proteome. *Cell*, 122(6):957–968, September 2005.

[203] G Stoesser, W Baker, A van den Broek, E Camon, M Garcia-Pastor, C Kanz, T Kulikova, V Lombard, R Lopez, H Parkinson, N Redaschi, P Sterk, P Stoehr, and M A Tuli. The EMBL nucleotide sequence database. *Nucleic Acids Research*, 29(1):17–21, January 2001. PMID: 11125039.

[204] Johan A K Suykens, Tony Van Gestel, Jos De Brabanter, Bart De Moor, and Joos Vandewalle. *Least Squares Support Vector Machines*. World Scientific Publishing Company, January 2003.

[205] Z Tan, X Sun, F-S Hou, H-W Oh, L G W Hilgenberg, E M Hol, F W van Leeuwen, M A Smith, D K O'Dowd, and S S Schreiber. Mutant ubiquitin found in alzheimer's disease causes neuritic beading of mitochondria in association with neuronal degeneration. *Cell Death Differ*, 14(10):1721–1732, June 2007.

[206] Leslie Michels Thompson. Neurodegeneration: A question of balance. *Nature*, 452(7188):707–708, April 2008.

[207] H Towbin, T Staehelin, and J Gordon. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proceedings of the National Academy of Sciences of the United States of America*, 76(9):4350–4354, September 1979. PMID: 388439 PMCID: 411572.

[208] Kevin Truong and Mitsuhiko Ikura. Domain fusion analysis by applying relational algebra to protein sequence and domain databases. *BMC Bioinformatics*, 4(1):16, 2003.

[209] Andrew S. Turnell, Grant S. Stewart, Roger J. A. Grand, Susan M. Rookes, Ashley Martin, Hiroyuki Yamano, Stephen J. Elledge, and Phillip H. Gallimore. The APC/C and CBP/p300 cooperate to regulate transcription and cell-cycle progression. *Nature*, 438(7068):690–695, December 2005.

[210] Paul R. Turner, Kate O'Connor, Warren P. Tate, and Wickliffe C. Abraham. Roles of amyloid precursor protein and its fragments in regulating neural activity, plasticity and memory. *Progress in Neurobiology*, 70(1):1–32, May 2003.

[211] Daisuke Uchida, Shigetsugu Hatakeyama, Akemi Matsushima, Hongwei Han, Satoshi Ishido, Hak Hotta, Jun Kudoh, Nobuyoshi Shimizu, Vassilis Doucas, Keiichi I. Nakayama, Noriyuki Kuroda, and Mitsuru Matsumoto. AIRE functions as an e3 ubiquitin ligase. *The Journal of Experimental Medicine*, 199(2):167–172, January 2004. PMID: 14734522 PMCID: 2211764.

[212] Peter Uetz, Loic Giot, Gerard Cagney, Traci A. Mansfield, Richard S. Judson, James R. Knight, Daniel Lockshon, Vaibhav Narayan, Maithreyan Srinivasan, Pascale Pochart, Alia Qureshi-Emili, Ying Li, Brian Godwin, Diana Conover, Theodore Kalbfleisch, Govindan Vijayadamodar, Meijia Yang, Mark Johnston, Stanley Fields, and Jonathan M. Rothberg. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627, February 2000.

[213] Sjoerd J L van Wijk, Sjoerd J de Vries, Patrick Kemmeren, Anding Huang, Rolf Boelens, Alexandre M J J Bonvin, and H Th Marc Timmers. A comprehensive framework of E2-RING e3 interactions of the human ubiquitin-proteasome system. *Mol Syst Biol*, 5, 2009.

[214] Alexei Vazquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Global protein function prediction from protein-protein interaction networks. *Nat Biotech*, 21(6):697–700, June 2003.

[215] Thiago M Venancio, S Balaji, Lakshminarayan M Iyer, and L Aravind. Reconstructing the ubiquitin network - cross-talk with other systems and identification of novel functions. *Genome Biology*, 10(3):R33, 2009.

[216] Kavitha Venkatesan, Jean-Francois Rual, Alexei Vazquez, Ulrich Stelzl, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Martina Zenkner, Xiaofeng

Xin, Kwang-Il Goh, Muhammed A Yildirim, Nicolas Simonis, Kathrin Heinzmann, Fana Gebreab, Julie M Sahalie, Sebiha Cevik, Christophe Simon, Anne-Sophie de Smet, Elizabeth Dann, Alex Smolyar, Arunachalam Vinayagam, Haiyuan Yu, David Szeto, Heather Borick, Amelie Dricot, Niels Klitgord, Ryan R Murray, Chenwei Lin, Maciej Lalowski, Jan Timm, Kirstin Rau, Charles Boone, Pascal Braun, Michael E Cusick, Frederick P Roth, David E Hill, Jan Tavernier, Erich E Wanker, Albert-Laszlo Barabasi, and Marc Vidal. An empirical framework for binary interactome mapping. *Nat Meth*, 6(1):83–90, January 2009.

[217] Arunachalam Vinayagam, Coral del Val, Falk Schubert, Roland Eils, Karl-Heinz Glatting, Sandor Suhai, and Rainer Konig. GOPET: a tool for automated predictions of gene ontology terms. *BMC Bioinformatics*, 7(1):161, 2006.

[218] Panagiotis J Vlachostergios, Anna Patrikidou, Danai D Daliani, and Christos N Papandreou. The ubiquitin-proteasome system in cancer, a major player in DNA repair. part 1: post-translational regulation. *Journal of Cellular and Molecular Medicine*, 13(9b):3006–3018, September 2009.

[219] Panagiotis J Vlachostergios, Anna Patrikidou, Danai D Daliani, and Christos N Papandreou. The ubiquitin-proteasome system in cancer, a major player in DNA repair. part 2: transcriptional regulation. *Journal of Cellular and Molecular Medicine*, 13(9b):3019–3031, September 2009.

[220] Christian von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G. Oliver, Stanley Fields, and Peer Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, May 2002.

[221] Albertha J. M. Walhout, Raffaella Sordella, Xiaowei Lu, James L. Hartley, Gary F. Temple, Michael A. Brasch, Nicolas Thierry-Mieg, and Marc Vidal. Protein interaction mapping in c. elegans using proteins involved in vulval development. *Science*, 287(5450):116–122, January 2000.

[222] H. Wang, F. Azuaje, O. Bodenreider, and J. Dopazo. Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In *Computational Intelligence in Bioinformatics and Computational Biology, 2004. CIBCB '04. Proceedings of the 2004 IEEE Symposium on*, pages 25–31, 2004.

[223] Alexander Weinstein. Homologous genes and linear linkage in drosophila virilis. *Proceedings of the National Academy of Sciences of the United States of America*, 6(11):625–639, November 1920. ArticleType: research-article / Full publication date: Nov. 15, 1920 / Copyright © 1920 National Academy of Sciences.

[224] Gary L Wenk. Neuropathologic changes in alzheimer's disease. *The Journal of Clinical Psychiatry*, 64 Suppl 9:7–10, 2003. PMID: 12934968.

[225] Ingrid E. Wertz, Karen M. O'Rourke, Honglin Zhou, Michael Eby, L. Aravind, Somasekar Seshagiri, Ping Wu, Christian Wiesmann, Rohan Baker, David L. Boone, Averil Ma, Eugene V. Koonin, and Vishva M. Dixit. De-ubiquitination and ubiquitin ligase domains of a20 downregulate NF-[kappa]B signalling. *Nature*, 430(7000):694–699, 2004.

[226] Andrew R. Willems, Michael Schwab, and Mike Tyers. A hitchhiker's guide to the cullin ubiquitin ligases: SCF and its kin. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1695(1-3):133–170, November 2004.

[227] C R Woese and G E Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, 74(11):5088–5090, November 1977. PMC432104.

[228] C R Woese, O Kandler, and M L Wheelis. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences of the United States of America*, 87(12):4576–4579, June 1990.

[229] Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. Probability estimates for multi-class classification by pairwise coupling. *J. Mach. Learn. Res.*, 5:975–1005, December 2004. ACM ID: 1016791.

[230] Xiaomei Wu, Lei Zhu, Jie Guo, Da-Yong Zhang, and Kui Lin. Prediction of yeast protein-protein interaction network: insights from the gene ontology and annotations. *Nucl. Acids Res.*, 34(7):2137–2150, April 2006.

[231] Hui Xiong, Danling Wang, Linan Chen, Yeun Su Choo, Hong Ma, Chengyuan Tang, Kun Xia, Wei Jiang, Ze'ev Ronai, Xiaoxi Zhuang, and Zhuohua Zhang. Parkin, PINK1, and DJ-1 form a ubiquitin e3 ligase complex promoting unfolded protein degradation. *The Journal of Clinical Investigation*, 119(3):650–660, March 2009. PMID: 19229105 PMCID: 2648688.

[232] Haiyuan Yu, Pascal Braun, Muhammed A. Yildirim, Irma Lemmens, Kavitha Venkatesan, Julie Sahalie, Tomoko Hirozane-Kishikawa, Fana Gebreab, Na Li, Nicolas Simonis, Tong Hao, Jean-Francois Rual, Amelie Dricot, Alexei Vazquez, Ryan R. Murray, Christophe Simon, Leah Tardivo, Stanley Tam, Nenad Svrzikapa, Changyu Fan, Anne-Sophie de Smet, Adriana Motyl, Michael E. Hudson, Juyong Park, Xiaofeng Xin, Michael E. Cusick, Troy Moore, Charlie

Boone, Michael Snyder, Frederick P. Roth, Albert-Laszlo Barabasi, Jan Tavernier, David E. Hill, and Marc Vidal. High-Quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, October 2008.

[233] Jiantao Yu, Maozu Guo, Chris J. Needham, Yangchao Huang, Lu Cai, and David R. Westhead. Simple sequence-based kernels do not predict protein–protein interactions. *Bioinformatics*, 26(20):2610 –2614, October 2010.

[234] Jingkai Yu and Russell L. Finley. Combining multiple positive training sets to generate confidence scores for protein-protein interactions. *Bioinformatics*, 25(1):105–111, January 2009.

[235] Wolfgang Zachariae, Andrej Shevchenko, Paul D. Andrews, Rafael Ciosk, Marta Galova, Michael J. R. Stark, Matthias Mann, and Kim Nasmyth. Mass spectrometric analysis of the Anaphase-Promoting complex from yeast: Identification of a subunit related to cullins. *Science*, 279(5354):1216 –1219, February 1998.

[236] Ronen Zaidel-Bar, Shalev Itzkovitz, Avi Ma'ayan, Ravi Iyengar, and Benjamin Geiger. Functional atlas of the integrin adhesome. *Nat Cell Biol*, 9(8):858–867, 2007.

[237] Nazar Zaki, Sanja Lazarova-Molnar, Wassim El-Hajj, and Piers Campbell. Protein-protein interaction based on pairwise similarity. *BMC Bioinformatics*, 10(1):150, 2009.

[238] Lan V Zhang, Sharyl L Wong, Oliver D King, and Frederick P Roth. Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics*, 5:38, 2004. PMC419405.

[239] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, February 2004. Mathematical Reviews number (MathSciNet): MR2051001; Zentralblatt MATH identifier: 02113743.

[240] Huiru Zheng, Haiying Wang, and D.H. Glass. Integration of genomic data for inferring protein complexes from global Protein–Protein interaction networks. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 38(1):5–16, 2008.

[241] Wenlai Zhou, Ping Zhu, Jianxun Wang, Gabriel Pascual, Kenneth A. Ohgi, Jean Lozach, Christopher K. Glass, and Michael G. Rosenfeld. Histone H2A monoubiquitination represses transcription by inhibiting RNA polymerase II transcriptional elongation. *Molecular Cell*, 29(1):69–80, January 2008.

[242] Erik S Zimmerman, Brenda A Schulman, and Ning Zheng. Structural assembly of cullin-RING ubiquitin ligase complexes. *Current Opinion in Structural Biology*, 20(6):714–721, December 2010.

APPENDIX A

INTRODUCCIÓN BIOLÓGICA

Este apéndice y los siguientes han sido incluido en la tesis para una mejor comprensión de los conceptos tan diferentes de los que se suelen usar en titulaciones de ingenierías en informática o asociadas a las tecnologías de la información y comunicaciones. Muchos de los conceptos usados en la tesis pueden encontrarse en el glosario, sin embargo tantos los apéndices como el glosario no son exclusivos, de hecho se ha realizado un gran esfuerzo haciendo acopio de todos los términos que se han considerado fundamentales para una comprensión relativamente aceptable de este trabajo de investigación para quienes no tuviesen una formación en dicho tema. En este apéndice se realiza una breve descripción de la teoría celular dando una pequeña reseña de los componentes que constituyen la célula de los diferentes organismos.

A.1 De la célula al ADN (ácido desoxirribonucleico)

A.1.1 La célula: Un poco de historia

Una definición básica de célula es la siguiente: la célula es la unidad funcional y estructural de todos los organismos vivos conocidos. Es la unidad más pequeña clasificada como viva. Algunos organismos, tales como la bacteria, son unicelulares (que están formados por una única célula). Mientras otros organismos, como el humano, son pluricelulares.

El término célula procede del latín cellula, diminutivo de cella. Fue acuñado por Robert Hooke en un libro en 1665 al recoger sus observaciones sobre los tejidos vegetales [50].

La teoría celular, fue postulada en 1839 por Matthias Jakob Schleiden and Theodor Schwann, que establecía que todos los organismos están compuestos por una o mas células. Todas las células provienen de células precedentes .Las funciones vitales de una organismos tienen lugar dentro de las células, y todas las células contienen la

información hereditaria necesaria para las funciones de regulación de la célula y para la transmisión de la información a la siguiente generación de células. Probablemente la novedad más importante de esta teoría con respecto a anteriores autores, es la de añadir un carácter fisiológico, y no sólo estructural, a la célula como unidad constituyente de los seres vivos [50].

Rudolf Virchow, uno de los fisiólogos más prominentes del siglo XIX, postuló que todas las células provienen de otras células, más en concreto, "Omnis cellula ex cellula" ("cada célula es derivada de otra célula [ya existente]"). Así pues a finales de los años 60 del siglo XIX, se había constituido el eje fundamental del estudio de los seres vivos.

A.1.2 Teoría celular

El concepto de célula como unidad anatómica y funcional de los organismos surgió entre los años 1830 y 1880, y por tanto la teoría celular que a partir de entonces se desarrolló afirma los siguientes postulados:

1. Que la célula es una unidad morfológica de todo ser vivo: es decir, que en los seres vivos todo está formado por células o por sus productos de secreción.

2. Este primer postulado sería completado por Rudolf Virchow con la afirmación "Omnis cellula ex cellula", la cual indica que toda célula deriva de una célula precedente (biogénesis). En otras palabras, este postulado constituye la refutación de la teoría de generación espontánea o ex novo, en la que se lanzó la hipótesis de que la posibilidad de que se generara vida a partir de elementos inanimados [161].

3. Un tercer postulado de la teoría celular indica que las funciones vitales de los organismos ocurren dentro de las células, o en su entorno inmediato, y son controladas por sustancias que ellas secretan. Cada célula es un sistema abierto, que intercambia materia y energía con su medio. En una célula ocurren todas las funciones vitales, de manera que basta una sola de ellas para tener un ser vivo (que será un ser vivo unicelular). Así pues, la célula es la unidad fisiológica de la vida.

4. Finalmente, el cuarto postulado de la teoría celular expresa que cada célula contiene toda la información hereditaria necesaria para el control de su propio ciclo y del desarrollo y el funcionamiento de un organismo de su especie, así como para la transmisión de esa información a la siguiente generación celular.

A.1.3 Definición de célula

Por tanto, el término célula agrupa a las células procariotas y a las células eucariotas. Por definición, la célula es la unidad más pequeña capaz de manifestar las propiedades del ser vivo. La célula sintetiza el conjunto, o casi, de sus constituyentes, utilizando elementos del medio extracelular. Crece y se multiplica. Está limitada por la membrana plasmática, que encierra un cierto número de orgánulos [133].

La célula eucariota contiene núcleo, orgánulo limitado por una envoltura que encierra el material genético en forma de ácido desoxirribonucleico (ADN), molécula fundamental de los cromosomas. Los protozoos están formados por una sola célula libre eucariota libre, a menudo capaz de moverse (amebas, paramecios); los metazoos son seres pluricelulares constituidos por células eucariotas agrupadas en tejidos (epiteliales -ver en glosario epitelio-, musculares, conjuntivos, de sostén -cartilaginoso, óseo-, nervioso). Con la excepción de los hematíes y las células nerviosas, la célula sufre un ciclo o alterna dos grandes fases, la fase de actividad funcional o interfase, y la fase de multiplicación o mitosis [133].

La célula procariota no posee un núcleo y nunca lo ha tenido; un solo cromosoma, formado también por ADN constituye el material genético; ninguna envoltura lo separa del citoplasma [133].

Los virus escapan a esta clasificación. Mientras están aislados, no manifiestan ninguna actividad vital. Por el contrario, cuando están en las células eucariotas o procariotas que infectan, su material genético (ADN o ARN - ácido ribonucleico-) se incorpora a la célula huésped, se replica y dirige la síntesis de proteínas virales [133].

También cabe mencionar las archaea (o arqueas), que son microorganismos unicelulares. Al igual que las bacterias, las archaea carecen de núcleo y por lo tanto son procariotas. Pero, las diferencias a nivel molecular entre arqueas y bacterias son tan enormes que se las clasifica en grupos distintos. Es por ello, estas diferencias son mayores de las que hay, por ejemplo, entre un animal y una planta. A día de hoy, se considera que las archaea están filogenéticamente más próximas a las eucariotas que a las bacterias. Viven en condiciones extremas.

Antes de continuar, debemos explicar los orgánulos que poseen las células.

A.1.4 La célula eucariota

En las células eucariotas, el ADN está separado del citoplasma por una envoltura que delimita el núcleo. Las células eucariotas poseen, además del núcleo, varios orgánulos característicos y específicos: retículo endoplasmático (RE), aparato de Golgi, mitocondrias, cloroplastos (en las células vegetales), endosomas, lisosomas, peroxisomas, citoesqueleto y centrosoma [133]. Las figuras A.1.4 representan las células eucariotas animales y vegetales.

A.1.4.1 Membrana Plasmática

La membrana plasmática es una estructura organizada y compleja y no sólo una simple interfaz que separa el medio intracelular del medio ambiente. Está compuesta por doble capa lipídica (asociada a proteínas intramembrana o periféricas). El papel de la membrana plasmática consiste en mantener la integridad de la célula que limita, permite o impide la entrada de moléculas del citoplasma. Posee las moléculas necesarias para la endocitosis (fagocitosis y pinocitosis) e interviene en el reconocimiento de aquellas que circulan en el medio extracelular, de desechos celulares u otras células

con las que se puede asociar (por medio de uniones intercelulares). Permite la comunicación de unas células con otras gracias a los receptores de membrana, que se unen específicamente a señales moleculares, los ligandos, elaborados y liberados por otras células. Además la membrana puede desarrollar microvellosidades en células especializadas en la absorción y uniones intercelulares allí donde sea necesaria una adhesión fuerte entre células o entre las células y la matriz extracelular. La membrana plasmática asociada al citoesqueleto participa en el mantenimiento de la forma y los movimientos de la célula [133].

A.1.4.2 Núcleo

El núcleo, estructura propia de los organismos eucariotas, es un orgánulo muy complejo que contiene, en forma de ADN, la información necesaria para el mantenimiento de las características y para la síntesis de proteínas específicas de la especie. El núcleo, redondo u ovalado, contiene, en el nucleoplasma, uno o dos nucléolos y los cromosomas (ADN). El ADN de las células eucariotas siempre está asociado a las proteínas histonas. El núcleo está limitado por la envoltura nuclear, una dependencia del retículo endoplasmático rugoso (RER) que separa el núcleo del citoplasma [133].

Los poros, aberturas de la envoltura, ofrecen a las sustancias exógenas o endógenas la posibilidad para transitar en sentido núcleo → citoplasma o en sentido citoplasma → núcleo. El ADN nuclear es el depositario de la mayor parte de la información genética ya que las mitocondrias también contienen ADN. En el núcleo se replica ADN, se corrigen los errores que se producen durante la replicación, y donde tiene lugar, en una molécula portadora, (el ARN mensajero o ARNm), la transcripción de la información que ha de ser descodificada y traducida en proteínas por los ribosomas en el citoplasma (proceso de traducción). El nucléolo contiene fibrillas y granos de ARNr que participan en la formación de ribosomas citoplasmáticos [133].

A.1.4.3 Retículo endoplasmático

El retículo endoplasmático (RE) es un conjunto de cavidades, consiste en una red de túbulos y sacos interconectados, que se extienden por todo el citoplasma y que llevan a cabo funciones celulares (incluida la síntesis proteica, la producción de esteroides, almacenamiento de glucógeno, etc.). Ocupa toda la célula, a excepción del exoplasma (parte periférica del citoplasma). El RE puede ser rugoso (RER), si los ribosomas se adhieren a la cara externa de su membrana, y liso (REL) si está desprovisto de ribosomas [133]. El REL procesa proteínas sintetizadas, y produce los lípidos de la célula que pasan a formar parte de las membranas celulares, principalmente. El retículo endoplasmático rugoso permite la síntesis de proteínas. Los ribosomas adheridos a él leen las moléculas de ARNm y, en función de la información descifrada, sintetizan proteínas destinadas a la exportación en vesículas de secreción, así como la mayor parte de proteínas de membrana. Los ribosomas libres traducen el ARNm en proteínas destinadas a citoplasma o a los distintos orgánulos [133].

Las proteínas sintetizadas por los ribosomas unidos al RER pueden ser de dos tipos:

- proteínas transmembrana, que se insertan directamente en la propia membrana del RER y que quedan con una orientación definitiva sea cual sea su membrana final de destino.
- proteínas que atraviesan completamente la membrana del retículo y son vertidas en su interior.

A.1.4.4 Mitocondrias

Las mitocondrias son elementos redondeados u ovalados de pequeñas dimensiones; están limitadas por dos membranas, una externa, periférica y lisa, y una interna, con crestas que incrementan su superficie. Las mitocondrias están constituidas por una matriz limitada por dos membranas.La membrana externa es muy permeable y la interna es poco permeable. La membrana interna dibuja crestas que se sumergen en la matriz y contienen la cadena respiratoria y el mecanismo para la síntesis de ATP (molécula que libera parte de la energía necesaria para el funcionamiento de la célula por una reacción química llamada hidrólisis). Contiene ADN específico, el ADN mitocondrial (ADNmt) [133].

A.1.4.5 Aparato de Golgi

El aparato de Golgi es un orgánulo que interviene tanto en la modificación (por eliminación o adición) de los azúcares de las glucoproteínas sintetizadas en el RER, como en la ordenación y clasificación de las proteínas procedentes del mismo. El aparato de Golgi está formado por una o más series de cisternas ligeramente curvas y aplanadas (son vesículas de pequeñas dimensiones) limitadas por membranas, y a este conjunto se le llama dictiosomas. El aparato de Golgi es el principal sitio de formación de nuevas membranas. Empaqueta proteínas en vesículas separadas y específicas para los productos que transportan [133, 50].

A.1.4.6 Endosomas

El endosoma es un orgánulo de las células animales delimitado por una sola membrana, que transporta material que se acaba de incorporar por endocitosis. Cuando a la misma se le introducen enzimas hidrolíticas (ver glosario hidrólisis) son transformados en lisosomas. Existen dos tipos de endosomas, dependiendo de la ubicación. Los endosomas, que incluyen los endosomas tempranos y los endosomas tardíos, son orgánulos permanentes, limitados por una membrana y situados en la ruta endocítica (ver en glosario endocitosis) [133].

El RE, el aparato de Golgi y los endosomas se agrupan funcionalmente en un sistema, el sistema de endomembranas, que reúne orgánulos y comportamientos celulares limitados por una membrana, y que se comunican entre ellos y con la membrana plasmática a través de vesículas.

A.1.4.7 Lisosomas

Los lisosomas son orgánulos limitados por una membrana que contienen un gran número de enzimas hidrolíticas (hidrolasas) capaces de lisar (ver glosario lisis) la mayoría de las moléculas en una célula. Los lisosomas son necesarios para la "digestión celular" [133].

A.1.4.8 Peroxisomas

Los peroxisomas son orgánulos esféricos u ovalados, de 0,3-1,5 μ, que están presentes tanto en células eucariotas de mamíferos como en protozoos y células vegetales. Los peroxisomas contienen peroxidasas que destruyen el peróxido de hidrógeno [133].

A.1.4.9 Centro celular o centrosoma

El centrosoma o centro organizador de microtúbulos (MTOC, microtubule organizing center) es un orgánulo de pequeñas dimensiones constituido por uno o dos centríolos sumergidos en una masa granulada; existe en todos las células animales susceptibles de dividirse, y generalmente ocupa una región muy cercana al centro de la célula [133].

A.1.4.10 Citoesqueleto

El citoesqueleto es una entidad constituida por el conjunto de MT (microtúbulos) , microfilamentos de actina (MF) y filamentos intermedios (FI). Interviene en el mantenimiento de la morfología celular, el transporte intracelular, la movilidad celular, la mitosis y la meiosis [133].

A.1.4.11 Citosol

El citosol es una solución acuosa (85 % de agua) de pH 7, homogénea, transparente, que no contiene estructuras visibles al MO (microscopio óptico) o al ME (microcospio electrónico). A veces se denomina hialoplasma (plasma transparente). El citosol es el sobrenadante obtenido después de varios fases de ultracentrifugación, que han eliminado el material particulado [133].

En el citosol se desarrollan los procesos siguientes [133]:

- Degradación (catabolismo) de moléculas proteicas, lipídicas y glucídicas.
- Síntesis (anabolismo) de moléssculas orgánicas (proteínas, glúcidos, lípidos, nucleótidos y algunos aminoácidos raros) destinadas a las membranas de los orgánulos.

A.1.5 La célula procariota

Etimológicamente, procariota (o protocariota) significa con "núcleo primitivo", pero carecen de núcleo celular. De hecho, el ADN de las células procariotas tiene forma de bucle cerrado y nunca está separado del citoplasma por una membrana. Los procariotas son seres unicelulares, o bien, aunque más raramente, seres pluricelulares (p.ej.; la

oscillaria) [133]. Como está descrito en [133], las células procariotas defieren de las eucariotas en :

- La presencia de una pared constituida por peptidoglicanos.
- Su tamaño (de 1 a 10 μm).
- Su molécula de ADN libre y circular, siempre en contacto con el citosol, que está desprovista de nucleosoma; sin embargo, este ADN está asociado a una histona denominada HU, que cumple varios funciones y, en particular, la de reparar el ADN.
- La ausencia de mitocondrias (la cadena respiratoria se localiza en la membrana plasmática de la bacteria) y de cualquier otro orgánulo limitado por membrana (Golgi, RE, lisosomas, peroxisomas, etc...)
- La ausencia de mitosis y de meiosis (los procariotas se reproducen por bipartición binaria).
- Los ribosomas que se parecen a los de mitocondrias y cloroplastos, y no a los del citoplasma de las células eucariotas, para cuya síntesis no es necesaria la presencia del nucléolo.

El diagrama A.2 representa la célula procariota.

Resumiendo, las células procariotas se caracterizan por la ausencia de núcleo, centrosoma, mitocondrias, aparato de Golgi, RE, lisosomas, peroxisomas, etc. Contienen una molécula de ADN circular (llamado plásmido), desprovista de nucleosomas), y ribosomas. Las células procariotas y eucariotas son estructuras complejas que aseguran la homeostasis del medio intracelular, permitiendo así que tendrán lugar, en las mejores condiciones, miles de reacciones. Bajo la influencia de factores extra e intracelulares, la célula no solamente se replica, sino que también es capaz de crecer. Las células eucariotas difieren de las procariotas en la presencia de un núcleo (envoltura nuclear que separa al ADN del citoplasma) y de orgánulos específicos (RE, Golgi, mitocondrias, cloroplastos, endosomas).

A.1.6 Arqueas

Las arqueas se distinguen por la organización de su pared, la existencia de ácidos grasos ramificados en la membrana plasmática y la presencia de nucleótidos particulares como la N1-metilonosina, encontrada específicamente en los ARNt. También difieren en la forma de los ribosomas y en la constitución de la ARN polimerasa. Dependiendo del autor, no se las considera un reino aparte, sino una bacteria, dividiéndose así en arqueobacterias (las que estamos llamando arqueas) y las eubacterias (las que nosotros consideramos bacterias). Por tanto, las archaea (o arqueas) son microorganismos unicelulares. Al igual que las bacterias, las archaea carecen de núcleo y es por ello

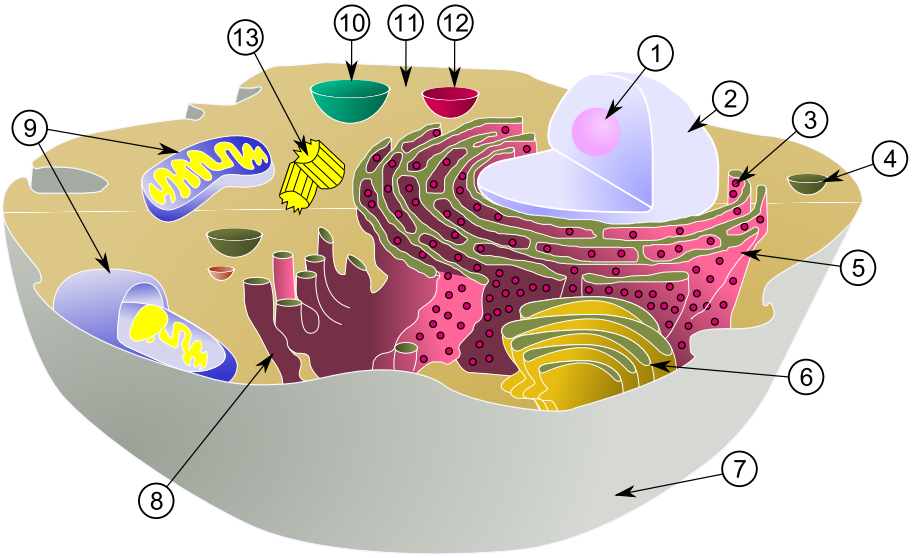
que son eucariotas. No obstante, las diferencias a nivel molecular entre arqueas y bacterias son enormes y por ello se las clasifica en grupos distintos. Actualmente se considera que las archaea están filogenéticamente más próximas a las eucariotas que a las bacterias [133].

Las archaea fueron descubiertas originariamente en ambientes extremos, pero desde entonces se las ha hallado en todo tipo de hábitats y podrían contribuir hasta el 20% de la biomasa [53]. Archaea constituye uno de los dominios en los que se dividen los seres vivos. Antiguamente se clasificaban como perteneciendo al reino Monera en la taxonomía tradicional de los cinco reinos. En 1990 se propuso considerarlos un dominio separado, según el sistema de tres dominios de Carl Woese [227, 228].

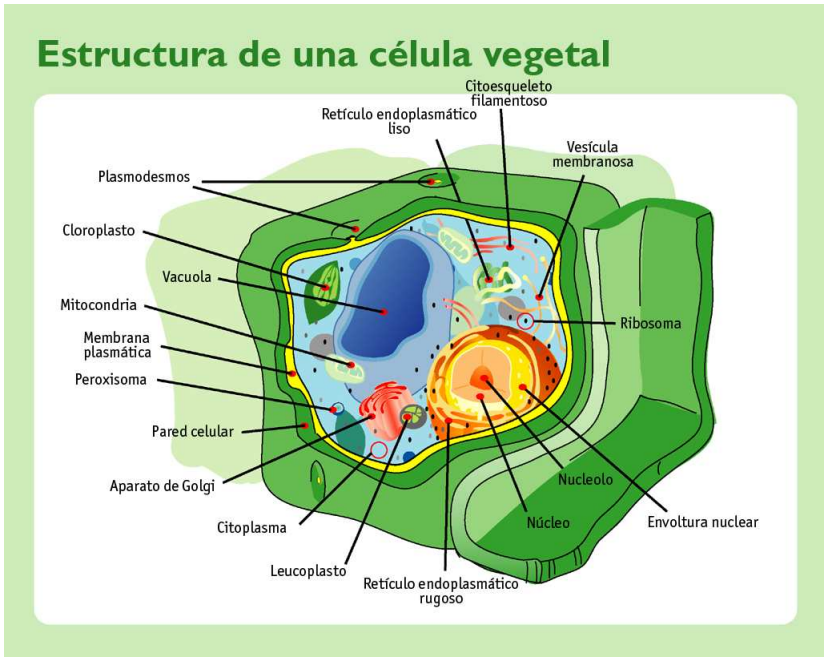
A.2 Referencias

Este apéndice ha sido realizado fundamentalmente usando varios libros, *Genética Moderna* [81], *Biología celular* [133] o *Citología e histología vegetal y animal* [161] y la documentación aportada en el *Máster de Bioinformática de la Universidad Internacional de Andalucía [50].*

Se hace una mención especial a www.wikipedia.es por las figuras de dominio público utilizadas y por aclaraciones de los conceptos básicos en la realización de estos apéndices escritos como ayuda de la lectura de esta tesis.



(a) Célula eucariota animal



(b) Célula eucariota vegetal

Fig. A.1: Diagrama de una célula animal, arriba (1. Nucleolo, 2. Núcleo, 3. Ribosoma, 4. Vesícula, 5. Retículo endoplasmático rugoso, 6. Aparato de Golgi, 7. Citoesqueleto (microtúbulos), 8. Retículo endoplasmático liso, 9. Mitocondria, 10. Vacuola, 11. Citoplasma, 12. Lisosoma. 13. Centríolos.); y de una célula vegetal, debajo. La imagen de célula animal se ha extraído de http://upload.wikimedia.org/wikipedia/commons/1/1a/Biological_cell.svg y la de vegetal de la siguiente dirección web http://upload.wikimedia.org/wikipedia/commons/7/73/Estructura_celula_vegetal.png bajo el auspicio de *Contenido Libre*.

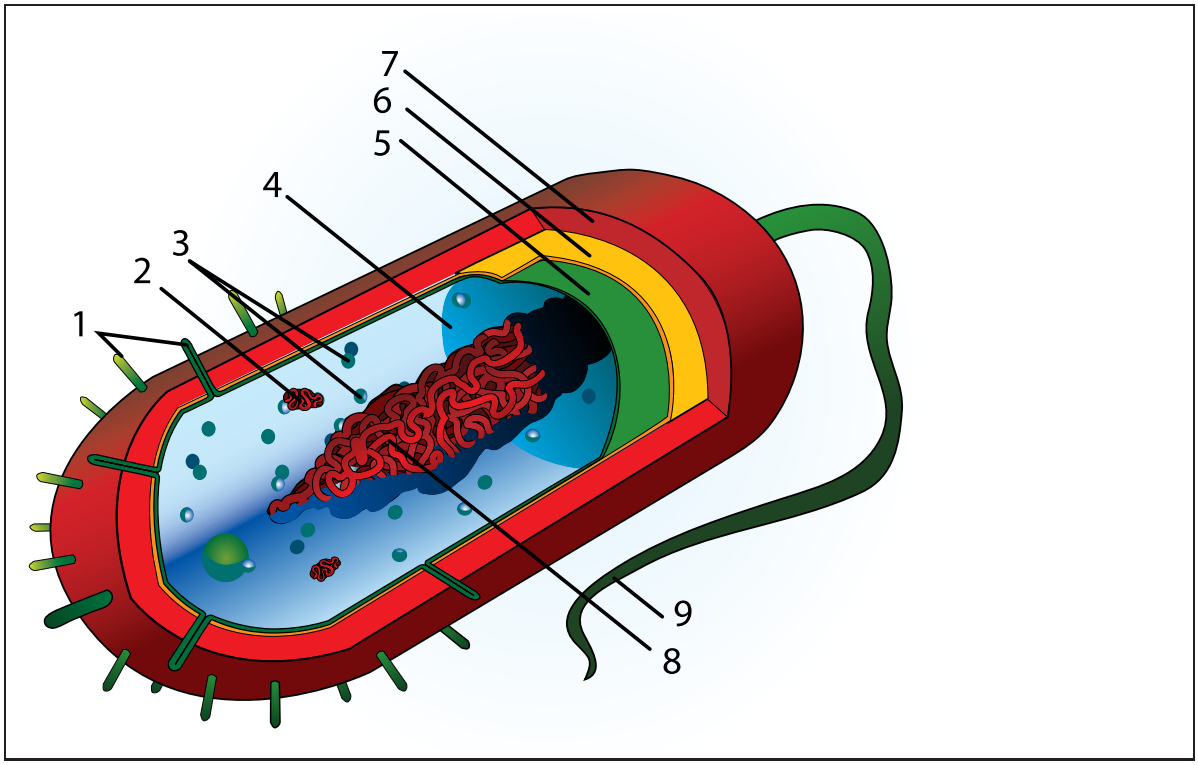


Fig. A.2: Estructura de la célula procariota: 1, pili; 2, plásmido; 3, ribosomas, 4, citoplasma; 5, membrana plasmática; 6, pared celular; 7, cápsula; 8, ADN; 9, flagelo bacteriano. Dicha figura ha sido tomado de http://upload.wikimedia.org/wikipedia/commons/7/72/Average_prokaryote_cell-_es.svg bajo *Contenido Libre*

APPENDIX B

BIOLOGÍA MOLECULAR Y BIOQUÍMICA

En este apéndice, se realizará una breve introducción a la biología molecular y bioquímica, explicando el material genético, el dogma central de la biología molecular, estructura del gen y los procesos de transcripción y traducción. Finalmente se realiza una explicación de los distintos ARN y qué es la homología.

B.1 Material Genético

Antes de adentrarnos en temas más profundos daremos una serie de nociones fundamentales, todo discurre centrado en el ADN (ácido desoxirribonucleico) que es el material genético. La información almacenada en el ADN permite la organización de moléculas inimitables en células vivas y funcionales, y en organismos que son capaces de regular su composición química interna, crecimiento y reproducción. Las unidades que gobiernan las características a nivel genético se llaman genes. Antes que nada, hay que entender la estructura química del ADN proporcionadas en los años 50, qué y como la información se pasa de una generación a la siguiente. El gen es considerado como la unidad de almacenamiento de información y unidad de herencia al transmitir esa información a la descendencia. Los genes se disponen, pues, a lo largo de cada uno de los cromosomas. Cada gen ocupa en el cromosoma una posición determinada llamada locus. El conjunto de cromosomas de una especie se denomina genoma. Los genes pueden aparecer en versiones diferentes, con variaciones pequeñas en su secuencia, denominadas alelos. Los alelos pueden ser dominantes o recesivos. Cuando una sola copia del alelo hace que se manifieste el rasgo fenotípico, el alelo es dominante. Cuando son precisas dos copias del alelo (una en cada cromosoma del par), el alelo es recesivo [76, 118].

Por tanto, un cromosoma es cada uno de los pequeños cuerpos en forma de bastoncillos en que se organiza la cromatina del núcleo celular en la mitosis y la meiosis,

cada uno de los cuales se divide longitudinalmente, dando origen a dos cadenas gemelas (iguales). Su número es constante para una especie determinada; en Homo sapiens sapiens (el ser humano) se tienen 46. De ellos 44 son autosómicos y 2 son sexuales o gonosomas [76].

Se llama cromatina al material microscópico constituido del ADN y de proteínas especiales llamadas histonas que se encuentra en el núcleo de las células eucariotas en las cuales los cromosomas se ven como una maraña de hilos delgados. Cuando la célula comienza su proceso de división (cariocinesis), la cromatina se condensa y los cromosomas se hacen visibles como entidades independientes. La unidad básica de la cromatina son los nucleosomas. Los cromosomas se suelen representar por pares, en paralelo con su homólogo [76].

El genotipo es el contenido genético (el genoma específico) de un individuo, en forma de ADN. Junto con la variación ambiental que influye sobre el individuo, codifica el fenotipo del individuo. De otro modo, el genotipo puede definirse como el conjunto de genes de un organismo y el fenotipo como el conjunto de rasgos de un organismo. se denomina fenotipo a la expresión del genotipo en un determinado ambiente. Los rasgos fenotípicos incluyen rasgos tanto físicos como conductuales. Es importante destacar que el fenotipo no puede definirse como la "manifestación visible" del genotipo, pues a veces las características que se estudian no son visibles de un individuo, como es el caso de la presencia de una enzima. Por tanto, el fenotipo está determinado fundamentalmente por el genotipo, o por la identidad de los alelos, los cuales, individualmente, cargan una o más posiciones en los cromosomas. Algunos fenotipos están determinados por los múltiples genes, y además influenciados por factores del medio. De esta manera, la identidad de uno, o de unos pocos alelos conocidos, no siempre permite una predicción del fenotipo [156, 76].

B.1.1 Nucleótidos

Los genes por sí mismos contienen su información como una secuencia específica de nucleótidos que son encontrados en las moléculas de ADN. Sólo 4 bases diferentes son usadas en las moléculas del ADN: guanina, adenina, timina y citosina (G, A, T y C). Cada base está unida a un grupo fosfato y un azúcar desoxiribosa para formar un nucleótido. Lo que hace que un nucleótido sea diferente de otro es la base nitrogenada que contiene. Toda la información que recibimos de los genes viene dada simplemente por el orden en los cuatro nucleóticos se encuentran a lo largo de las moléculas de ADN. Los complicados genes pueden estar formados por muchos miles de nucleóticos, y todas las instrucciones de genéticas de un organismo, su genoma, puede estar contenido en millones o incluso miles de millones de nucleótidos [118].

Las cadenas de nucleótidos pueden estar unidas unas a otras formando una cadena muy larga de polinucleótidos, y cuando se considera a una escala mayor, estamos hablando de cromosomas. Las uniones entre dos nucleótidos están siempre formadas por una enlace fosfodiéster que conecta el grupo fosfato de un nucleótido con un azúcar

desoribosa (desoxirribosa) de otro nucleótido. Todos los organismos vivos tienen este enlace fosfodiéster de la misma forma. Tanto en el ADN como en el ARN, el enlace fosfodiéster es el vínculo entre el átomo de carbono 3' y el carbono 5' del azúcar ribosa. La diferencia entre las terminaciones 5' y 3' de la cadena de polinucleótidos puede ser muy delicada, tanto que la orientación en las que las moléculas de ADN se disponga es importantísimo para la célula, para entender el contenido de su información [118].

Un tema común que se encuentran completamente en todos los sistemas biológicos y a todos los niveles es la idea de que la estructura y la función están íntimamente relacionados. Watson y Crick nos mostraron la estructura en doble hélice de las moléculas de ADN, y dieron la pista de cómo el ADN actúa como material genético. En 1953, se demostró que la información contenida en una hebra es esencialmente redundante a la información contenida en la otra. El ADN podía ser replicado y transmitido de generación en generación simplemente separando las dos hebras que conforman el ADN, y usando cada hebra como un patrón para la síntesis de una nueva hebra [118].

El contenido de una molécula de ADN viene dada por una secuencia específica de sus nucleótidos. Mientras que el contenido de información de cada hebra de la molécula doble-hélice de ADN es redundante no es exactamente la misma, es decir, es complementaria una hebra de otra. Por cada G localizada en una hebra, se encuentra una C en la otra y viceversa. La interacción entre G's y C's y entre A's y T's son específicas y estables. Las interacciones entre sus grupos químicos forma pares de bases estables [118].

Aunque las dos hebras de una molécula de ADN son complementarias no tienen la misma orientación 5'/3'. Por tanto, se dice que dos hebras son antiparalelas una respecto de la otra cuando la terminación 5' de una hebra corresponde a la terminación 3' de su complementaria y viceversa. Las secuencias que vienen dadas desde 5' a un punto en particular se les dicen que están descritas "upstream" mientras que aquellas que vienen dadas desde 3' se les dicen descritas "downstream" [118].

B.1.2 El Dogma central de la biología molecular

Mientras que la secuencia específica de nucleótidos en una molécula de ADN puede contener información de vital importancia para la célula, es de hecho las proteínas las que realizan el trabajo de alterar químicamente la célula actuando como catalizadores biológicos, se las llama enzimas. El término gen se suele usar de diferentes formas, pero una de sus definiciones más simples es que los genes explican con detalle las instrucciones necesarias para hacer que la enzima catalice. El dogma central de la biología molecular es el proceso por el que la información que se extrae de una secuencia de nucleótidos de un gen y se usa posteriormente para hacer que un gen sea esencialmente el mismo en todos los organismos de la Tierra. La información almacenada en el ADN se usa para hacer una cadena simple de polinucleótidos llamada ARN (ácido ribonucleíco) que es temporal y que se usa para crear proteínas. El proceso

de creación de una ARN copia de un gen se le llama transcripción y la actividad enzimática la lleva a cabo la ARN polimerasa. Hay una correspondencia uno a uno entre los nucleótidos que se usan para crear el ARN (G, A, U, y C donde U es la abreviación para Uracilo) y la secuencia de nucleótidos en el ADN (G, A, T y C respectivamente). El proceso de convertir la información de una secuencia de nucleótido en ARN a la secuencia de aminoácidos que crear una proteína se le llama traducción y es llevado a cabo por el complejo de proteínas y ARN llamado ribosomas [118].

B.1.3 Estructura del gen y del contenido de información

Gen se puede definir simplemente una región de ADN capaz de transcribirse en una molécula de ARN funcional (para la mayoría de los genes, un mARN). Podemos añadir que este ARN debe producirse en el momento correcto y en el lugar adecuado. El gen, por tanto, es realmente funcional. Para que todo esto ocurra, un gen contiene una región reguladora, es decir, un segmento de ADN con una secuencia específica de nucleótidos que le permita recibir y responder a señales de otras partes del genoma o del ambiente celular. Las señales de activación se convierten en proteínas reguladoras que se unen a la región reguladora del gen e inician la transcripción de la región adyacente que tiene la capacidad para formar ARN. En el otro extremo del gen existe una región encargada de terminar la transcripción [81].

Los genes eucariotas contienen segmentos de ADN llamados intrones que se encuentran intercalados en la región transcrita del gen. Los intrones no contienen información para la formación del producto génico correspondiente (por ejemplo, la proteína). Se transcriben junto con las regiones codificantes (llamadas exones) pero son luego eliminados del transcrito inicial. Los intrones por tanto deben ser considerados como parte de un gen. La figura B.1 muestra la estructura de un gen [81].

B.1.3.1 Expresión génica

La expresión génica es el proceso de usar la información almacenada en el ADN para producir una molécula de ARN y posteriormente su correspondiente proteína. En todos los organismos, hay dos pasos principales en la separación de la codificación de la proteína por un gen y su proteína: primero, el ADN en el que el gen reside debe ser transcrito del ADN a un ARN mensajero (ARMm), y segundo, debe ser traducido del ARNm a una proteína. El proceso de producir una molécula biológicamente funcional de ARN o una proteína se se le llama expresión génica, y la molécula resultado por sí misma se la llama producto de gen [81, 76, 75].

Por tanto, el código genético es el conjunto de reglas con las que un gen se traduce a una proteína funcional, Cada gen consiste en una secuencia específica de nucleótidos codificados en una hebra ADN (o a veces en ARN); una correspondencia entre los nucleótidos (los bloques básicos de construcción del material genético), y aminoácidos (los bloques básicos de las proteínas), deben estar establecidos por genes que se traducen a proteínas funcionales. Los conjuntos de tres nucleótidos, se conocen

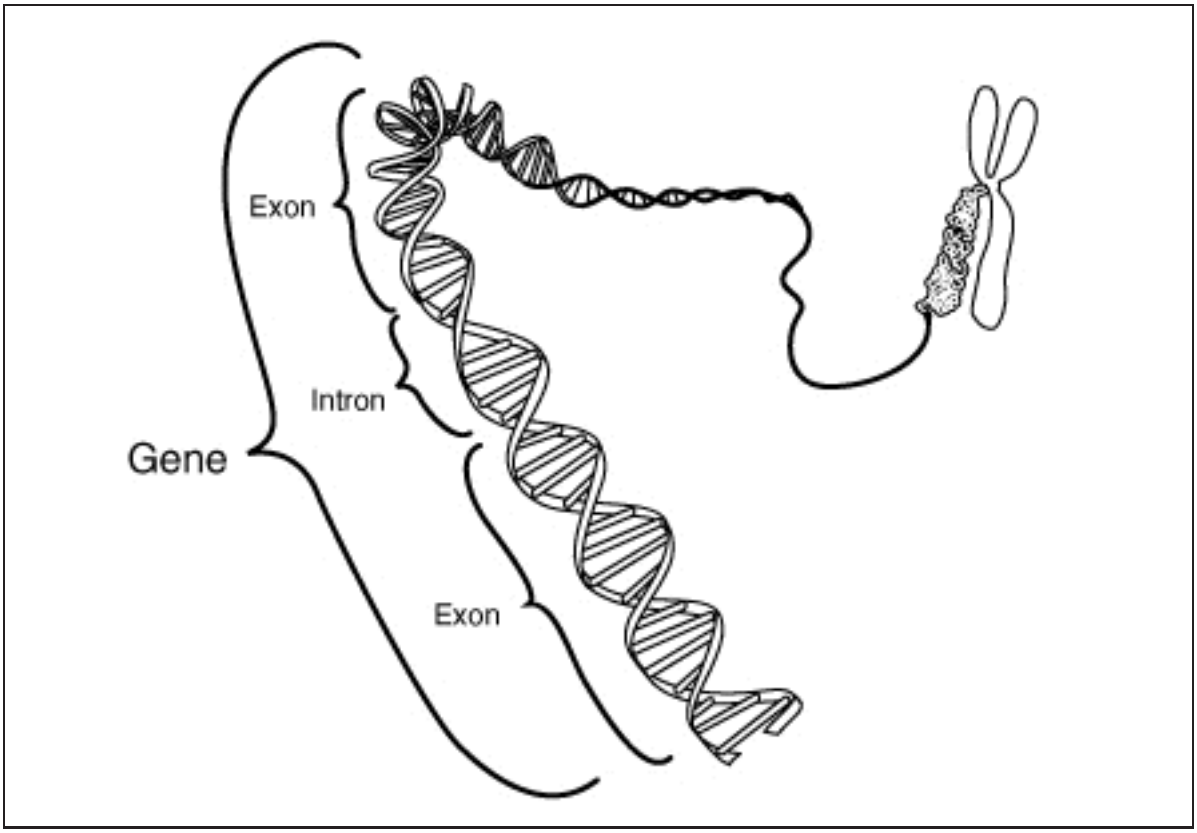


Fig. B.1: Genes en relación a la estructura de doble hélice del ADN y a un cromosoma (derecha). Los intrones son regiones que normalmente se encuentran en genes eucariotas y que son eliminados en el proceso de "splicing" o ayuste (después de que el ADN sea transcrito a ARN): solamente los exones codifican la proteína. El diagrama etiqueta una región de sólo 40 bases aprox. como un gen. En realidad la mayoría de los genes son cientos de veces más grandes. Fuente: figura obtenida de http://upload.wikimedia.org/wikipedia/commons/0/07/Gene.png cortesía del National Human Genome Research Institute (EEUU).

como codones, cada uno corresponde a un aminoácido específico o a una señal; hay tres codones se conocen como "codones de stop o parada" y, en lugar de especificar un nuevo aminoácido, alertan la maquinaria de traducción que el fin del gen se ha alcanzado. Existen 64 codones posibles (cuatro nucleótidos posibles con tres posiciones, por consiguiente 43 codones posibles) y solamente 20 aminoácidos estándar; así pues, el código es redundante y múltiples codones pueden especificar los mismos aminoácidos. La correspondencia entre codones y aminoácidos es casi universal entre todos los organismos vivos, el diagrama B.3 representa un codón. También es conveniente mencionar que en genética se llama marco abierto de lectura (siglas ORF del inglés Open reading frame) a cada una de las secuencias de ADN comprendida entre un

codón de inicio (ATG) de la traducción y un codón de terminación, descontando las secuencias que corresponden a los intrones en caso de haberlas. Se encuentra acotado por los UTRs, o secuencias no traducidas representado por la figura B.2 [76, 75].

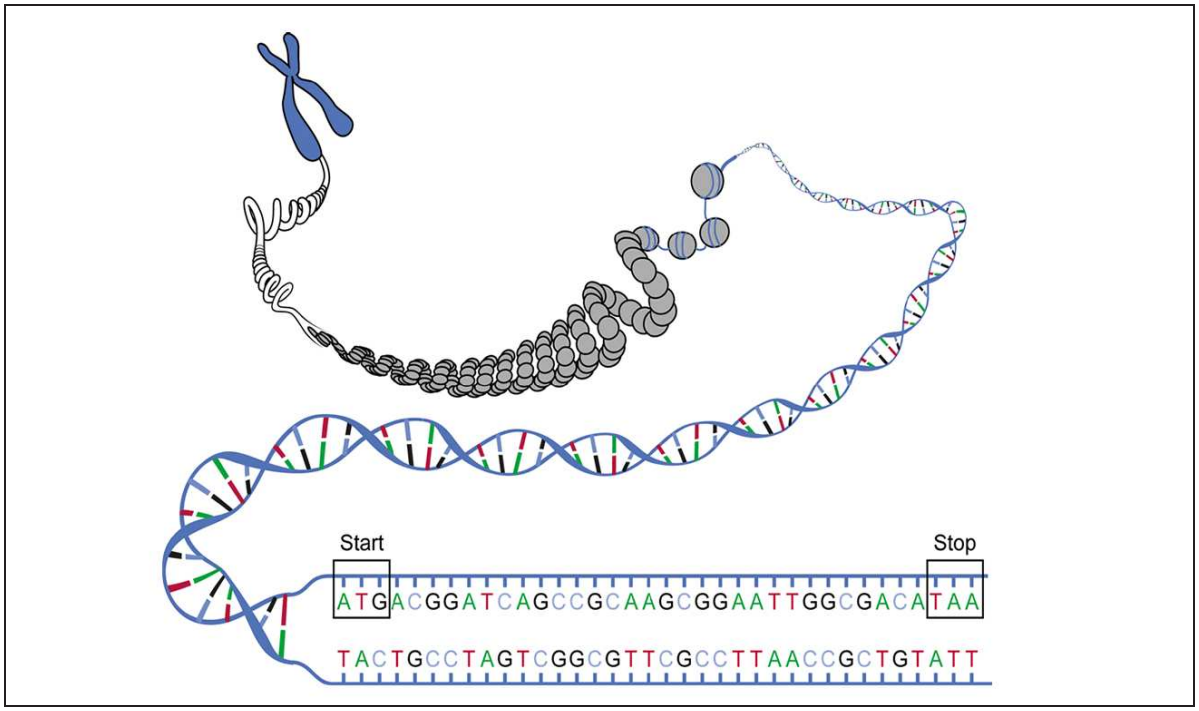


Fig. B.2: Esquema de un marco abierto de lectura (ORF), que incluye el codón de inicio (o start) y de parada (o stop). Figura extraída de http://upload.wikimedia.org/wikipedia/commons/b/bb/DNA_ORF.gif bajo dominio público por cortesía de National Human Genome Research Institute (EEUU).

B.1.3.2 Transcripción

El proceso de transcripción genética produce una molécula de ARN de una sola hebra conocida como ARN mensajero, cuya secuencia de nucleótidos es complementaria a la del ADN del cuál fue transcrito. La hebra de ADN cuya secuencia se ajusta con la del ARN se conoce como hebra codificante y la hebra a partir de la cual el ARN fue sintetizado es la hebra patrón. La transcripción se lleva a cabo por una enzima llamada ARN polimerasa, la cual lee la hebra patrón en la dirección 3' a 5' y sintetiza el ARN de 5' a 3'. Para iniciar la transcripción, la polimerasa primero reconoce y enlace una región promotor del gen (El promotor de un gen es la sección de ADN que controla la iniciación de la transcripción del ARN como producto de ese gen). Así pues, el mecanismo más importante de la regulación de genes es el bloqueo o aislando la región promotor, o uniendo fuertemente por moléculas represoras que físicamente bloquean la polimerasa, u organizando el ADN para que la región promotor no sea accesible

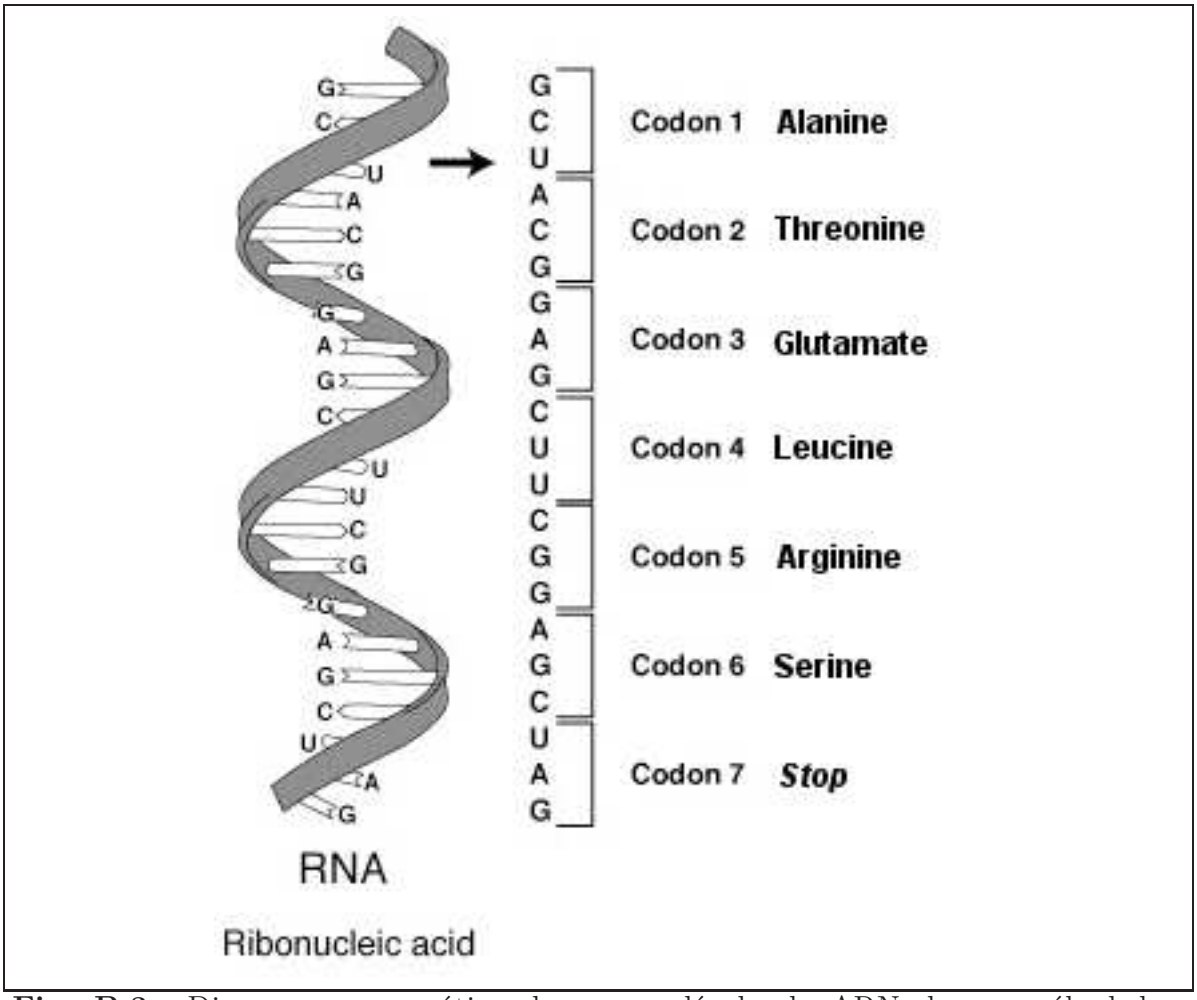


Fig. B.3: Diagrama esquemático de una molécula de ARN de una sóla hebra ilustrando la posición de los codones de 3 bases. Además podemos observar los aminoácidos correspondientes a los codones junto con el codón de stop o parada. Figura obtenida de http://upload.wikimedia.org/wikipedia/en/1/1d/Rna-codons-protein.png bajo dominio público por cortesía deL National Institutes of Health (EEUU).

[76, 75].

En procariotas, la transcripción tiene lugar en el citoplasma; La traducción puede empezar en la terminación 5' del ARN mientras que la terminación 3' está siendo transcrita. En eucariotas, la transcripcción tiene lugar necesariamente en el núcleo, donde el ADN de la célula está confinado; la molécula de ARN producida por la polimerasa se conoce como transcripción primaria (primary trasncript) y debe sufrir modificaciones post-transcripcionales antes de ser exportada al citoplasma para la traducción. El proceso de unión de los intrones presentes (splicing o ayuste) en la región transcrita es una modificación que sólo ocurre en eucariotas; los mecanismos

de splicing alternativos pueden resultar en transcripciones maduras a partir del mismo gen teniendo secuencias diferentes y por lo tanto codificación de proteínas diferentes. Esta es la principal forma de regulación en célula eucariotas [76, 75].

B.1.3.3 Traducción

La traducción es el proceso por el que una molécula ARNm maduro se usa como un patrón para sintetizar una nueva proteína. La traducción se lleva a cabo por los ribosomas, grandes complejos de ARN y proteínas responsables de las reacciones químicas que añaden nuevo aminoácidos haciendo crecer la cadena de polipéptidos mediante la formación de enlaces peptídicos. El código genético se lee cada vez de tres nucléotidos en tres nucléotidos, en unidades llamadas codones, mediante interacciones con moléculas de ARN especializadas llamadas ARN trasnferente (ARNt). Cada ARNt tiene 3 bases no emparejadas conocidas como el anticodon que son complementarias al codo que leen; el ARNt tiene también carga covalente del aminoácido para la nueva cadena de polipéptidos, la cual se sintetiza a partir de la terminación amino hasta la terminación carboxilo. Durante y después de su síntesis, la nueva proteína debe plegarse hasta lograr su estructura tridimensional activa antes de que pueda llevar a cabo su función celular [76, 75].

La figura B.4 representa los pasos principales que hemos mencionado, transcripción y traducción.

B.1.4 Genes y ARN

Como hemos discutido previamente, los productos iniciales de todos los genes son ácidos ribonucleicos (ARN). El ARN se sintetiza mediante un proceso que copia la secuencia nucleótidica del ADN (transcripción). Aunque tanto el ADN como el ARN son ácidos nucleicos, el ARN difiere en varios aspectos fundamentales [81]:

- El ARN está constituido por una cadena de nucleótidos, no es una hélice doble. Por tanto el ARN adopta más estructuras tridimiensionales complejas que el ADN de cadena doble.
- El ARN tiene un azúcar ribosa en sus nucleótidos en lugar de desoxirribosa.
- Los nucleótidos del ARN pueden tener las bases adenina, guanina y citosina, pero la base pirimidínica uracilo (U) sustituye a la timina.

B.1.5 Clases de ARN

Los ARN pueden agruparse en dos clases principales. Algunos ARN actúan de intermediarios en el proceso de descodificación de los genes a cadenas polipeptídicas (ver glosario, péptido). Nos referimos a estos ARN con el término de ARN “informativos”. En la otra clase, los propios ARN son los productos finales, funcionales. Por ello, nos referimos a estos ARN como ARN funcionales [81, 76, 75].

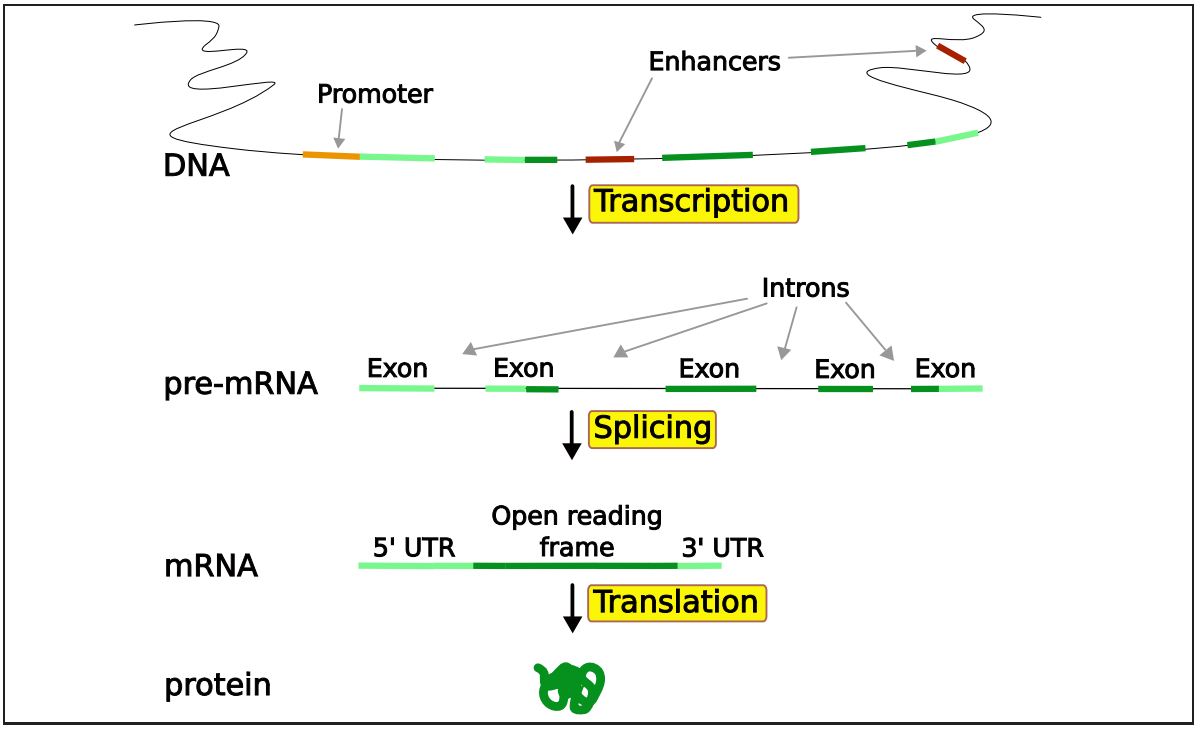


Fig. B.4: Diagrama del "típico" gen eucariota codificante de proteína. Promotores and "enhancers" determinan qué porciones de ADN será transcritos un ARNm precursor (pre-ARNm). El pre-ARNm es entonces ayustado (spliced) a un ARN mensajero (ARNm) que más tarde será traducido a una proteina. Figura obtenida de http://upload.wikimedia.org/wikipedia/commons/a/a7/Gene2-plain.svg bajo dominio público.

B.1.5.1 ARN informativos

Para la mayoría de los genes, el ARN es sólo un intermediario en la síntesis del producto funcional definitivo, que es una proteína. El ARN informativo de esta inmensa mayoría de genes es siempre ARN mensajero (ARNm) [81, 76, 75]. En procariotas, el transcrito (transcript), tal como es sintetizado a partir del ADN (transcrito primario), es el ARNm. En eucariotas, el transcrito primario sufre modificaciones en los extremos 5' y 3', y la eliminación de los intrones.

B.1.5.2 ARN funcionales

Los ARN funcionales ejecutan directamente sus propias funciones; nunca se traducen a polipéptidos. Las principales clases de ARN funcionales actúan en varios pasos del proceso de conversión de la información presente en el ADN. En todos los organismos se han encontrado dos clases de ARN funcionales de este tipo: ARN transferentes y ARN ribosómicos [81, 76, 75].

Las moléculas de ARN transferente (ARNt) funcionan como transportadores que

llevan los aminoácidos hasta el ARNm durante el proceso de traducción (síntesis proteica). Los ARN ribosómicos (ARNr) son componentes de los ribosomas, complejos macromoleculares que actúan de guías coordinando el ensamblaje de la cadena aminoacídica de una proteína. Los ribosomas están constituidos por varios tipos de ARNr y alrededor de 100 proteínas diferentes.

Existen otras clases de ARN funcionales implicados en el procesamiento de la información que son específicos de eucariotas. Los ARN nucleares pequeños (ARNsn) están implicados en el procesamiento de los pre-ARNm a ARNm en el núcleo de las células eucarióticas. Varios ARNsn diferentes, junto con varias subunidades proteicas, forma un complejo macromolecular designado snRNP (partícula ribonucleoproteica pequeña). Los ARN citoplasmáticos pequeños (ARNsc) están involucrados en el transporte de proteínas dentro de las células eucarióticas [81].

También cabe mencionar por su relación el Spliceosoma (también es posible llamarlo Complejo de corte y empalme) es un complejo formado por cinco ribonucleoproteínas nucleares pequeñas (snRNP, del inglés small nuclear ribonucleoproteins) capaz de eliminar los intrones (secuencias no codificantes) de los precursores del mRNA; este proceso se denomina splicing (ayuste de ARN. Las snRNP son complejos formados por unas diez proteínas más una pequeña molécula de ARN, rica en uracilo (U), que es la encargada de reconocer al intrón mediante apareamiento complementario de bases.

B.2 Genética y homología

A partir de los años 20 la homología comenzó a considerarse desde una perspectiva genética. En 1920 Alexander Weinstein acuñó el término genes homólogos para referirse a genes de especies distintas con expresiones fenotípicas similares [223]. En 1934 Alan Boyden reivindicó la genética como herramienta para el reconocimiento de homologías, a las que consideró, por primera vez, como un "fenómeno genético"[29]. Al principio, el concepto de homología, asociado a la genética mendeliana, individualizó los genes a partir de su función y no de su ascendencia común [32]. Dos alelos eran considerados homólogos sólo si compartían la misma expresión fenotípica. Este concepto de homología fue muy utilizado hasta los años sesenta (Mayr en 1963). A partir de entonces se impuso la concepción de homología genética manejada actualmente, según la cual, dos genes de dos especies distintas son homólogos si se derivan de un mismo gen ancestral [76].

La secuencia de nucleótidos de un gen es transmitida de padres a hijos y es lo que principalmente cambia con la evolución. Cuando examinamos el genoma de dos especies esperamos encontrar los genes equivalentes en ambas, con una secuencia algo diferente, más cuanto más remoto en el tiempo el antepasado común. La expresión homología de secuencia se refiere a la correspondencia entre las cadenas nucleotídicas (nucleótido) de esos dos genes, que es precisamente la que permite reconocer que son homólogos [76].

Dentro de la homología de secuencia se distinguen dos tipos de homología: **la ortología y la paralogía**. Llamamos genes ortólogos a los que son semejantes por pertenecer a dos especies que tienen un antepasado común. Existen además genes parálogos, que son aquellos que se encuentran en el mismo organismo, y cuya semejanza revela que uno procede de la duplicación del otro. La ortología requiere que se haya producido especiación, mientras que esta no es necesaria en el caso de la parología, que puede producirse sólo en los individuos de una misma especie [76].

Podemos también hablar de la xenología, cuando se produce una transferencia horizontal de gen (HGT - horizontal gene transfer) entre dos organismos. Los xenólogos pueden tener diferentes funciones, si el nuevo entorno es enormemente diferente que el gen transferido horizontalmente. Normalmente, los xenólogos tienen funciones similares en ambos organismos.

La **transferencia de genes horizontal** (TGH), también conocida como **transferencia de genes lateral** (TGL), es un proceso en el que un organismo transfiere material genético a otra célula que no es descendiente. Por el contrario, la transferencia vertical ocurre cuando un organismo recibe material genético de sus ancestros, por ejemplo de sus padres o de una especie de la que ha evolucionado. La mayoría de los estudios sobre genética se han centrado en la prevalencia de la transferencia vertical, pero hay un sentimiento actualmente de que la transferencia horizontal es un fenómeno significante. La transferencia artificial de genes horizontal es una forma de ingeniería genética [76].

Cabe reseñar que en procariotas, es muy común la **"transducción"**, el proceso por el que el ADN de una bacteria a otra a través de un virus bacteriano (un bacteriófago). Y para eucariotas, la transfección consiste en la introducción de material genético externo en células eucariotas mediante plásmidos, vectores víricos (en este caso también se habla de transducción) u otras herramientas para la transferencia. El término transfección para métodos no-virales es usado en referencia a células de mamífero, mientras que el término transformación se prefiere para describir las transferencias no-virales de material genético en bacterias y células eucariotas no animales como hongos, algas o plantas [76].

La duplicación génica es un fenómeno evolutivo importante. Una vez ocurrida, los genes repetidos evolucionan separadamente, pudiendo dar lugar a productos distintos y abriendo campo a nuevas adaptaciones. En biología molecular, la paralogía es el equivalente de la homología serial. Son parálogos, por ejemplo los genes que determinan las distintas clases de hemoglobinas que se producen a lo largo de la vida fetal y adulta. La hemoglobina consiste en un grupo hemo y cuatro globinas. En los vertebrados primitivos estas cuatro cadenas globinas eran del mismo tipo, pues se producían a partir de un mismo gen. Sin embargo, en los vertebrados superiores la hemoglobina consiste en dos cadenas de globina α y β, debido a la ocurrencia de una duplicación genética que condujo a dos copias del gen de globina original. Ambas copias divergieron a lo largo de la evolución, dando lugar a dos genes de globina especializados distintos

y a sus productos [76].

GenBank es una base de datos en la que se almacenan todas las secuencias de ADN. Para hacer test de homología se realiza una búsqueda llamada BLAST. Se introduce una secuencia y se obtiene una lista de todas las secuencias almacenadas que se parecen a la secuencia introducida, ordenada de mayor a menor grado de similitud [76].

B.3 Referencias

Este apéndice, además de las referencias indicadas de la bibliografía, ha sido completado fundamentalmente gracias a información de los artículos que se encuentran "online" en el portal especializado de *quimica.es* (o su versión en inglés http://www.chemeurope.com/en) de la empresa alemana CHEMIE.DE (CHEMIE.DE Information Service GmbH, All rights reserved) que gestiona portales científicos especializados, ofreciendo servicios y software para laboratorios. Exclusivamente online y desde hace más de diez años. Los artículos principales fueron:

- http://www.quimica.es/enciclopedia/Cromosoma.html
- http://www.quimica.es/enciclopedia/Gen.html
- http://www.quimica.es/enciclopedia/Genotipo.html
- http://www.quimica.es/enciclopedia/Fenotipo.html
- http://www.chemeurope.com/en/encyclopedia/Transcription_genetics.html
- http://www.chemeurope.com/en/encyclopedia/Translation_biology.html
- http://www.chemeurope.com/en/encyclopedia/Genetic_code.html
- http://www.quimica.es/enciclopedia/Ácido_ribonucleico.html
- http://www.quimica.es/enciclopedia/Homología_(biología).html (sección B.2).
- http://www.quimica.es/enciclopedia/Transferencia_horizontal_de_genes.html

Como se ha comprobado hay traducciones del libro de *Fundamental Concepts of Bioinformatics* [118]. Se ha hecho uso también de *Genética Moderna* [81]. Y para completar, se ha usando de referencia los libros *Biología celular* [133] o *Citología e histología vegetal y animal* [161], aunque estos últimos, han sido usados más extensamente en otros apéndices.

Se hace una mención especial a la documentación del Máster de Bioinformática de la Universidad Internacional de Andalucía [50] y a www.wikipedia.es por las figuras de dominio público utilizadas y por aclaraciones de los conceptos básicos en la realización de estos apéndices escritos como ayuda de la lectura de esta tesis. Sobretodo el artículo

más utilizado fue http://en.wikipedia.org/wiki/Gene extensamente revisado. Otros fueron:

- http://en.wikipedia.org/wiki/Homology_(biology)
- http://en.wikipedia.org/wiki/Horizontal_gene_transfer
- http://es.wikipedia.org/wiki/Marco_abierto_de_lectura

APPENDIX C

PROTEÍNAS

En este apéndice se realiza una breve introducción a la estructura de las proteínas y sus propiedades.

C.1 Estructura proteica

Una proteína es un polímero compuesto por monómeros denominados aminoácidos. Es una cadena de aminoácidos a la que en ocasiones nos referimos con el término polipéptido. Todos los aminoácidos se ajustan a la fórmula general C.1 [81]:

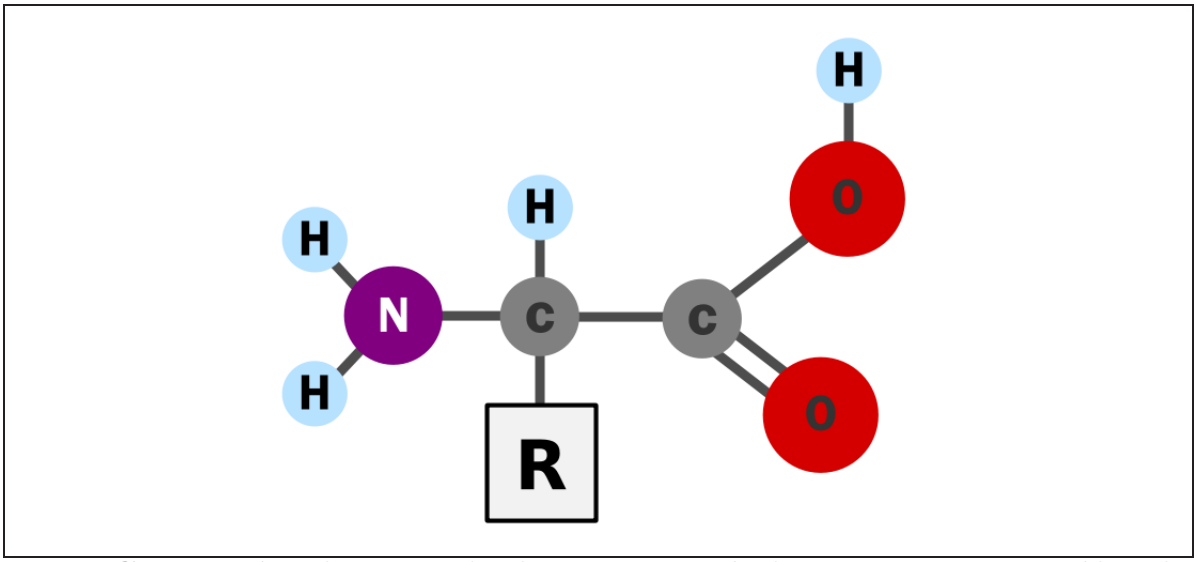


Fig. C.1: Fórmula general de un aminoácido. Figura extraída de http://upload.wikimedia.org/wikipedia/commons/c/ce/AminoAcidball.svg bajo domino público.

La cadena lateral, o grupo R (reactivo), puede ser cualquier grupo desde un átomo de hidrógeno hasta un anillo complejo. Hay 20 aminoácidos que pueden ser constituyentes de las proteínas, cada uno con un grupo R diferente que les confiere propiedades específicas. En las proteínas, los aminoácidos se mantienen unidos mediante enlaces covalentes denominados enlaces peptídicos [81].

El enlace peptídico se produce por una reacción de condensación durante la cual se elimina una molécula de agua. Debido al mecanismo de formación del enlace peptídico, una cadena polipeptídica siempre tiene un extremo amino (NH_2) y un extremo carboxilo (COOH) como se observa en C.1.

Las proteínas presentan una estructura compleja que consta de cuatro niveles de organización. La secuencia lineal de aminoácidos de la cadena polipeptídica constituye la **estructura primaria** de la proteína. Las proteínas adoptan una forma específica debido al establecimiento de varios tipos de enlaces débiles entre los átomos de una misma cadena polipeptídica, o de cadenas diferentes. La **estructura secundaria** de una proteína surge de las interacciones entre aminoácidos que se encuentran próximos en la secuencia lineal, dando lugar a una hélice α en la mayoría de los casos. Los motivos más comunes son la hélice alfa y la beta lámina en la estructura secundaria. La **estructura terciaria** se origina por el plegamiento de la hélice o de otras estructuras secundarias. En algunas proteínas, se reúnen dos o más estructuras terciarias para generar una **estructura cuaternaria**. El diagrama C.2 muestra los cuatro niveles de organización mencionados [81].

La asociación cuaternaria puede ocurrir entre polipéptidos diferentes o idénticos. Muchas proteínas adoptan estructuras compactas y se denominan proteínas globulares. Las enzimas y los anticuerpos se encuentran entre las proteínas globulares más importantes. Las proteínas con una forma lineal, denominadas proteínas fibrilares, son componentes importantes de estructuras tales como el pelo y el músculo [81].

La forma de una proteína es una propiedad esencial de una proteína debido a que es dicha forma específica la que le permite desempeñar una función específica en la célula. Por ejemplo, una enzima tiene un "bolsillo" específico (el sitio activo) donde encaja el sustrato. La forma está determinada principalmente por la secuencia aminoacídica primaria. En última instancia, incluso la estructura primaria de sus componentes aminoácidos. La secuencia de aminoácidos también dicta qué grupos R se encuentran disponibles para permitir que la proteína se una a otros componentes celulares específicos [81].

Más ampliamente, podemos decir que, la estructura de las proteínas,por tanto, jerarquizarse en una serie de niveles, interdependientes [76]. Estos niveles corresponden a:

1. Estructura primaria, que corresponde a la secuencia de aminoácidos.

2. Estructura secundaria, que provoca la aparición de motivos estructurales.

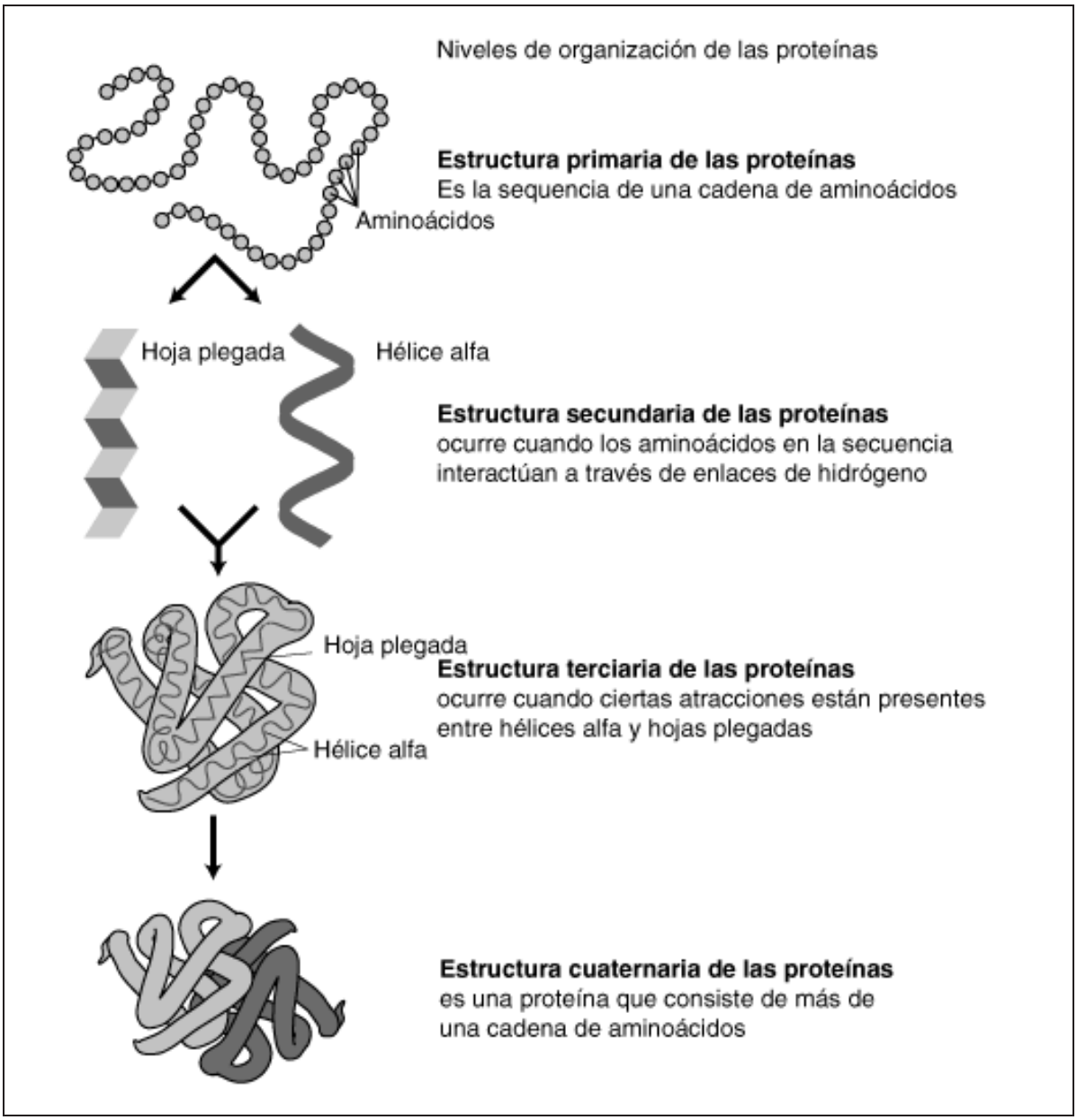


Fig. C.2: Niveles de organización de las proteínas. Figura extraída de http://upload.wikimedia.org/wikipedia/commons/2/25/Estructura_proteínas.png bajo dominio público cortesía del National Human Genome Research Institute.

3. Estructura terciaria, que define la estructura de las proteínas compuestas por un sólo polipéptido.

4. Estructura cuaternaria, si interviene más de un polipéptido.

C.1.1 Estructura primaria

La estructura primaria de las proteínas se refiere a la secuencia de aminoácidos, es decir, la combinación lineal de los aminoácidos mediante un tipo de enlace covalente, el enlace peptídico. Los aminoácidos están unidos por enlaces peptídicos siendo una de sus características más importantes la coplanaridad (en el mismo plano) de los radicales constituyentes del enlace [76].

La estructura lineal del péptido definirá en gran medida las propiedades de niveles de organización superiores de la proteína. Este orden es consecuencia de la información del material genético: Cuando se produce la traducción del RNA se obtiene el orden de aminoácidos que van a dar lugar a la proteína. Se puede decir, por tanto, que la estructura primaria de las proteínas no es más que el orden de aminoácidos que la conforma [76].

C.1.2 Estructura secundaria

La estructura secundaria de las proteínas es el plegamiento que la cadena polipeptídica adopta gracias a la formación de enlaces de hidrógeno entre los átomos que forman el enlace peptídico, es decir, un tipo de enlace no covalente [76].

Los motivos más comunes son la hélice alfa y la beta lámina.

Hélice alfa

Los aminoácidos en una hélice α están dispuestos en una estructura helicoidal dextrógira (que gira en el mismo sentido que las agujas del reloj en contraposición al sentido levógiro), con unos 3.6 aminoácidos por vuelta. Cada aminoácido supone un giro de unos 100° en la hélice, y los carbonos α de dos aminoácidos contiguos están separados por 1.5 $\mathring{A}$. La hélice está estrechamente empaquetada, de forma que no hay casi espacio libre dentro de la hélice. Todas las cadenas laterales de los aminoácidos están dispuestas hacia el exterior de la hélice [76].

El grupo amino del aminoácido (n) puede establecer un enlace de hidrógeno con el grupo carbonilo del aminoácido (n+4). De esta forma, cada aminoácido (n) de la hélice forma dos puentes de hidrógeno con su enlace peptídico y el enlace peptídico del aminoácido en (n+4) y en (n-4). En total son 7 enlaces de hidrógeno por vuelta. Esto estabiliza enormemente la hélice. Esta dentro de los niveles de organización de la proteína [76].

Lámina beta

La beta lámina se forma por el posicionamiento paralelo de dos cadenas de aminoácidos dentro de la misma proteína, en el que los grupos amino de una de las cadenas forman

enlaces de hidrógeno con los grupos carbonilo de la opuesta. Es una estructura muy estable que puede llegar a resultar de una ruptura de los enlaces de hidrógeno durante la formación de la hélice alfa. Las cadenas laterales de esta estructura están posicionados sobre y bajo el plano de las láminas. Dichos sustituyentes no deben ser muy grandes, ni crear un impedimento estérico, ya que se vería afectada la estructura de la lámina [76].

C.1.3 Estructura terciaria

Es el modo en que la cadena polipeptídica se pliega en el espacio, es decir, cómo se enrolla una determinada proteína, ya sea globular o fibrosa. Es la disposición de los dominios en el espacio [76].

La estructura terciaria se realiza de manera que los aminoácidos apolares se sitúan hacia el interior y los polares hacia el exterior en medios acuosos. Esto provoca una estabilización por interacciones hidrofóbicas, de fuerzas de van der Waals y de puentes disulfuro [122] (covalentes, entre aminoácidos de cisteína convenientemente orientados) y mediante enlaces iónicos [76].

C.1.4 Estructura cuaternaria

La estructura cuaternaria deriva de la conjunción de varias cadenas peptídicas que, asociadas, conforman un ente, un multímero, que posee propiedades distintas a las de sus monómeros componentes. Dichas subunidades se asocian entre sí mediante interacciones no covalentes, como pueden ser puentes de hidrógeno, interacciones hidrofóbicas o puentes salinos. Para el caso de una proteína constituida por dos monómeros, un dímero, éste puede ser un homodímero, si los monómeros constituyentes son iguales, o un heterodímero, si los monómeros no son iguales [76].

C.1.5 Propiedades de las proteínas

Las propiedades son las siguientes [76]:

- Solubilidad: Se mantiene siempre y cuando los enlaces fuertes y débiles estén presentes. Si se aumenta la temperatura y el pH, se pierde la solubilidad.

- Capacidad Electrolítica: Se determina a través de la electrólisis (descomposición mediante una corriente eléctrica) , en la cual si las proteínas se trasladan al polo positivo es porque su radical tiene carga negativa y viceversa.

- Especificidad: Cada proteína tiene una función específica que está determinada por su estructura primaria.

- Amortiguador de pH: (conocido como efecto tampón)Actúan como amortiguadores de pH debido a su carácter, anfotero, es decir, pueden comportarse como ácidos (soltando electrones(e-)) o como bases (tomando electrones).

C.1.6 Desnaturalización

Las proteínas pueden desnaturalizarse al perder todas sus estructuras menos la primaria. Al desnaturalizarse una proteína, esta pierde solubilidad en el agua y precipita. La desnaturalización se produce por cambios de temperatura o variaciones de pH, sales de metales pesados, radiación UV, rayos X. En algunos casos, las proteínas desnaturalizadas pueden volver a su estado original a través de un proceso llamado renaturalización [76].

C.2 Referencias

Fundamentalmente para la realización de este apéndice se ha utilizando el libro de *Genética Moderna* [81], y gracias a información de los artículos que se encuentran "online" en el portal especializado de *quimica.es* (o su versión en inglés http://www.chemeurope.com/en) de la empresa alemana CHEMIE.DE (CHEMIE.DE Information Service GmbH, All rights reserved) que gestiona portales científicos especializados, ofreciendo servicios y software para laboratorios. o de *quimica.es*. Los artículos principales fueron:

- http://www.quimica.es/enciclopedia/Proteínas.html
- http://www.quimica.es/enciclopedia/Estructura_de_las_proteínas.html

Se quiere hacer especial mención a www.wikipedia.es de donde se extrajeron las figuras de éste apéndice bajo dominio público y que han permitido aclarar los conceptos expuestos (ver artículos de proteína y de estructura de las proteínas).

APPENDIX D

PROTEÓMICA

D.1 Proteómica y genómica

El objetivo de la genómica es obtener el catálogo de genes de los organismos vivos. Dicho catálogo contendría información sobre la ubicación del gen, sus elementos promotores, reguladores, composición del gen en exones e intrones, su secuencia de bases, función, proteína que sintetiza, etc. Es pues una lista detallada, finita y digamos que estática. La proteómica complementa la visión genómica con el entendimiento de la dinámica del genoma: el proteoma, o el conjunto de las proteínas expresadas en un organismo en un determinado momento bajo unas determinadas condiciones. Es decir, más que el conocimiento de cuáles son los genes en un organismo, el interés se centra en comprender como esos genes se expresan bajo determinadas condiciones ambientales, y especialmente, como cambia ese comportamiento cuando las condiciones se modifican [99].

La regulación de los procesos celulares se basa en la modificación de las proporciones relativas de las proteínas presentes en la célula. Por ejemplo, cuando una célula recibe una señal, digamos un factor de crecimiento inmediatamente modifica sus niveles de proteínas. Los receptores de la superficie celular se activan y modifican, se envía el mensaje desde el receptor al núcleo (lo que involucra también movimiento de proteínas). El resultado final es que la célula incremente o disminuya la presencia de determinadas proteínas (en el ejemplo, los genes requeridos para la división celular aumentarán su actividad) [99].

El aspecto clave en proteómica (caracterización y estudio del proteoma) es la identificación de esas variaciones. Interesa también conocer, por ejemplo, qué genes se expresan más o menos (diferencialmente) en determinados tejidos, cuáles son las diferencias en la expresión de determinadas proteínas cuando se trata de un tejido sano o de un tejido enfermo. Este conocimiento es de vital importancia como mecanismo de

cura de enfermedades o para poder diagnosticarlas[99].

D.1.1 Estudios Analíticos

Las áreas de aplicación de los estudios proteómicos son [99]:

- Identificación a gran escala de proteínas y micro-caracterización de sus modificaciones post-traducionales.
- Proteómica de "expresión diferencial" para la comprobación de niveles de proteínas con aplicación potencial en un amplio abanico de enfermedades.
- Estudio de las interacciones proteína-proteína mediante técnicas de afinidad, espectrometría de masas o el sistema doble híbrido de levaduras.

Cualquier estudio que se realice en torno a estas áreas de aplicación de la proteómica, la finalidad del mismo es la identificación y caracterización de una serie de proteínas que intervienen en dicho estudio. Para ello, se establece un modelo más o menos homogéneo de análisis, que comprende las fases de separación de proteínas, digestión proteolítica, identificación y caracterización [99].

D.1.1.1 Separación de proteínas

Normalmente la muestra a analizar se compone de una mezcla compleja de varias proteínas, de la cual conocemos bastante poco. Por tanto, una vez que esta muestra se ha tratado con una serie de compuestos (preparación de la muestra), se dispone a someterla a una técnica de separación de proteínas para, en la medida de lo posible, realizar análisis separados por cada proteína presente [99].

Las técnicas más utilizadas para la separación de proteínas son:

1. Cromatografía líquida de alta resolución (HPLC) (High Perfomance Liquid Chromatography) Donde la separación de las proteínas se realiza en función de su tamaño. El complejo se deposita en una columna que contiene un polímero entrecruzado con poros de tamaño determinado. Las proteínas mayores se desplazan más rápidamente que las de menor tamaño ya que son demasiado grandes para penetrar en los poros de la cuentas y siguen una ruta más directa a través de la columna. las más pequeñas penetran en los poros y pasan más despacio [99].

2. Electroforesis bidimensional Dentro de esta técnica la separación se realiza en función del punto isoeléctrico (pH en donde la carga de la proteína es neutra. Primera dimensión) y de la masa molecular (segunda dimensión) de cada proteína. El gel bidimensonal lo podemos interpretar como una matriz donde cada mancha o spot que observamos representa una proteína, como se puede observar en la figura D.1 [99].

Por tanto, la electroforesis es una técnica para la separación de moléculas (proteínas o ácidos nucleicos) según la movilidad de estas en un campo eléctrico a través de una matriz porosa, la cual finalmente las separa por tamaños moleculares y carga eléctrica, dependiendo de la técnica que se use. Los ácidos nucleicos ya disponen de una carga eléctrica negativa, que los dirigirá al polo positivo, mientras que las proteínas se cargan con sustancias como el SDS (detergente) que incorpora cargas negativas de una manera dependiente del peso molecular. Para la separación se usa un gel de agarosa o poliacrilamida (fibras cruzadas, como una malla). Al poner la mezcla de moléculas y aplicar un campo eléctrico, éstas se moverán y deberán ir pasando por la malla, por la que las pequeñas se moverán mejor, más rápidamente. Así, las más pequeñas avanzarán más y las más grandes quedarán cerca del lugar de partida [99].

La electroforesis en gel es un grupo de técnicas empleadas por los científicos para separar moléculas basándose en propiedades como el tamaño, la forma o el punto isoeléctrico. La electroforesis en gel se utiliza generalmente con propósitos analíticos, pero puede ser una técnica preparativa para purificar moléculas parcialmente antes de aplicar una espectroscopia de masas, una PCR (reacción en cadena de la polimerasa, típica para amplificar un fragmento de ADN), una clonación o una secuenciación de ADN [99].

Uno de los problemas típicos es la falta de fijación en ambas dimensiones, apareciendo desplazamientos, este fenómeno se le llama streak o streaking [99].

D.1.2 Digestión proteolítica

Esta etapa es necesaria para el análisis de nuestras proteínas por espectrometría de masas. La misma consiste en fraccionar cada una de las mismas en los distintos péptidos que la componen. La cadena polipeptídica no se fracciona al azar, sino que por el contrario, la podemos fraccionar en un número restringido de puntos y producir un número limitado de fragmentos bien definidos gracias al empleo de unas enzimas, las proteasas, que son las encargadas de romper los enlaces existentes entre los distintos péptidos [99].

Existen dos grandes categorías, según la forman en que "atacan" (hidrólisis de los enlaces peptídicos) la proteína: endoproteasas y exoproteasas. Las endoproteasas o endopeptidasas atacan dentro de la proteína, dando lugar a péptidos relativamente largos, mientras que las peptidasas o exopeptidasas atacan los extremos o fragmentos de una proteína dando como resultado péptidos pequeños o incluso aminoácidos [99].

D.1.3 Identificación de proteínas

Después de la separación y digestión de las distintas proteícas presentes en la muestra, debemos realizar los análisis necesarios para la identificación de cada una de las mismas. La identificación de las proteínas se pueden contemplar de dos formas. Por una parte

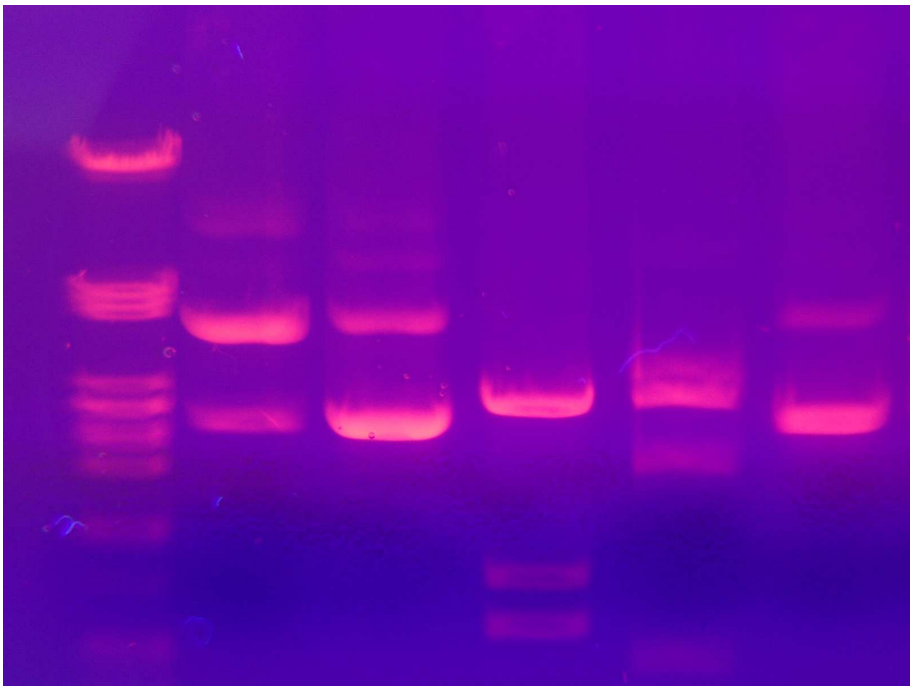


Fig. D.1: Representación matricial de las proteínas por su punto isoeléctrico y masa molecular. Es un gel bidimensional de poliacrilamida. Figura extraída de http://upload.wikimedia.org/wikipedia/commons/6/60/Gel_electrophoresis_2.jpg bajo dominio público.

cuando las proteínas a identificar están registradas y anotadas en bases de datos, de tal forma que con algoritmos complejos de búsqueda se pueden localizar; y por otra, cuando las proteínas que pretendemos determinar no se encuentran introducidas en ninguna base de datos (por ejemplos usar espectrometría de masas) [99].

D.1.4 Caracterización de proteínas

Esta fase parte casi siempre de la identificación previa de la proteína, para determinar ciertas modificaciones post-traduccionales que se observan en un conjunto determinado de proteínas de nuestra, y también para determinar la información relevante de las mismas, o sus posibles interacciones con otras macromoléculas [99].

D.2 Espectrometría de masas

La Espectrometría de masas es una técnica analítica que ha experimentado un gran desarrollo tecnológico en los últimos años. Permite estudiar compuestos de naturaleza diversa: orgánica, inorgánica o biológica y obtener información cualitativa o cuantitativa. Mediante el análisis por Espectrometría de masas es posible obtener información de la masa molecular del compuesto analizado así como obtener

información estructural del mismo. Para ello es necesario ionizar las moléculas y obtener los iones formados en fase gaseosa. Este proceso tiene lugar en la fuente de ionización y actualmente, existen diferentes técnicas que permiten llevarlo a cabo como Impacto Electrónico (EI), Bombardeo con átomos rápidos (FAB), Ionización Química a Presión Atmosférica (APCI), Desorción/Ionización por Láser Asistida por Matriz (MALDI) ó Electrospray (ESI). Los iones generados son acelerados hacia un analizador y separados en función de su relación masa/carga (m/z) mediante la aplicación de campos eléctricos, magnéticos ó simplemente determinando el tiempo de llegada a un detector. Los iones que llegan al detector producen una señal eléctrica que es procesada, ampliada y enviada a un ordenador. El registro obtenido se denomina Espectro de masas y representa las abundancias iónicas obtenidas en función de la relación masa/carga de los iones detectados. Esta definición y más información se encuentra disponible en http://www.uam.es/investigacion/servicios/sidi/especifica/masas.html de la documentación del Laboratorio de Espectrometría de Masas de la Universidad Autónoma de Madrid.

La espectrometría de masas es una tecnología analítica esencial en el contexto de la proteómica actual. Básicamente los espectrómetros de masas utilizados en el análisis de proteínas o péptidos pueden ser divididos en dos partes: la primera constituye una fuente de iones que genera e introduce los iones analitos en el instrumento y la segunda un detector para medir las masas de los iones generados. Existen dos tipos de técnicas de ionización que se utilizan habitualmente: la ionización por electronebulización (Electrospray Ionisation, ESI), y la deserción/ionización mediante láser inducida por matriz (Matrix-Assited Laser Desoption/Ionization, MALDI) [99].

En la figura D.2 se muestra un ejemplo de los resultados de una espectrometría de masas de un espectrómetro ESI-Q-TOF .

D.2.1 Espectrometría de masas MALDI-TOF

MALDI-TOF es una técnica de ionización suave utilizada en espectrometría de masas. Se denomina MALDI por sus siglas en inglés Matrix-Assisted Laser Desorption/Ionization (desorción/ionización láser asistida por matriz) y TOF por el detector de iones que se acopla al MALDI y cuyo nombre procede también de sus siglas en inglés Time-Of-Flight [76].

Permite el análisis de biomoléculas (biopolímeros como las proteínas, los péptidos y los azúcares) y moléculas orgánicas grandes (como los polímeros, los dendrímeros y otras macromoléculas) que tienden a hacerse frágiles y fragmentarse cuando son ionizadas por métodos más convencionales [76].

Los espectrómetros de masas MALDI generan iones en fase gaseosa mediante la vaporización inducida por láser de una mezcla sólida de analito y la matriz. La matriz (generalmente un compuesto orgánico pequeño y cristalino) actúa como aceptor de la energía transmitida por el láser y la transmite a las moléculas de analito, produciendo la sublimación de los iones de la matriz y del analito a la fase gaseosa. Estos iones

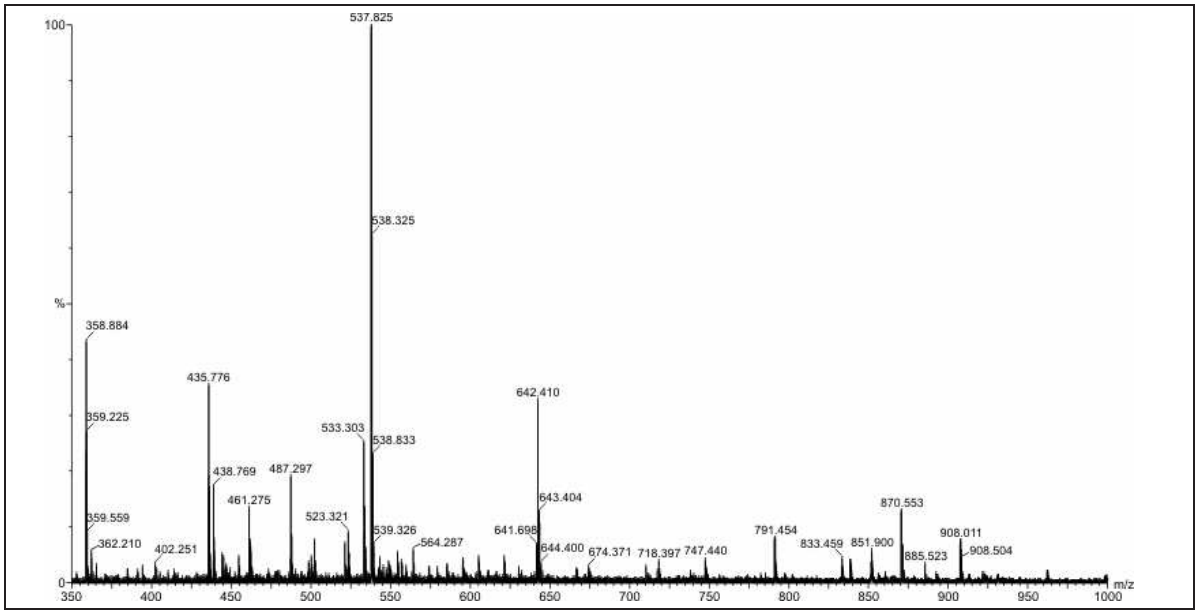


Fig. D.2: Ejemplo del resultado de un espectrómetro de masas, representa masa (m) respecto la carga (z). Figura extraída de http://upload.wikimedia.org/wikipedia/commons/7/73/WidmoMS.gif. Esta figura se encuentra bajo la licencia Creative Commons Genérica de Atribución/Compartir-Igual 3.0

son acelerados en un campo eléctrico para que, a continuación, penetren en un tubo de recorrido o vuelo libre sin campo eléctrico alguno. Para un voltaje de aceleración dado, el tiempo de vuelo (Time-Of-Flight, TOF) que lleva a un ion alcanzar el detector es proporcinal a su relación masa/carga (m/z) de tal forma que los iones más pesados son más lentos que los más ligeros, y por consiguiente su registro en el detector es posterior [76].

La proteína analizada se digiere completamente con una endoproteasa y los péptidos resultantes de la digestión se extraen y se preparan para el análisis por espectrometría. El espectro de masas de la digestión de los péptidos es una relación de masas moleculares de cada uno de los péptidos producidos, que conforma un patrón característico para dicha cadena polipeptídica, lo que llamamos huella peptídica (fingerprinting) [76, 99].

D.2.2 Espectrometría de masas tándem (MS/MS)

El espectrómetro de masas MALDI-TOF es una herramienta muy potente a la hora de identificar un alto número de proteínas pero con pocas o ninguna modificación. De hecho, esta carencia, es debido a que MALDI, no identifica proteínas en base a la secuencia de aminoácidos que la compone, y por tanto cualquier variación de la misma complica el proceso de identificación [99].

Esto, junto con el problema que puede originarse por el modo de ionización, se vuelve necesario cumplimentar el uso del MALDI-TOF, con otras técnicas como la degradación de Edman, en desuso, o de otros espectrómetros de masas que permitan trabajar a nivel de secuencia de proteínas [99].

En esta dirección, y como posible técnica complementaria a MALDI, utilizamos la espectrometría de masas en tandem (en dos fases masa/masa -MS/MS-), como el sistema compuesto por electronebulización con trampa iónica (ESI), en donde, el analito es introducido en solución y los iones se forman por desolvatación (liberar las moléculas o iones de un soluto, disociándose de las del disolvente) de las microgotas de líquido en un campo eléctrico intenso a presión atmosférica [99].

Concluida esta fase, los iones son dirigidos al analizador, que en nuestro caso es una trampa iónica (Ion trap), que gracias a un campo magnético es capaz de retener todos los iones obtenidos en la fase anterior, y permitir sólo el paso de aquellos que formar parte de un determinado péptido [99].

D.3 Métodos para la detección y análisis de interacciones proteína-proteína

Esta sección ha sido extraída principalmente del artículo de Tord Berggård et al. [18], realizando una traducción al español de buena fé de todos los métodos que se exponen a continuación.

Antes de centrarnos en el estudio de la bioinformática aplicada la detección de interacciones proteína-proteína, conviene realizar una explicación sobre los diferentes métodos en laboratorio para la detección y análisis de interacciones proteína-proteína, que habitualmente aparecen en la literatura relacionada con este campo, y al menos, es necesario tener una cierta idea de lo que realmente se está hablando.

El estudio de una típica interacción proteína-proteína comienza con una prueba de detección (screen o muestreo) inicial. Existen tres técnicas que pueden ser usadas para esta finalidad [18]:

- proteínas “affinity-tagged” (afinidad-etiquetadas)
- sistemas de doble híbrido (“two-hybrid”)
- Algunas técnicas proteómicas cuantitativas que pueden ser usadas en combinación con , por ejemplo, cromatografía de afinidad y coinmunoprecipitación para las pruebas de detección (screening o muestreo) de las interacciones proteína-proteína.

Existen cuatro estrategias más usadas para la validación de las interacciones proteína-proteína:

- Microcopio confocal para la colocalización intracelular de proteínas.
- Coinmunoprecipitación.

- Surface Plasmon Resonance (SPR - Resonancia de Plasmón de Superficie).

- Estudios de espectroscopia.

El genoma humano consiste de 20000-30000 genes que codifican alrededor de 500000 proteínas diferentes de las cuales más de 10000 pueden ser producidas por la célula en cualquier momento dado (el "proteoma" celular). Se ha estimado que el 80% de las proteínas no operan solas sino en complejos. Estas interacciones proteína-proteína se regulan por medio de varios mecanismos. Muchas interacciones proteína-proteína son parte de redes celulares más grandes de interacciones de proteína-proteína. Se piensa que las redes celulares de interacciones proteína-proteína son construidas por nodos altamente conectados (llamados hubs -centros-) y muchos nodos débilmente conectados [18]. Cada nodo recibe entradas y genera una o más salidas específicas de modo similar a unidades computacionales (p.ej.: algunos complejos importantes son los ribosomas, spliceosome (o ayustosoma), y el complejo de poro nuclear). La arquitectura básica de la red de interacción proteína-proteína es similar en todas las células. Así, los hubs son esenciales para la supervivencia de la célula y son iguales en todasls celulas, y por tanto, las diferencias de células específicas pueden ser encontradas a nivel regulatorio. La mayoría de proteínas celulares están directamente o indirectamente asociadas a grandes redes de interacción proteína-proteína celulares, y por tanto, existe implicaciones en el modo en se definen los caminos o rutas celulares (cellular pathways). Por tanto, una ruta celular (p.ej., tráfico en vesículas, apoptosis - muerte celular- o control del ciclo de la célula) puede ser visto como un subsistema que está altamente interconectado a otras rutas [18].

En los últimos años se ha innovado tremendamente en métodos de identificación y caracterización de interacciones proteína-proteína que han sido presentadas y muchas de ellas están actualmente en uso en laboratorios de todo el mundo. Así por ejemplo, actualmente las técnicas de espectrometría de masas pueden ser usadas no sólo para identificar proteínas individuales, sino también para caracterizar una serie de datos biológicos. Algo sorprendente, es que un número muy bajo de interacciones están avaladas por más de un método [220]. Debido a la cantidad de métodos diferentes hay poco solapamiento entre los subconjuntos de interacciones obtenidos, sin embargo, en subconjuntos de interacciones proteína-proteína identificados por el mismo método, el solapamiento es muy bajo (p.ej.: comparar resultados de métodos doble-híbrido de levadura (Y2H - yeast two-hybrid) from Ito et al. [101] y Uetz et al. [212]). Aunque el número de métodos que se han desarrollado últimamente combaten la tasa de falsos positivos, es claramente importante validad las interacciones proteína-proteína por varios métodos.

D.3.1 Uso de etiquetas (tags) de afinidad para purificación de complejos de proteínas in vivo

Esta sección está extraída del artículo de Tord Berggård et al. [18] cuyo título en inglés de la sección sería *Use of affinity tags for purification of protein complexes in vivo*.

Primero, las células son transfectadas (ver transfección en el glosario) mediante un plásmido que codifica una proteína cepo fusionada (Proteínas de fusión) a una etiqueta (de afinidad). Después de un periodo apropiado de expresión, las células son descompuestas por lisis (rotura de la membrana celular) y el cepo etiquetado, junto con proteínas a las que está unida, se aislan usando una producto químico específico o mediante algún ligando biológico (iones o moléculas que rodean a un metal en un complejo) sujetas a un soporte sólido. Las proteínas extraídas por un disolvente (eluted) son entonces separadas por gel-electrofóresis y especialmente las proteínas unidas se identifican por espectrometría de masas. Es un método excelente para purificar e identificar complejos de multiproteina. Muchos métodos muy usados, como el sistema doble-híbrido de la levadura, no son capaces de detectar más proteínas involucradas en una interacción (proteína-proteína). Los métodos bajados en afinidad estás sesgados por las proteínas que interaccionan con gran afinidad y con baja cinética de disociación. Puede afectar a la obtención de proteínas que interaccionan débilmente con las proteínas etiquetadas [18]. Los datos de interacción obtenidos por métodos de alto-rendimiento (High-throughtput) que usan etiquetas (tags) adolecen de una gran sesgo en favor de las proteínas que más abundan, esto no ocurre con métodos tales como los basados en doble-híbrido o chips de proteoma muestran menor sesgo en este caso. La proteína cepo etiquetada está ligeramente sobre-expresada si se comparan los niveles de expresión. Además la proteína cepo está unida a una etiqueta artificial externa (exógena), introducida en la célula, puede causar problemas, por ejemplo, la etiqueta se encierra dentro de un complejo y no se encuentra, se puede decrementar la afinidad dentro de un complejo o la etiqueta puede afectar a la localización intracelular de la proteína cebo [18].

D.3.2 Tandem affinity purification (TAP)-tags

Uno de los métodos mejor conocidos para purificar complejos de proteínas es el método TAP. El concepto básico de TAP del llamado TAP-tag, el cual se une a una proteína específica. Los TAP-tag consisten en dos tags (etiquetas) de afinidad secuencial espaciados por un lugar de segmentación (cleavage site) de la proteasa TEV (tobacco etch virus, virus del grabado de tabaco). La proteasa TEV descompone una determinada secuencia muy poco común en proteínas de mamíferos (Glu-X-X-Tyr-X-Gln/Ser). El uso de la proteasa TEV minimiza por tanto el riesgo de separar la proteína cebo y/o proteína asociadas [18].

Una importante ventaja de las técnicas TAP es que la cantidad de enlaces no específicos se reduce comparado a otras técnicas basadas en afinidad. Esto se logra

gracias al uso de dos pasos de purificación. Primero, la proteína TAP-etiquetada (TAP-tagged) es expresada en una linea de célula u organismo adecuado y se permite que se asocie con sus objetivos endógenos. Posteriormente a la lisis de la célula, la proteína TAP-tagged se permite que se enlace via la primera parte de la etiqueta TAP (TAP-tagged, p.ej.: proteína A) a una columna específica (p. ej.: inmunoglobulinas inmobilizadas). Después del enjuague, la proteinasa-TEV es añadida y el TAP-tag es segmentado, dejando la primera etiqueta (tag) de afinidad en la columna. La proteína cebo, todavía fusionada con la segunda parte del TAP-tag, se une entonces a la segunda columna, la cual es enjuagada y separada por lavado químico (eluted). Aunque los métodos TAP-tag son altamente sensibles y selectivos, un problema potencial del método es que la pureza incrementada tiene un precio; las interaciones proteína-proteína de una naturaleza más transitoria pueden perderse durante la serie de pasos de la purificación. Otro problema es que se necesita una cantidad de gran material para comenzar. Por ello los métodos TAP se ha usado con limitado éxito con célula de mamíferos. Sin embargo, un método eficiente TAP-tag basado en la proteína G llamado GS-TAP ha sido desarrollado para la purificación de complejos de proteínas de células de mamíferos [18]. En la figura D.3 se muestra un esquema simple de TAP.

D.3.2.1 Strep-tag III (Estreptococos-etiquetado III)

A veces el uso de tags (etiquetas) simples se prefiere al uso de TAP-tags porque puede incrementar el rendimiento de objetivos enlazados de baja-especificadad. Sin embargo, la cantidad de fondo suele también incrementarse. Un método alternativo a los TAP-tags para la purificación de complejos de proteínas para células de mamíferos es Strep-tag III y ha sido comercializado por la IBA (Göttingen, Alemania). La elución (lavado químico para separar materiales) de la proteína de fusión Strep-tag III (ver glosario Proteínas de fusión) se realiza gracias a la adición biotina. Este método pretende minimizar la interferencia con la formación de complejos [18].

D.3.3 Identificación de interacciones de proteína-protínas mediante proteómica cuantitativa

Uno de los retos más grandes cuando se trabaja con interacciones proteína-proteína es distinguir enlaces específicos de enlaces no específicos. Esto es particularmente importante cuando se usa espectrometría de masas (MS) como método de identificación final ya que estos métodos son tan sensibles con cualquier prueba de detección (screen o muestreo) de interacción proteína-proteínas resultará en un gran número de proteínas identificadas contaminantes. Tap puede decrementar el número de enlaces no específicos pero las interacciones débiles se pierden y por tanto, a veces, interesa usar métodos que las preserven (p.ej.: coinmunoprecipitación, enfoques de etiqueta -tag- única, o cromatografías de afinidad). Es importante identificar proteínas contaminantes. Una herramienta excelente para la identificación de falsos positivos es la proteómica cuantitativa, es decir, técnicas que etiquetan las proteínas de forma

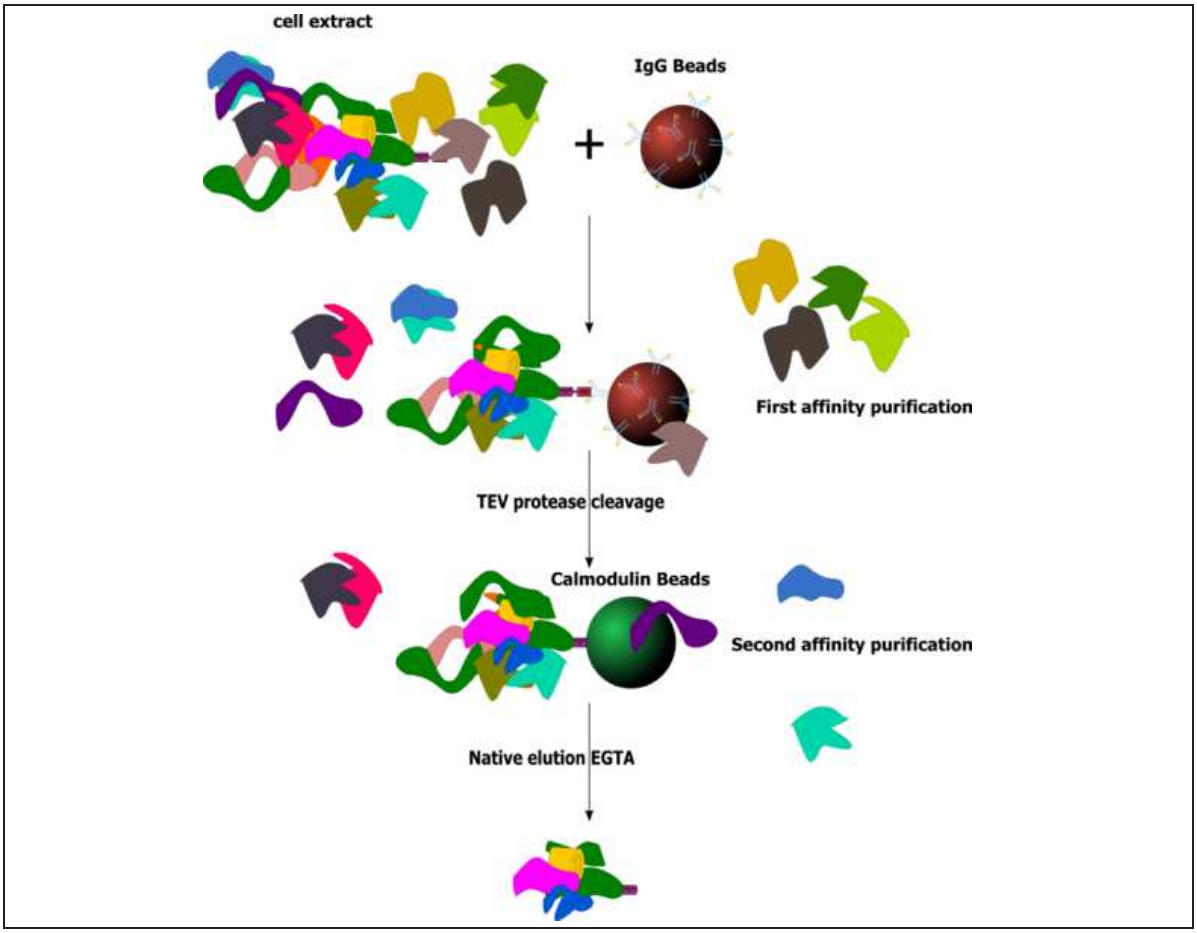


Fig. D.3: Esquema simple de un experimento de purificación de complejos de proteínas usando el método TAP (Tandem Affinity Purification). Figura extraída de http://upload.wikimedia.org/wikipedia/commons/e/e9/Taptag_simple.png bajo Creative Commons Attribution-Share Alike 3.0 Unported license.

diferente con isótopos estables y que comparar su abundancia relativa en diferentes muestras usando MS [18]. Podemos clasificarlos en dos tipos diferentes [18]:

- Etiquetado metabólico, los isótopos estables se incorporan, por ejemplo, en los aminoácidos en diversos puntos en el crecimiento de la célula.

- Etiquetado químico, agentes químicos pesados y ligeros se unen a aminoácidos reactivos de proteínas "post-cosecha". En un experimento típico, se usan dos muestras, la primera muestra se etiqueta con isótopos de masa ligera, mientras que por el contrario, las proteínas de la segunda muestra se etiquetan con isótopos pesados. Se mezclan las dos muestras, y las proteínas se digieren enzimáticamente y se analizan mediante MS. Así se pueden distinguir los péptidos que contienen

péptidos pesados o ligeros basados en la diferencia de masas. Otros métodos de etiquetado químico son el SILAC [158] o el ICAT [176].

D.3.4 Enlazado químico (Chemical crosslinking)

El crosslinking hace referencia a los enlaces covalentes formados entre cadenas de polímeros tanto dentro de la cadena como entre cadenas. Un enfoque clásico para determinar la interacción proteína-proteína ha sido por crosslinking químico. Uno de los primeros enfoques para determinar cada gran complejo (p.ej. ribosomas) mediante el uso de re-agentes homobifuncionales. Estos agentes son generalmente mucho menos selectivos y permite la independencia de enlaces covalentes (crosslinking) en la secuencia de dominios enlazados [18].

D.3.5 Los sistemas doble híbrido

El sistema doble híbrido es uno de los más ampliamente usados para prueba de detección (screening, muestreo) o para confirmar las interacciones proteína-proteína. Se basa en el hecho de muchos factores de transcripción eucariota están compuestas de dos dominios funcionalmente distintos que median entre la activación transcripcional y el enlazado de ADN respectivamente. Las ventajas del sistema Y2H (doble híbrido de la levadura) es que es simple de montar, no es caro, requiere poca optimización y detecta interacciones de proteínas in vivo. Tiene también varias desventajas, por ejemplo, sólo se pueden estudiar las interacciones binarias y el método genera un gran número de falsos positivos. No se sabe la tasa de falsos positivos pero se estima que es mayor de 50% de las interacciones identificadas. Existen multitud de variantes de este sistema en la literatura [18].

En la figura D.4 se muestra un diagrama simple del sistema doble híbrido.

D.3.6 Verificación de interacciones

Un objetivo potencial puede pasar todos los experimentos de control, pero, la interacción puede todavía haberse obtenido de forma indirecta (indirect via a common interaction partner) o ser simplemente no fisiológica. Por lo tanto, es importante verificar la interacción, incluso, en algunos casos, con el propio método usado [18].

D.3.6.1 Microscopio confocal

La co-localización se define como la presencia de dos o más diferencias moleculares que residen en la misma localización física en la célula. Si las proteínas interaccionan in vivo, se espera que sean co-localizadas o al menos, que se muestre una distribución que se solape dentro de la célula. Por lo tanto, la localización intracelular de dos (o más) proteínas puede ser estudiada por microscopia confocal. Las células son primero transfectadas (ver glosario transfección) en un plásmido de expresión que codifica la primera proteína unida a una etiqueta (tag) específica y la segunda proteína se une a otra etiqueta. Finalmente, el espécimen se incuba con anticuerpos etiquetados con

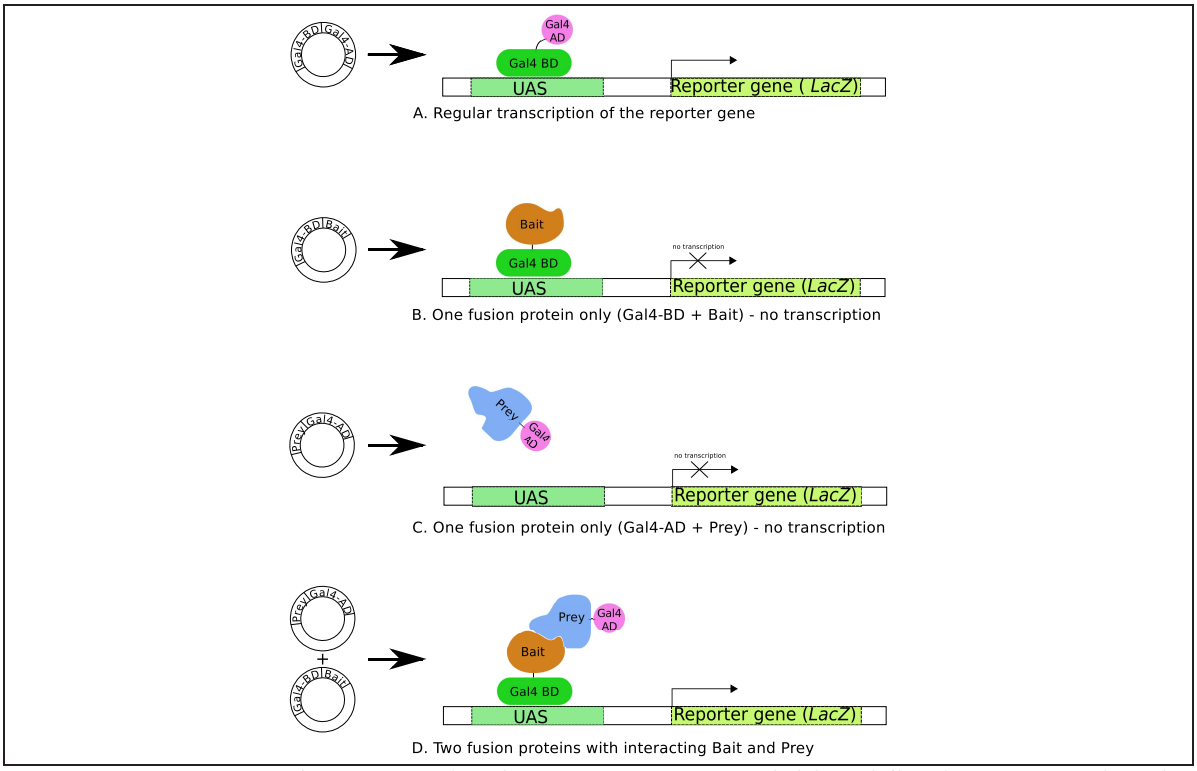


Fig. D.4: Visión general de un sistema doble híbrido comprobando la interacción dos proteínas (bait-prey, cebo-presa). Figura extraída de http://upload.wikimedia.org/wikipedia/commons/6/61/Two_hybrid_assay.svg bajo Creative Commons Attribution-Share Alike 3.0 Unported, 2.5 Generic, 2.0 Generic and 1.0 Generic license.

diferentes fluoróforos (que hace que una molécula sea fluorescente). Si dos proteínas son co-localizadas, las sondas fluoroforas estarán también co-localizadas. Esto es representado en una imagen generado por el microscopio confocal con píxeles que contienen ambas contribuciones de colores [18]. La sobreexpresión y/o el etiquetado puede a veces provocar que una proteína no llegue a estar bien localizada. Para poder solucionarlo, se usa como alternativa expresión temporal (transient expression) de proteínas etiquetadas, la localización de proteínas endógenas puede estudiarse entonces. Pero puede ocurrir que los componentes de los complejos de proteínas no sean los suficientemente expresado en las líneas de la célula que se está estudiando [18].

D.3.6.2 Co-IP: coinmunoprecipitación

Uno de los métodos más comunes de verificación es el co-IP. En el experimento típico, los complejos cebo se capturan a partir de por ejemplo un anticuerpo específico de un célula que se aplicó lisis. El anticuerpo es entonces inmovilizado usando la proteína A o la proteína G covalentemente unida a unas “gotas” (-beads- ciertas estructuras

periféricas) de sefarosa. Después del lavado de las gotas, el anticuerpo, el cebo y las proteínas asociadas al cebo puede ser identificadas por MS o por immunoblotting (es un método immunológico para aislamiento y medida cuantitativa de substancias inmunoreactivas) [18]. Puede llevarse a cabo de varios modos [18]:

1. Los experimentos de co-IP de línea de células (cell-lines) o de tejidos que expresan sus proteínas endóngenas pueden ser llevados a cabo. La ventaja es que se pueden estudiar los complejos de proteínas endógenas y se evita la sobreexpresión y los efectos negativos las etiquetas de afinidad (affinity tags).

2. También es posible usar células que se transfectan (ver glosario transfección) con un plásmido que codifica una proteína cebo etiquetada. Por tanto, un anticuerpo es dirigido directamente contra la etiqueta puede ser usado en los experimentos co-IP. Una ventaja de este enfoque es uno que puede ser relativamente confiable ya que el anticuerpo dirigido contra la etiqueta específica, la reacción no se cruza con otras proteínas.

3. Alternativamente, uno puede realizar experimentos co-IP usando células transfectadas con versiones etiquetadas con dos "compañeros" (partners) de interacción putativos.

Se comercializa en kits y pueden encontrar en el mercado con varios sistemas disponibles [18].

D.3.6.3 Estudios SPR (Surface plasmon resonance)

Los experimentos SPR (Surface plasmón resonance) ofrecen varias ventajas en la verificación de PPIs (protein-protein interactions- interacciones proteína-proteína). Las principales ventajas son las pequeñas cantidades de datos de muestra que son necesarias y no es necesario etiquetar. Además, el método proporciona información no solo en afinidades sino también en los índices de asociación y disociación que biológicamente son muy importantes, por ejemplo, en redes regulatorias. El método es genérico para todas las proteínas, porque depende del fenómeno de SPR atribuido a las superficies (películas) de los metales (normalmente a la plata y al oro), y a las señales grabadas (el ángulo de mínima luz de reflexión) que depende del índice de refracción cerca de la superficie. Con esta técnica, uno de los compañeros de la interacción (interaction partners) se inmoviliza en una superficie de una película de oro y el otro compañero se inyecta sobre la superficie. La cantidad enlazada es monitorizada continuamente en fucnión del tiempo, y después de conectarse a un flujo buffer (buffer flow, ver glosario buffer), la disociación del complejo puede ser monitorizado. Los índices constantes de disociación se obtienen ajustando una o múltiples bajadas exponenciales de los datos de disociación. Si la concentración de la proteína inyectada es conocida, los índices constantes de disociación se pueden obtener del ajuste los datos de la fase

de asociación. Para más información de cómo se realiza, recurrir a [154, 159]. Para proteínas multidominio, también modificaciones de este enfoque. Hay varias clases de superficies que se comercializan actualmente Una de las limitaciones importantes del SPR es que la proteína inmovilizada puede estar inactiva o que la interacción estudia esté cerca de la superficie por lo que los parámetros no reflejan la solución. Por tanto, existen experimentos de control que tienen en cuenta estas limitaciones [18].

D.3.6.4 Estudios de espectrometría

Las técnicas de espectrometría pueden ser usadas para verificas las interacciones proteína-proteína. El requisito es que los complejos muestren una diferencia relativa con los componentes libres en cualquiera de los parámetros de espectrometría (p.ej.: longitud de onda máximo o polarización). Si se produce un cambio en cualquier método, esto suministra una verificación de una interacción. Una ventaja de estas técnicas que permiten usar sondas fluorescentes y la interacción puede ser estudiada in vivo. La expresión de uno o dos compañeros que interaccionan como fusión de proteínas con diferentes derivadas de proteína fluorescente verde permite la verificación de las interacción en células vivas [18].

D.3.7 Métodos Blot de hibridación molecular

La hibridación molecular es de las metodologías fundamentales comúnmente utilizada en laboratorio en el ámbito de la biología molecular y ampliamente conocida en proteómica, es por ello que se dedica una sección a estos métodos aunque directamente no todos se usen para la detección de proteínas. La mayor parte de dichas técnicas han sido diseñado para identificar determinadas secuencias en los ácidos nucleicos. El término de hibridación molecular hace referencia al emparejamiento que de forma específica tiene lugar entre cadenas de ácidos nucleicos con secuencias complementarias. El proceso de hibridación es similar a la reacción reacción antígeno-anticuerpo se usan sondas en lugar de anticuerpos [142]. Las sondas son fragmentos cortos de ARN o ADN sintetizados in vitro que son marcadas con sustancias radiactivas fluorescentes o de otro tipo con el objetivo de posibilitar su posterior detección y así identificar la secuencia de ADN o ARN en estudio [39, 142, 182].

En los siguientes apartados se describe las técnicas de hibridación tipo blot que se utilizan normalmente en laboratorio. La diferencia entre ellas reside en el tipo de ácido nucleico que es usado, el soporte para la hibridación y cómo se distribuyen los ácidos en la membrana [182].

D.3.7.1 Southern blot

Es una técnica utilizada para la detección de secuencias específicas de ADN. Para ello, se aisla el ADN de un tejido o línea celular, posteriormente se purifica y se realiza una digestión con enzimas de restricción específicas. Los fragmentos obtenidos son separadas mediante electroforesis y luego son transferidos a la membrana que están

siendo usado de soporte (gel de agarosa). Dicho gel de agarosa se coloca sobre papel filtro que ha sido previamente remojado en una solución solución salina concentrada (llamada solución de transferencia). En el siguiente paso, la membrana es situada sobre el gel y encima de la membrana una pila de papel filtro. Por tanto, el papel filtro o seca y por capilaridad la solución de transferencia es atraída, entonces el ADN es arrastrado hacia la membrana donde se queda inmovilizado. Así se conserva la posición relativa del ADN sobre el gel. Es ahí entonces, donde el ADN puede hibridarse en la membrana con una sonda para marcarla [39, 45, 182].

Esta técnica se conoce como Southern blot porque Edwin Mellor Southern [197] propuso usar el término blot o blotting para secado (en inglés) y para referirse a esta técnica. Por eso se conoce a día de hoy como Southern blot [142]. Esta técnica es muy utilizada en el diagnóstico y caracterización de diversas inmunodeficiencias [39, 182].

D.3.7.2 Northern blot

Este método es una variante del Southern blot, de hecho la metodología es similar al Southern blot donde en lugar de usar ADN se utiliza ARN como el sustrato de estudio. El nombre fue nombrado así por analogía al anterior [39, 182, 25].

D.3.7.3 Western blot

Este método no realiza directamente un análisis de los ácidos nucleicos, sino de las proteínas que son el producto de la expresión génica. Además comprobamos que sigue el nombrado anterior de los métodos. La diferencia con respecto a los otros métodos anteriormente descritos es que no hay hibridación en la metodología, sino que se realiza una identificación de las proteínas mediante anticuerpos marcados. Así se puede conservar el principio de la transferencia a la membrana que permite la electrofóresis, para pasar luego a la identificación de proteínas [207].

Western blot se usa en ensayos para distinguir diferentes tipos de virus que infecta a células, en la variabilidad de los niveles individuales de determinadas enzimas o en la caracterización de genes a nivel molecular [39, 142, 182].

En la figura D.5 se muestra un diagrama de flujo simple del northern blotting.

D.3.7.4 Southwestern blot

Es un método es otra variante del Southern blot que fue descrito por primera vez por B. Bowen et al. en 1980 [28]. Dicho método se utiliza para la identificación y caracterización de las proteínas que se enlazan al ADN (ver definición en http://www.nlm.nih.gov/cgi/mesh/2011/MB_cgi?mode=&term=Southwestern+Blot), esto es debido a su capacidad para enlazarse con sondas específicas de oligonucleótidos. Las proteínas se separan mediante electroforesis y posteriormente se transfieren a membranas de nitrocelulosa como en otros métodos tipo blot.

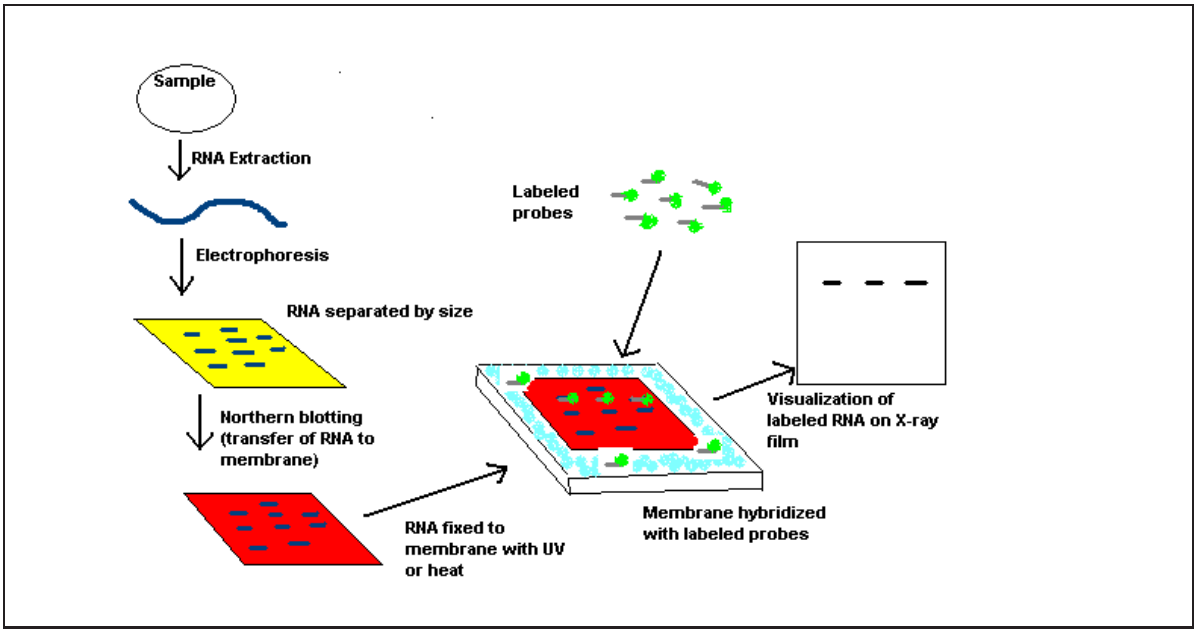


Fig. D.5: Diagrama de flujo que esboza el procedimiento general de detección de ARN por northern blotting. Figura extraída de http://upload.wikimedia.org/wikipedia/commons/e/e8/Northern_Blot_Scheme.PNG bajo domino público.

D.3.7.5 Dot Blot

Un Dot blot o Slot blot es una técnica usada para la detección de biomoléculas. El procedimiento es una simplificación de los métodos northern blot, Sourthern blot o western blot. En esta técnica, las biomoléculas (ARN) a ser detectadas no son separadas al comienzo por estrofóresis o cromatografía. La diferencia es que el ARN que se aplica directamente sobre la membrana. Para ello se utiliza un molde asociado a succión con vacío para colocar el ARN sobre la membrana y esto puede producir puntos (dot blot) o hendiduras (slot blot). Esta técnica permite un ahorro en tiempo (al no aplicar estrofóresis o cromatografía) y permite no permite ofrecer información sobre el tamaño de las bandas del ARN que es hibridado [182].

D.4 Referencias

Este apéndice, además de las referencias indicadas de la bibliografía, ha sido completado fundamentalmente gracias a información de los artículos que se encuentran "online" en el portal especializado de *quimica.es* (o su versión en inglés http://www.chemeurope.com/en) de la empresa alemana CHEMIE.DE (CHEMIE.DE Information Service GmbH, All rights reserved) que gestiona portales científicos especializados, ofreciendo servicios y software para laboratorios. Exclusivamente online y desde hace más de diez años. Los artículos principales fueron:

- http://www.quimica.es/enciclopedia/MALDI-TOF.html
- http://www.quimica.es/enciclopedia/SELDI.html
- http://www.quimica.es/enciclopedia/Espectrógrafo_de_masas.html

Se hace una mención especial a la documentación del curso de "Bioinformática Aplicada a la Proteómica" impartido por Integromics S.L. [99] y a www.wikipedia.es por las figuras de dominio público utilizadas y por aclaraciones de los conceptos básicos en la realización de estos apéndices escritos como ayuda de la lectura de esta tesis. El artículo más usado para la escritura de este apéndice fue http://es.wikipedia.org/wiki/Espectrómetro_de_masas.

También se ha utilizado la monografía realizada por Dr. Vladimir Ruiz Álvarez, Especialista de Primer Grado en Laboratorio Clínico. Máster en Bioquímica General. Profesor Asistente de Bioquímica y Laboratorio Clínico del Instituto de Nutrición e Higiene de los Alimentos de La Habana (Cuba) y el Dr. Rafael Menéndez Lorenzo, Especialista de Primer Grado en Medicina Interna, Profesor Asistente de Medicina Interna, Instituto superior de Ciencias Médicas de La Habana, Facultad de Ciencias Médicas Calixto García (Cuba) en la dirección http://www.monografias.com/trabajos20/biologia-molecular/biologia-molecular.shtml [182].

APPENDIX E

ENFERMEDADES NEURODEGENERATIVAS

En este apéndice se describen las enfermedades neurodegenerativas explicando las relaciones entre las enfermedades desde el punto de vista de mecanismos intracelulares como de genética.

La neurodegeneración es un término que tiene un sentido amplio para referirse desde la pérdida progresiva de las funciones o estructuras de las neuronas hasta la muerte de dichas neuronas. Dichas enfermedades pasando por las de Parkinson, Alzheimer o Huntington aparecen como resultado de un proceso neurodegenerativo.

El conocimiento de las enfermedades neurodegenerativas ha avanzando de forma muy notable en los últimos años gracias a los avances científicos tanto en los diferentes aspectos clínicos, terapéuticos como en la investigación genómica y proteómica. Hace tres décadas se conocía muy poco sobre las causas o de los mecanismos de estas enfermedades, y así desarrollar terapias más efectivas. Por todo ello, a día de hoy, se sabe que las enfermedades neurodegenerativas se deben a consecuencia de anormalidades ocurridas en el proceso de determinadas proteínas que intervienen en el ciclo celular, produciendo un acúmulo de dichas proteínas en las neuronas o en las proximidades de éstas, reduciendo o inhibiendo la función de esas neuronas [147].

El principal avance en este sentido ha sido el descubrimiento del prión en 1982 por Stanley B. Prusiner (Premio Nobel en 1997). El prión es un tipo de proteína que actúa como agente patógeno causando degeneración del sistema nervioso, pero que a diferencia del resto de agentes infecciosos (por ejemplo bacterias o virus), que contienen ácidos nucleicos (ya sea ADN, el ARN, o ambos), un prion carece de estos al estar compuesto solamente por aminoácidos. Por tanto, el conocimiento sobre las enfermedades neurodegenerativas, algunas de ellas producidas por estos priones, ha mostrado nuevas líneas de investigación en enfermedades que puede ser genéticas o infecciosas [147].

E.1 Enfermedades neurodegenerativas por proteopatías

Los últimos avances sobre biología molecular han permitido obtener un mayor conocimiento sobre las causas de las enfermedades neurodegenerativas por proteopatías o alteración de proteínas. El genoma humano dispone de unos 35000 genes, de estos genes se codificaran muchas proteínas que solo se expresa en el sistema nervioso. Asimismo, se codificarán proteínas tendrás diferente grado de expresión en determinados tipos de neuronas. Así pues, ciertos grupos de neuronas son muy susceptibles de cualquier alteración producido bien sea por variaciones genéticas o por factores ambientales o incluso ambos. Por este motivo, para una neurona, que es una célula de larga vida, cualquier leve perturbación puede tener consecuencias irreversibles [147].

El primer factor de riesgo es la edad. Las neuronas pierden gradualmente su función conforme la enfermedad progresa con la edad. El envejecimiento masivo de la población hace que las enfermedades como las de Alzheimer y la de Parkinson aparezcan no sólo en países más desarrollados sino también los que están en vías de desarrollo [147].

El análisis de los cuadros clínicos de los diferentes enfermos de estas enfermedades permite establecer aspectos comunes. El principal aspecto común es la presencia de determinadas proteínas que no pueden ser eliminadas correctamente por diversos motivos de las neuronas o su entorno. Según Prusiner, en las enfermedades neurodegenerativas [147]:

1. Reducción del “splicing“ (acoplamiento) de las proteínas degradadas.

2. Alteraciones en los genes que intervienen en el ”splicing‘.

3. Expresión génica impropia.

4. Plegamiento anormal de las proteínas.

5. Perturbaciones en el proceso degradación de proteínas (proteolisis).

6. Alteraciones en las modificaciones post-traduccionales (post-translational en inglés) de proteínas que acaban ser sintetizadas.

Resumiendo, las proteínas que se procesan de forma defectuosa, se acumulan cuando los mecanismos celulares de eliminación no funcionan correctamente. De esta forma, las proteínas que son muy específicas en determinadas rutas celulares y son mal procesadas producen por tanto el mal funcionamiento de determinadas neuronas. De forma progresiva se pierden selectivamente y simétricamente las neuronas de los sistemas motor sensorial y cognitivo. De todo ello se ha procedido a una clasificación nosológica nosológica de estas enfermedades neurodegenerativas basadas en los patrones celulares concretos [147]:

1. Enfermedad de Huntington muestra nivel muy bajo o ausencia de ácido gamma-aminobutírico del neostriado (región del cerebro).

2. Enfermedad de Alzheimer viene determinada por pérdidas neuronales, placas seniles (son depósitos extracelulares de beta-amiloide en el cerebro), ovillos neurofibriales (conglomerado anormal de proteínas compuesto por pequeñas fibrillas) y deficiencia en acetilcolina.

3. Enfermedad de Parkinson es definida por los cuerpos de Lewy y por la deplección (pérdida) de dopamina.

4. La esclerosis lateral amiotrófica se caracteriza porque los axones motores se hinchan y se encuentran inclusiones celulares (sustancias extrañas en el citoplasma).

En las siguientes secciones describiremos con más detalle las relaciones entre las enfermedades neurodegenativas y los mecanismos celulares que se ven afectados [147].

E.2 Relaciones entre las enfermedades neurodegenerativas

E.2.1 Genéticas

Muchas de las enfermedades neurodegenerativas están causadas por mutaciones genéticas, de hecho, la mayoría de ellas se localizan en genes que no tienen relación alguna. Sin embargo, el gen mutado tiene una característica común: la repetición del triplete de nucleótidos CAG. Este triplete codifica el aminoácido glutamina, y una repetición de CAG produce la cadena PolyQ o poliglutamina (en inglés polyglutamine) [147].

Las enfermedades en el que se produce este fenómeno se le llaman enfermedades heredadas de la poliglutamina. Una repetición de triplete CAG produce una patogénesis dominante. Los residuos extras de la glutamina puede tomar propiedades tóxicas de varias formas, entre las que podemos destacar el plegamiento anómalo de proteínas y degradación de rutas metabólicas, localización subcelular alterada, e interacciones anormales con otras proteínas celulares [137].

Otro factor a tener en cuenta es la alfa-sinucleína que puede agregarse formando fibrilas indisolubles en condiciones patogénicas caracterizadas por la aparición cuerpos de Lewy en enfermedades como la de Parkingson, demencia por cuerpos de Lewy y atrofia sistémica múltiple. La alfa-sinucleína es el principal componente estructural de fibrilas de cuerpos de Lewy. Además, un fragmento de la alfa-sinucleína, llamado componente no-Abeta (en inglés non-Abeta component o NAC), se encuentra en las placas de amiloide en la enfermedad del Alzheimer.

E.2.2 Mecanismos Intracelulares

E.2.2.1 Degradación de las rutas metabólicas de la proteína

La enfermedad de Parkinson y la de Huntington son de aparición tardía y están asociadas con la acumulación de proteínas tóxicas intracelulares. Las enfermedades causadas por la agregación de proteínas se les conocen como proteopatías, y están causadas en una primera instancia por agregados que aparecen en las siguientes estructuras [31]:

- citosol, por ejemplo, en el caso de la enfermedad de Parkinson y la de Huntington.
- núcleo, por ejemplo, la ataxia espinocerebelar tipo 1 (en inglés Spinocerebellar ataxia type 1 o SCA1)
- retículo endoplasmático (ER), es el caso de la encefalopatía familiar con cuerpos de inclusión de neuroserpina.
- proteinas excretadas extracelularmente, amiloide-β en enfermedad del Alzheimer.

En las células eucariotas, hay dos principales vías de eliminación de proteínas y orgánulos que están estorbando o produciendo conflictos en el interior de la célula:

- **Sistema ubiquitina-proteasoma:** La ubiquitina junto con grupo de enzimas especializas que trabajan en cascada es la clave para la degradación de muchas proteínas que causan proteinopatías incluyendo expansiones de poliQ y de alfa-sinucleínas. La investigación en este aspecto indica que las enzimas del proteasoma pueden no ser capaces de partir correctamente estas proteínas irregulares, lo cual podría ser tóxico en determinadas especies. Para más información ver sección 8.2.
- **Ruta metabólica autofago-lisosoma:** es una forma de muerte celular programada (en inglés programmed cell death o PCD). Esta ruta es usada cuando una proteína es propensa a formar agregados siendo de un sustrato pobre para el proteasoma. Hay dos formas de autofagia o autodigestión celular:
  - **Macrofagia:** está involucrado con el reciclaje de macromoléculas en condiciones de inanición, en ciertas rutas de apóptosis, y si está ausente entonces se produce la formación de inclusiones ubiquitinadas.
  - **Autofagia mediado por chaperona (en inglés chaperone-mediated autophagy CMA)**. Los defectos en esta forma de autofagia produce neurodegeneración. Las investigaciones han demostrado que las proteínas mutantes se unen a los receptores de la ruta CMA en membranas del lisosoma produciendo el bloqueo de su degradación así como de la degradación de otros substratos [31].

E.2.2.2 Degradación de las rutas metabólicas de la proteína

La forma más común de de muerte celular en neurodegeneración es a través de la ruta de apoptósis mitocondrial. Esta ruta controla la activacón de la caspasa-9 regulando la liberación el citocromo c desde el espacio intermembranal de la mitocondria. Las especies reactivas al oxígeno (ERO o en inglés ROS por reactive oxygen species) en el que se incluyen radicales libres, peróxidos tanto orgánicos como inorgánicos e iones de 0_2, son productos del metabolismo oxidativo normal generados por mitocondrias. Los niveles de ROS pueden aumentar en gran manera en épocas de estrés ambiental, y por tanto se puede producir daños significativos a las estructuras celulares, lo que se conoce como estrés oxidativo. Así pues, la sobreproducción de ERO es el principal rasgo de todas las enferemedades neurodegenativas. Además, la enfermedad mitocondrial puede llevar a neurodegeneración de la misma manera y en el mismo nivel a todas las funciones que estén relacionadas con la mitocontria: homeostais de calcio, PCD, fisión y fusión mitocondrial, concentración de lípidos en las membranas de la mitocondria, y transición de permeabilidad mitocondrial. Por consiguiente queda patente el hecho del efecto directo que la disfunción de la mitocondria tienen en las enfermedades neurodegenerativas.

E.2.2.3 Transporte axónico

El transporte de organelos, enzimas, agregados macromoleculares y metabolitos, es una función de axoplasma (citoplasma contenido dentro del axón) en la cuál intervienen directamente los microtúbulos. La tumefacción (hinchazón) de los axones y la aparición de "esferoides" o bolas de retracción axonal son comunes en muchas enfermedades neurodegenerativas. Esto sugiere que los axones afectados no están solo presentes en neuronas enfermas sino que puede pueden causar ciertas patologías a causa de la acumulación de orgánulos. Cuando el transporte axónico se interrumpe gravemente se produce la degeneración de Wallerian [61]. Para mayor información visitar la dirección web de histología en http://escuela.med.puc.cl/.

E.2.3 Muerte Celular programada

La muerte celular programada (en inglés programmed cell death o PCD) es un proceso de autodestrucción celular controlada que permite al organismo su correcta morfogénesis (es el proceso biológico que lleva a que un organismo desarrolle su forma), así como su renovación y la eliminación de las células que amenacen su supervivencia [74, 114, 104, 174]. Definiciones y más información en http://www.ciencia.cl/CienciaAlDia/volumen3/numero3/articulos/articulo5.html.

E.2.3.1 Apóptosis (tipo I)

Es una forma de muerte celular programada de organismos multicelulares. Hay dos rutas de apóptosis:

- **Ruta de apóptosis extrínsica**: tiene lugar cuando los factores externos de la célula activan los receptores de muerte celular desde la superficie (por ejemplo Fas) lo que resulta en la activacion de las caspasas 8 ó 10 [206].
- **Ruta de apóptosis intrínseca**: es el resultado de liberación por parte de la mitocondria de la citocromo c o del mal funcionamiento del retículo endoplasmático o de ambos. Todo ello lleva a la activación de la caspasa-9. El núcleo y el aparato de Golgi también puede activar la apóptosis si tienen los sensores dañados [206, 224].

Las caspasas (en inglés caspases) son un grupo de proteínas que pertenecen al grupo de las cisteín-proteasas (en inglés cysteine-proteases) y realizan la división o segmentación de residuos (aminoácidos) muy específicos. Las caspasas son mediadores fundamentales de los procesos de apóptosis celular y hay dos tipos: caspasas iniciadoras y caspasas efectoras. Las iniciadoras (por ejemplo caspasas 8 y 9) procesan las caspasas efectoras (por ejemplo las caspasas 3 y 7), segmentando las formas inactivas para activarlas. Las caspasas efectoras procesas (segmentando) otras proteínas o substratos proteícos que intervienen en la apóptosis [206].

E.2.3.2 Autofágico (tipo II)

La autofagia es una forma de fagocitosis intracelular, altamente conservado en células eucariotas, en el que la célcula consume los orgánulos dañados o proteínas que están mal plegadas encapsulándolas en la autofagosoma (vesículas de doble membrana) que capturan material citoplasmático y lo transportan hasta los compartimentos acídicos (lisosomas para las células de mamífero o vacuolas en el caso de levaduras), donde son degradados por enzimas hidrolíticas. Muchas enfermedades neurodegenerativas muestran un nivel inusual de agregados que pueden ser resultados por un defecto en el mecanismo de autofagia [206].

E.2.3.3 Citoplasmático (tipo III)

Muerte celular citoplasmática se produce a través de procesos no-apoptosicos, lo que se considera tipo III. Sin embargo, hay otras formas de muerte celular programada que no se comprenden completamente o que no se han sido aceptados por la comunidad científica por ejemplo es el caso de la aponecrosis que es una combinación de apoptosis y necrosis. No está clara exactamente la combinación de apoptosis, procesos non-apoptosicos y necrosis que podría causar diferentes clases de aponecrosis [206].

E.2.4 Muerte celular programada y neurodegeneración

Las diferentes enfermedades están involucradas en las rutas de apóptosis como no-apoptósicas en algún punto quedando patente su interdependencia con la muerte celular [206]. Generalmente, la muerte celular en neurodegeneración se debe a la apoptosis y lo más habitual es a través de la ruta intrínseca mitocondrial [128]. A continuación se explica brevemente la neurodegeneración en las diferentes enfermedades.

E.2.4.1 Enfermedad de Alzheimer

La enfermedad del Alzheimer (Alzheimer's disease) se describió por Alois Alzheimer en 1907 que señaló las alteraciones anatomopatológicas características de dicha enfermedad que fueron los ovillos neurofibriales y las placas seniles. Los ovillos neurofibriales son agregaciones de la proteína microtubular TAU que se encuentra hiperfosforilada. Las placas seniles son la consecuencia del acúmulo de varias proteínas en una reacción inflamatoria alrededor de los depósitos de la sustancia denominada β-amiloide [147].

Por tanto, todo lo anterior hace que esta enfermedad se caracterice por la pérdida de neuronas y sinapsis en la corteza cerebral y en ciertas regiones subcorticales provocada por la acumulación de proteínas TAU (ovillos neurofibriales) y A-beta (beta-amiloide placas) en el cerebro [171]. Esta pérdida provoca la atrofia grave de las regiones afectadas, ya sea el lóbulo temporal como el lóbulo parietal y partes de la corteza frontal y de la circunvolución del cíngulo [210].

E.2.4.2 Enfermedad de Huntington

La enfermedad de Parkingson se caracteriza por la progresiva demencia producida por graves pérdidas neuronales [147]. La enfermedad de Huntington está relacionada con el cromosoma 4p15.3 donde el gen HD contiene una repetición CAG inestable en su primer exon. El gen HD codifica una proteína denominada huntingtina. Sin embargo, cuando hay un número elevado de repeticiones CAG en el gen HD se expresa una proteína huntingtina anormal alargada de 40 a 150 residuos de glutamina.

Las pérdidas neuronales se deben a una acumulación anormal de la proteína huntingtina y ubiquitina en las regiones afectadas, formando inclusiones intranucleares. Dichas inclusiones son la consecuencia de que el proceso de transporte de proteínas (huntingtina, ubiquitina y otros componentes proteasómicos) hacia el núcleo falla y es ahí donde se encuentran estos depósitos, en las célculas afectadas. Así pues, las neuronas mueren por apoptosis y se eliminan [147].

E.2.4.3 Enfermedad de Parkinson

Es la segunda enfermedad neurodegenerativa más frecuente y sus síntomas son la rigidez, bradiquinesia y temblor provocada por la pérdida de neuronas y por la deplección de dopamina en el estriado (región del núcleo del cerebro) [147].

La enfermedad de Parkinson se caracteriza por una acumulación anormal de la proteína alfa-sinucleína unida a la ubiquitina en las células dañadas. De hecho, el complejo alfa-sinucleina-ubiquitina no puede ser degradado por proteasoma produciendo acumulaciones de proteínas que forman inclusiones citoplasmática que se llaman cuerpos de Lewy [147, 130].

E.2.4.4 Esclerosis lateral amiotrófica

La esclerosis lateral amiotrófica (en inglés amyotrophic lateral sclerosis ALS) es el proceso neuromotor más frecuente. En un enfermo típico, los músculos de la vía motora se atrofian y mueren las segundas neuronas (espinales) [147]. La causa de pérdida de neuronas motoras en la esclerosis lateral amiotrófica es desconocida. Hay un subgrupo de enfermos de la forma familiar (menos del 20%) que presentan mutaciones en el gen del cromosoma 21, superóxido dismutasa tipo 1 (SOD1 en inglés superoxide dismutase 1) que codifica una proteína que interviene en al regulación de los radicales libres intracelulares (enzima antioxidante de Cu/Zn). Sin embargo, dicha información está en discusión por la comunidad científica [147].

E.2.4.5 Envejecimiento y neurodegeneración

Uno de los mayores riesgos de las enfermedades neurodegenerativas es la edad. Las mutaciones mitocondriales junto a con el estrés oxidativo que contribuyen al envejecimiento [51]. Muchas de estas enfermedades son de aparición tardía en edad adulta, lo que quiere decir que existe algún factor que cambia con la edad en cada una de estas enfermedades [31]. El factor común de dichas enfermedades es que las neuronas gradualmente pierden su función conforme la enfermedad progresa con la edad.

E.3 Referencias

Esta sección ha sido extraída de la monografía de los doctores Francisco Mora Teruel (Catedrático de Fisiología Humana de la Facultad de Medicina de la Universidad Complutense de Madrid y Catedrático Adscrito de Fisiología Molecular y Biofísica de la Facultad de Medicina de la Universidad de Iowa en Estados Unidos), José María Segovia de Arana (catedrático de Medicina Interna, Universidad Autónoma de Madrid) en el documento llamado *segovia-neurodegenerativas-01.pdf* con título "Enfermedades Neurodegenerativas" de la url http://www.imsersomayores.csic.es/, usar búsqueda de documentación de la biblioteca del portal de mayores realizada entre el IMSERSO (Instituto de Mayores y Servicios Sociales) y el CSIC (Consejo Superior de Investigación Científicas de España) con el auspicio del Ministerio de Sanidad, Política Social e Igualdad. Fundamentalmente se usó el prólogo y el capítulo I de dicho trabajo [147].

Se hace una mención especial a www.wikipedia.es por los recursos y aclaraciones de este tema tan complejo. Fundamentalmente se usado el artículo en inglés http://en.wikipedia.org/wiki/Neurodegeneration. Otros artículos usados han sido:

- http://es.wikipedia.org/wiki/Especie_reactiva_de_oxígeno
- http://es.wikipedia.org/wiki/Ovillo_neurofibrilar
- http://es.wikipedia.org/wiki/Placa_senil

GLOSSARY

árbol filogenético Un árbol filogenético es una representación en árbol que muestra las relaciones evolutivas entre varias especies u otras entidades, de las cuales se cree que tienen una ascendencia común. El análisis molecular de secuencias también nos ha enseñado que hay una división en la raíz misma del árbol de la vida que es más fundamental que la división de 5 reinos que se enseña normalmente. En lugar de los dos tipos celulares canónicos, los procariotas y eucariotas, hay tres tipos principales de células, las arqueobacterias, las eubacterias y los eucariotas. Este nuevo árbol recibe el nombre de árbol filogenético universal. Fuente: Antonio Barbadilla. Departamento de Genética y Microbiología. Universidad Autónoma de Barcelona http://biologia.uab.es/divulgacio/evol.html. 36

anticuerpos Un anticuerpo es una sustancia producida en el organismo animal por la presencia de un antígeno, contra cuya acción reacciona específicamente. Fuente: Diccionario de la Lengua Española. Real Academia Española. http://drae2.es/. 208, 224

ayuste El "ayuste" (en inglés, splicing) sería por tanto el proceso de corte de los intrones y empalme de los exones en ARN durante la transcripción (ver ayustosoma). Fuente: Páginas de Bioquímica y Biología Molecular desarrollado por Dr. Ángel Herráez de la Universidad de Alcalá de Henares http://biomodel.uah.es/.. 199, 202

ayustosoma El "ayustosoma" o cuerpo de empalme (en inglés, spliceosome) es un complejo ribonucleoproteico responsable del proceso de corte de intrones y empalme de exones, una de las modificaciones postranscripcionales que sufre el RNA en eucariotas. Está formado por varias moléculas de RNA (del tipo nuclear pequeño, snRNA) asociadas con varias moléculas de proteína.. 220

biotina Es la Vitamina H, B7 o B8; interviene en el metabolismo de los hidratos de carbono, grasas, aminoácidos y purinas; alivia dolores musculares, el eczema y la dermatitis y también ayuda a combatir la depresión y la somnolencia. Fuente: Diccionario médico-biológico, histórico y etimológico de la Universidad de Salamanca. http://dicciomed.eusal.es ISBN: 8478005722 . 222

bradiquinesia es un trastorno caracterizado por el enlentecimiento de todos los movimientos voluntarios y el habla, como la causada por el parkinsonismo, otras alteraciones extrapiramidales y ciertos tranquilizantes. Fuente: Diccionario del Instituto de Química y Biología (IQB) http://www.iqb.es/diccio/diccio1.htm.. 237

buffer Un tampón (en inglés buffer) es una solución amortiguadora o solución reguladora es la mezcla en concentraciones relativamente elevadas de un ácido débil y su base conjugada, es decir, sales hidrolíticamente activas. Tienen la propiedad de mantener estable el pH de una disolución frente a la adición de cantidades relativamente pequeñas de ácidos o bases fuertes. Se puede entender esta propiedad como consecuencia del efecto ion común y las diferentes constantes de acidez o basicidad: una pequeña cantidad de ácido o base desplaza levemente el equilibrio ácido-base débil, lo cual tiene una consecuencia menor sobre el pH. Fuente: Morcillo, Jesús (1989). Temas básicos de química (2ª edición). Alhambra Universidad. p. 270-272. ISBN 9788420507828. 226

chaperona Se aplica a las proteínas pertenecientes a la maquinaria de síntesis celular que contribuyen al correcto plegamiento de las proteínas recién sintetizadas.. Fuente: Diccionario médico-biológico, histórico y etimológico de la Universidad de Salamanca. http://dicciomed.eusal.es ISBN: 8478005722. 234

citocromo Cualquiera de los pigmentos respiratorios hemoproteinicos intracelulares que son enzimas en el transporte de electrones. El citocromo oxidasa, una porfirina con hierro, es un enzima importante en la respiración. Fuente: Diccionario médico-biológico, histórico y etimológico de la Universidad de Salamanca. http://dicciomed.eusal.es ISBN: 8478005722. 234

cuerpos de Lewy son inclusiones (sustancias extrañas) citoplasmáticas eosinofílicas, constituidas por neurofilamentos que se acumulan por un defecto de la fosforilización. Están constituidos fundamentalmente por a-sinucleína y algunas otras proteínas como la ubiquitina, gelsolina y quinasas. Su acumulación en las neuronas de la corteza cerebral y de otros núcleos subcorticales ocasiona la llamada demencia con cuerpos de Lewy. Fuente: Diccionario del Instituto de Química y Biología (IQB) http://www.iqb.es/diccio/diccio1.htm.. 233

dopamina Es una hormona que actúa como neurotransmisor; es precursor inmediato de la noradrenalina. Fuente: Diccionario médico-biológico, histórico y etimológico de la Universidad de Salamanca. http://dicciomed.eusal.es ISBN: 8478005722. 237

e-value E-value (valor esperado o esperanza matemática) se usa comúnmente en pruebas de alineamiento de secuencias. Estas pruebas de significancia incluyen el Valor E, el cual representa el número de alineamientos con un score equivalente o superior que se pueden presentarse por azar. De esta forma la interpretación del Valor E, indica que tan confiable es el alineamiento obtenido, siendo los valores cercanos a cero los que indican menor posibilidad de obtención del resultado por azar. Fuente: Documentación de un curso del Centro de Bioinformática del Instituto de Biotecnología de la Universidad Nacional de Colombia http://bioinf.ibun.unal.edu.co/cbib/ . 55

endocitosis La endocitosis es el proceso de incorporación de moléculas al interior celular englobadas en vesículas. Fuente: Atlas de Histología Vegetal y Animal de la Universidad de Vigo http://webs.uvigo.es/mmegias/5-celulas/5-endocitosis.php . 185, 187

enzima Una enzima es una proteína que actúa como catalizador de una reacción química acelerándola. Las enzimas son protagonistas fundamentales en los procesos del metabolismo celular. Las enzimas unen su sustrato en el centro reactivo o catalítico, que suele estar protegido del agua para evitar interacciones no deseadas. Fuente: Glosario de Medicina Molecular del Portal de Gestión del Conocimiento de FIBAO (Fundación Pública Andaluza para la Investigación Biosanitaria de Andalucía Oriental Alejandro Otero) http://www.medmol.es/glosario.cfm.. 61, 187, 194, 198, 208, 215

epitelio Los epitelios constituyen uno de los cuatro tejidos fundamentales de los animales. El tejido epitelial recibe distintos nombres según donde se localize. Por ejemplo, en la piel se denomina epidermis, cuando recubre cavidades internas como la cavidad cardíaca, pulmonar o abdomen se llama mesotelio, y el epitelio que forma la superficie interna de los vasos sanguíneos y linfáticos es el endotelio. Fuente: Atlas de Histología Vegetal y Animal de la Universidad de Vigo http://webs.uvigo.es/mmegias/guiada_a_epitelios.php . 185

espectrometría de masas Espectrometría de masas es una técnica analítica que ha experimentado un gran desarrollo tecnológico en los últimos años. Permite estudiar compuestos de naturaleza diversa: orgánica, inorgánica o biológica y obtener información cualitativa o cuantitativa. Mediante el análisis por Espectrometría de masas es posible obtener información de la masa molecular del compuesto analizado así como obtener información estructural del mismo. Fuente: Documentación del

Laboratorio de Espectrometría de Masas de la Universidad Autónoma de Madrid. http://www.uam.es/investigacion/servicios/sidi/especifica/masas.html . 50, 215

fagocitosis La fagocitosis es una endocitosis especializada en incorporar grandes partículas como bacterias, virus y restos celulares. Fuente: Atlas de Histología Vegetal y Animal de la Universidad de Vigo en la página web de la siguiente dirección http://webs.uvigo.es/mmegias/5-celulas/5-endocitosis.php . 185

fenotipo Rasgos o características visibles de un organismo, por ejemplo, el color del cabello, el peso o la presencia o ausencia de una enfermedad. Los rasgos fenotípicos no son necesariamente genéticos. Fuente: Glosario en español del Instituto Nacional de Investigaciones del Genoma Humano (National Human Genome Research Institute). National Institutes of Health (EEUU). http://www.genome.gov/sglossary.cfm?ID=111. 56, 66, 194

fluoróforo Fluoróforo o fluorocromo es una sustancia que tiene la propiedad de convertir en fluorescentes los objetos que impregna (técnica de la microscopía fluorescente). Fuente: Definición de Dr. Alberto Martín Lasa. Portales Médicos S.L. http://www.portalesmedicos.com/diccionario_medico. 224

hidrólisis Desdoblamiento de la molécula de ciertos compuestos orgánicos por la acción del agua; reacción ácido-base entre una sustancia, típicamente una sal, y el agua. Fuente: Diccionario médico-biológico, histórico y etimológico de la Universidad de Salamanca. http://dicciomed.eusal.es ISBN: 8478005722. 187

ligando Un ligando es una molécula capaz de ser reconocida por otra provocando una respuesta biológica. Por ejemplo, podemos encontrar ligandos como las hormonas implicados en la señalización intercelular, o ligandos que actúan como reguladores de la actividad enzimática uniéndose a las enzimas que modulan. Fuente: Glosario de Medicina Molecular del Portal de Gestión del Conocimiento de FIBAO (Fundación Pública Andaluza para la Investigación Biosanitaria de Andalucía Oriental Alejandro Otero) http://www.medmol.es/glosario.cfm. . 61, 186, 221

lisis Destrucción o disolución de células o bacterias. Fuente: Diccionario médico-biológico, histórico y etimológico de la Universidad de Salamanca. ISBN: 8478005722 en la siguiente web http://dicciomed.eusal.es . 188, 221, 222

nosológica (De noso, enfermedad y del griego logos, tratado). Estudio de los caracteres distintivos que permiten definir las enfermedades. Fuente: Definición de Dr. Vicens Nieto. Portales Médicos S.L. http://www.portalesmedicos.com/diccionario_medico. 232

nucleoplasma El nucleoplasma o carioplasma es una masa o contenido fundamental del núcleo en el que están los corpúsculos: nucléolos, cromosomas o sus restos. Fuente: Diccionario médico-biológico, histórico y etimológico de la Universidad de Salamanca. http://dicciomed.eusal.es ISBN: 8478005722. 186

nucleosoma Los nucleosomas son una formación nuclear en la que el ADN se enrolla alrededor de proteínas de tipo histona. Es el primer nivel de enrollamiento de ADN. Para transcribir el ADN que lo forma hay que deshacer el nucleosoma. Fuente: Glosario de Medicina Molecular del Portal de Gestión del Conocimiento de FIBAO (Fundación Pública Andaluza para la Investigación Biosanitaria de Andalucía Oriental Alejandro Otero) http://www.medmol.es/glosario.cfm. 189

nucleótido Molécula orgánica formada por una base nitrogenada, una pentosa y ácido fosfórico, que constituye la unidad estructural de los ácidos nucleicos. Según que la pentosa sea la ribosa o la desoxirribosa, el nucleótido resultante se denomina ribonucleótido o desoxirribonucleótido. Fuente: Diccionario médico-biológico, histórico y etimológico de la Universidad de Salamanca. http://dicciomed.eusal.es ISBN: 8478005722. 202

nucléolo El nucléolo es una región del nucleoplasma donde se sintetiza el ARN ribosómico, se procesa y se ensambla con proteínas para formar las subunidades ribosómicas. Morfológicamente contiene varias regiones: centro fibrilar, componente fibrilar denso y componente granular. Tanto el número como el tamaño de los nucléolos puede variar en función del tipo celular y del estado fisiológico en el que se encuentre la célula. Fuente: Atlas de Histología Vegetal y Animal de la Universidad de Vigo http://webs.uvigo.es/mmegias/5-celulas/4-nucleolo.php . 186

ORF Marco de lectura abierta - Open reading frame (sinónimo: ORF) Todos los exones de un gen que contribuyen a los productos proteicos del gen. Fuente: Glosario de Genética del Instituto Roche de Investigación Biomédica. La dirección web es la siguiente http://www.institutoroche.org/Recursos_glosario/Va.html. 65

PCR Una técnica rápida y económica para hacer un número ilimitado de copias de cualquier fragmento del ADN. La PCR, o fotocopiado molecular como se le dice algunas veces, ha tenido un inmenso impacto en biología y medicina, y especialmente en la investigación genética. Fuente: Glosario en español del Instituto Nacional de Investigaciones del Genoma Humano (National Human Genome Research Institute). National Institutes of Health (EEUU). http://www.genome.gov/sglossary.cfm?ID=163. . 215

peptidoglicano El peptidoglicano es un heteropolímero formado por dos derivados de azúcares N-acetilglucosamina y N-acetilmurámico, y un pequeño grupo de

aminoácidos que incluyen L-alanina, D-alanina, D-glutámico, y lisina o en otros casos ácido diaminopimélico (DAP). Estos componentes se unen entre sí para formar una estructura que se repite a lo largo de la pared celular. Fuente: Tesis Doctoral de Dra. Mari Luz Mohedano Bonillo titulada CARACTERIZACIÓN FUNCIONAL DEL REGULADOR ESENCIAL YycF DE "STREPTOCOCCUS PNEUMONIAE". UNIVERSIDAD COMPLUTENSE DE MADRID. Madrid 2010. ISBN: 978-84-692-9929-6 eprints.ucm.es/9980/1/T31284.pdf . 189

pinocitosis El término pinocitosis se refiere a este tipo de endocitosis inespecífica de moléculas disueltas. Fuente: Atlas de Histología Vegetal y Animal de la Universidad de Vigo http://webs.uvigo.es/mmegias/5-celulas/5-endocitosis.php . 185

plasmón Un plasmón, en física, es un cuanto de oscilación del plasma (es decir, el estado de la materia). Los plasmones de superficie de superficie son, por definición, oscilaciones de la densidad de carga electrónica que puede existir en una interfase metal/dieléctrico. Asociado a éstos existe un campo electromagnético que se propaga a lo largo de la interfase, cuya amplitud decae exponencialmente en la dirección normal a la superficie. Fuente: Carlos Martínez Hernández / Víctor M. Coello Cárdenas. MODELADO DEL ESPARCIMIENTO ELÁSTICO DE PLASMONES DE SUPERFICIE Ciencia UANL, julio-septiembre 2005, vol. VIII, número 003. Universidad Autónoma de Nuevo León Monterrey, México pp. 346-350 redalyc.uaemex.mx/pdf/402/40280307.pdf. 226

plásmido Los plásmidos son fragmentos extracromosómicos de ácidos nucléicos (ADN o ARN) que aparecen en el citoplasma de algunos procariotas. Son de tamaño variable aunque menor que el cromosoma principal. Cada bacteria puede tener uno o varios a la vez. Fuente: Glosario de Medicina Molecular del Portal de Gestión del Conocimiento de FIBAO (Fundación Pública Andaluza para la Investigación Biosanitaria de Andalucía Oriental Alejandro Otero) http://www.medmol.es/glosario.cfm. . 189, 221

promotor El promotor de un gen es una región del ADN con unas características especiales que determina el punto en el que la ARN polimerasa comienza a transcribir un gen. Las características del promotor también influyen en la eficiencia de la transcripción. Esta región incluye las secuencias de unión a factores de transcripción y a otros elementos que participan en la transcripción. Fuente: Glosario de Medicina Molecular del Portal de Gestión del Conocimiento de FIBAO (Fundación Pública Andaluza para la Investigación Biosanitaria de Andalucía Oriental Alejandro Otero) http://www.medmol.es/glosario.cfm. . 213

Proteínas de fusión La tecnología de ADN recombinante ha hecho posible la concepción, diseño y elaboración de moléculas artificiales llamadas proteínas de fusión, en

las cuales, motivos estructurales y/o funcionales, provenientes de dos o más proteínas naturales, son combinados. Fuente: Documentación del tema "Anticuerpos monoclonales de segunda generación" del sistema Academia Biomédica Digital VITAE del Centro de Análisis de Imágenes Médicas Computarizadas (CAIBCO). Instituto de Medicina Tropical. Facultad de Medicina de la Universidad Central de Venezuela. Web: http://caibco.ucv.ve/. 36, 56, 71, 221, 222

péptido Un péptido esta formado por dos o más aminoácidos ensamblados por un enlace peptídico. Fuente: Glosario en español del Instituto Nacional de Investigaciones del Genoma Humano (National Human Genome Research Institute). National Institutes of Health (EEUU). http://www.genome.gov/sglossary.cfm?ID=163. 200

t-student La Prueba t-student es uno de los análisis estadísticos más comunes en la práctica es probablemente el utilizado para comparar dos grupos independientes de observaciones con respecto a una variable numérica. Fuente: Pértega Díaz S., Pita Fernández S. Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Universitario de A Coruña (España) CAD ATEN PRIMARIA 2001; 8: 37-41. Métodos paramétricos para la comparación de dos medias. t de Student http://www.fisterra.com/mbe/investiga/t_student/t_student.asp . 55

transfección La técnica de transfección génica mediante vectores virales aprovecha la capacidad infectiva de los virus para introducir fragmentos de ADN de interés en células de mamíferos. Fuente: Glosario de Medicina Molecular del Portal de Gestión del Conocimiento de FIBAO (Fundación Pública Andaluza para la Investigación Biosanitaria de Andalucía Oriental Alejandro Otero) http://www.medmol.es/. 226

ubiquitina proteína pequeña y altamente conservada presente en todas las células eucariotas que se une covalentemente a las lisinas de otras proteínas, marcando dicha proteína para que sea degradada por proteolisis intracelular en un proteosoma. Fuente: Diccionario médico-biológico, histórico y etimológico de la Universidad de Salamanca. ISBN: 8478005722, en la web http://dicciomed.eusal.es . 234