



UNIVERSIDAD DE CHILE  
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS  
DEPARTAMENTO DE INGENIERÍA QUÍMICA Y BIOTECNOLOGÍA

**USO DE ALGORITMOS DE CLUSTERING PARA PREDECIR EL  
COMPORTAMIENTO DE PROTEÍNAS EN CROMATOGRAFÍAS DE  
INTERACCIÓN HIDROFÓBICA Y SISTEMAS DE DOS FASES ACUOSAS**

TESIS PARA OPTAR A LOS TÍTULOS DE INGENIERO CIVIL EN BIOTECNOLOGÍA Y  
MAGÍSTER EN CIENCIAS DE LA INGENIERÍA MENCIÓN QUÍMICA

**JORGE ENRIQUE UGARTE HUMERES**

PROFESOR GUÍA:  
**DR. J. CRISTIAN SALGADO HERRERA**

PROFESOR CO-GUÍA:  
DR. ÁLVARO OLIVERA NAPPA

MIEMBROS DE LA COMISIÓN:  
DRA. BÁRBARA ANDREWS FARROW  
DRA. M. ELENA LIENQUEO CONTRERAS  
DRA. ANDREA MAHN OSSES

SANTIAGO DE CHILE  
MAYO 2012

## Resumen

El principal coste en la industria biotecnológica se produce en I+D, alcanzando un 53% de los ingresos en USA y 63% en Europa (1997-1999). Esto se explica por la complejidad de las técnicas utilizadas, como en algunos procesos de separación y purificación de proteínas. Para disminuir los costes en I+D, se puede reducir el tiempo de diseño de éstas utilizando modelos.

Dos técnicas utilizadas extensamente para la separación de proteínas son la cromatografía de interacción hidrofóbica (HIC) y los sistemas de dos fases acuosas (ATPS), para las cuales existen diversos modelos predictivos que se basan en una o más propiedades del sistema y/o la molécula a separar. Las principales limitantes de estos modelos son la capacidad predictiva, y la cantidad y coste de la información requerida. En los modelos que utilizan hidrofobicidad, una limitante adicional es la escala de hidrofobicidad utilizada. Por esto, el presente trabajo tiene como objetivo generar nuevas escalas de hidrofobicidad que mejoren el poder predictivo de modelos reportados para el tiempo de retención adimensional (DRT) de proteínas en HIC, y el coeficiente de partición (K) de proteínas en cuatro tipos de sistemas ATPS.

Se analizó un grupo reportado de 74 escalas de propiedades aminoacídicas (APVs), mediante los siguientes algoritmos de *clustering*: *Growing Neuronal Gas* (GNG), *Growing Grid* (GG), *Hierarchical Clustering*, *Bisection Algorithm*, *Restricted Neighbourhood Search Algorithm*, y *Markov Clustering Algorithm*. Se utilizó también el algoritmo de optimización *Genetic Algorithm* (GA). Para la predicción de DRT y K, en cada caso se utilizó un modelo que requiere la estructura tridimensional de las proteínas y tres modelos que solo requieren la composición aminoacídica, los que calculan o predicen la hidrofobicidad superficial media (ASH). El poder predictivo de los modelos se calculó mediante validación cruzada de Jackknife.

A través de la metodología empleada se obtuvo 308.000 nuevas escalas, de las cuales un 93% se generó con GNG, GG y GA, incluyendo las escalas más exitosas. En general, la utilización de las nuevas escalas permitió desarrollar modelos con un mejor poder predictivo que los basados en escalas reportadas en literatura. Estas mejoras se reflejaron en un aumento del poder predictivo entre un 11% y un 99,6% en un 81% de los casos con respecto al caso base. De forma simultánea, dentro de los modelos con aumento del poder predictivo se obtuvo mejoras en el nivel de ajuste, medido a través del Coeficiente de Pearson, de un 4% a un 300% en 28 de 42 casos (67%).

A partir del estudio de las mejores escalas obtenidas y los APVs, se concluyó que existe transferencia de propiedades desde estos últimos a las escalas generadas con GNG y GG. Por otro lado, se descartó transferencia de propiedades a las escalas generadas con GA, sin embargo, se validó su uso.

Se determinó que las mejores escalas contienen información de APVs asociados a estudios de: hidrofobicidad en sistemas físicoquímicos (HIC y ATPS), hidrofobicidad de aminoácidos en proteínas, y propensión conformacional de aminoácidos en proteínas. Los resultados obtenidos sugieren que incluir APVs del tipo conformacional permite mejorar las escalas obtenidas, disminuyendo el sesgo introducido por el uso de la ASH. Lo anterior sugiere que una escala que refleje la probabilidad de ocurrencia de cada aminoácido en distintos tipos de estructuras-configuraciones existentes en la superficie de las proteínas, y que incorpore el potencial hidrofóbico de cada de éstas, podría ser útil para mejorar el poder predictivo de los modelos.

En conclusión, a través del uso de algoritmos de *clustering* y *optimización* se logró un aumento significativo del poder predictivo de los modelos para HIC y ATPS, el que incluso es mayor al que se obtiene con otros modelos que incorporan directamente más información experimental, lo que permite reducir costes en I+D. La contribución realizada postula nuevas interrogantes y sugiere caminos que amplían y perfeccionan la búsqueda de metodologías para generar mejores modelos predictivos del comportamiento de proteínas en sistemas de separación, que requieren sólo la composición aminoacídica de las proteínas.

## Summary

The main costs in the biotech industry arise from R&D expenditure, which accounts for 53% of revenues in the US and 63% in Europe (1997-1999). The reason for these costs is that the techniques used, such as, for example protein separation and purification, are complex. In order to minimise R&D costs, the time spent on technique design may be reduced by using models.

Two of the most widely used protein separation/purification techniques are: hydrophobic interaction chromatography (hereinafter HIC) and aqueous two-phase systems (hereinafter ATPS). There are various predictive models based on one or more of the system's and or molecule's properties to be analysed. The main limitations of these predictive models are their predictive ability and the cost of the required information. In hydrophobicity models, the hydrophobicity scale used further limits an accurate prediction. In light of the foregoing, the present study aims to generate new hydrophobicity scales for enhancing the predictive power of reported models in relation to protein dimensionless retention time (DRT) for proteins separated/purified by HIC and partition coefficient (K) of proteins separated/purified by using four types of two-phase aqueous systems.

In order to generate the new scales, a group of 74 known amino acid property scales (hereinafter APVs) were analysed with the following clustering algorithms: Growing Neural Gas (GNG), Growing Grid (GG), Hierarchical Clustering, Bisection Algorithm, Restricted Neighborhood Search Algorithm and Markov Clustering Algorithm. Further, an optimisation algorithm, referred to as Genetic Algorithm, was used. DRT and K were predicted by means of one model that requires 3D structure of proteins and three models that require only amino acid composition of proteins, which calculate or predict the average surface hydrophobicity (ASH). The predictive power of each model was determined by Jackknife cross-validation.

The methodology used resulted in 308.000 new scales, 93% of which were generated by GNG, GG and GA, including the most successful scales. In general, the use of these new scales facilitated the development of models having a better predictive power than those based on scales which have been reported in the relevant literature. The improvements were reflected in an increase of between 11% and 99.6% predictive power in 81% of the cases. Simultaneously, within the 81% of cases in which predictive power improvements were observed, levels of fit also improved, as measured by a Pearson coefficient of 4% to 300% in 28 of 42 cases (i.e. 67%).

By analysing the best scales and APVs, it was concluded that properties of the latter are transferable to scales generated with GNG and GG. On the other hand, no transference of properties to those scales generated with GA was observed; nevertheless, their use was validated.

It was determined that the best scales include information of APVs associated with study of: hydrophobicity in physicochemical systems (HIC and ATPS), hydrophobicity of amino acids in proteins and conformational propensity of protein amino acids. The results obtained suggest that including conformational APVs improves scales obtained by decreasing the bias introduced by ASH. This suggests that a scale reflecting the probability of occurrence of each amino acid in different types of structures-configurations on the protein surface, and which incorporates the hydrophobic potential of each structure-configuration, may improve the predictive power of the models.

In conclusion, a significant increase in predictive power for HIC and ATPS models was achieved through use of clustering and optimisation algorithms. The increase obtained was, in fact, greater than that obtained with other models directly including more experimental information; accordingly, thereby reducing R&D expenditure. The results obtained raise new questions and suggest new ways to broaden and refine behavioural predictive models for protein separation/purification systems, which only require information on amino acid composition of the relevant proteins.

## Dedicatoria

*Algún día en cualquier parte, en cualquier lugar  
indefectiblemente te encontrarás a ti mismo, y ésa, sólo ésa,  
puede ser la más feliz o la más amarga de tus horas.*

**Pablo Neruda**

Vivir se podría reducir a la búsqueda constante de la felicidad, y las oportunidades y obstáculos que aparecen en el camino. Sobre cómo actuamos para conseguir la felicidad, aprovechamos las oportunidades y vencemos los obstáculos; inciden nuestro orgullo, prejuicios y escasez de visión. Esto muchas veces nos limita a no soñar suficientemente alto, no creer y desarrollar nuestras habilidades, y por último, a no buscar apoyo cuando lo necesitamos. Por lo tanto, limita nuestro éxito en la búsqueda de la felicidad.

Menciono esto porque durante una primera etapa de mi vida creo que no soñé lo suficiente, no creí suficiente en mis capacidades, y me conformé muchas veces con ser uno más.

Hoy, después de más de cuatro años desarrollando esta tesis, que fué un gran reto, y después de vencer muchos obstáculos simultáneos que me permiten escribir estas palabras desde Madrid, a raíz del Programa Para Formación de Directivos del Grupo Agbar en que participo, miro hacia atrás y veo que tengo mucho que agradecer.

Esta tesis me ha demandado mucho esfuerzo, constancia, y sacrificio. Es por esto que se la dedico a las tres personas que más influencia han tenido a lo largo de mi vida en mi éxito.

Primero, dedico esta tesis a mi difunta madre, María Carolina Humeres, que quizás de forma inocente y muy maternal, siempre me dijo que podría llegar a ser lo que yo quisiera en esta vida. También de forma inocente, yo le creí y aún le creo.

Luego, dedico esta tesis a mi padre, Juan Enrique Ugarte, quien sistemáticamente ha dado prioridad a sus hijos por sobre a si mismo, y cuyo apoyo incondicional y empuje durante la duración de esta tesis me ha sido de gran ayuda, sobre todo cuando más lo necesitaba. Tengo el orgullo de decir que cada día me parezco más a él.

Durante mis últimos años de colegio y primeros de universidad, hubo una persona que me motivó fuertemente a buscar mi propio camino, Patricia Santander, a quién también quisiera dedicarle esta tesis. Recorrer mi propio camino ha requerido un mayor esfuerzo, pero también me ha acercado un paso más a los valores que me definen el día de hoy. El obstáculo más difícil de todos es atreverse a soñar, y ser suficientemente persistentes para, en el día a día, hacer aquello que sabemos nos conduce hacia nuestras metas. Cumplir los sueños más ambiciosos difícilmente se debe a eventos fortuitos o esporádicos, sino a un esfuerzo constante, estrategia, y mucho valor.

## **Agradecimientos**

*Al final, lo que importa no son los años de vida, sino la vida de los años*

**Abraham Lincoln**

Quisiera comenzar agradeciendo al profesor Cristian Salgado, que fue un apoyo y empuje constante durante la realización de este trabajo, y con quien logramos encontrar la luz del conocimiento donde solo había oscuridad; y al profesor Álvaro Olivera, quien prestó una importante ayuda y consejo hacia el final de la tesis. Ambos fueron claves para el éxito y calidad de este trabajo.

También quisiera agradecer a mi comisión: Bárbara Andrews, Andrea Mahn, y M. Elena Lienqueo; por el esfuerzo, rapidez y diligencia; con que leyeron y me hicieron sus comentarios. Sin duda han realizado un aporte valioso a la calidad de esta tesis.

Agradezco a toda mi familia, mis amigos, compañeros, y novias; que a pesar del poco tiempo que tuve para dedicarles, siempre estuvieron ahí apoyando, preocupados, y son o fueron la vida de mis años.

Finalmente, agradezco a Juan E. Ugarte, Fernando Martínez, Cristian Fournies, Carolina Santander y Arturo Givovich; quienes, sin ser entendidos en el tema, de forma desinteresada y sólo por amistad, tuvieron la voluntad y el valor de leer mi tesis, y aportaron con correcciones y comentarios.

# Contenido

Resumen.....	i
Summary.....	ii
Dedicatoria.....	ii
Agradecimientos.....	iv
1 Introducción.....	1
2 Motivación.....	3
3 Antecedentes Generales.....	5
3.1 Cromatografía de Interacción Hidrofóbica (HIC).....	5
3.1.1 Principales Variables Fijas de Operación de HIC.....	6
3.1.2 Modelación de HIC.....	7
3.2 Sistemas de Dos Fases Acuosas (ATPS).....	9
3.2.1 Principales Variables Fijas de Operación.....	10
3.2.2 Modelación de ATPS.....	11
3.3 Métodos de Clustering.....	12
3.3.1 Etapas Involucradas en la Clasificación no Supervisada ( <i>Clustering</i> ).....	15
3.3.2 Características de los Algoritmos de Clustering.....	17
3.3.3 Clasificación de los Algoritmos de Clustering.....	21
3.3.4 Descripción de las Distintas Clases de Algoritmos.....	22
3.3.5 Comparación Entre Algoritmos de Clustering.....	28
3.4 Enfoque Escogido para la Predicción.....	29
3.5 Elección de los Algoritmos a Utilizar.....	30
4 Objetivos.....	32
4.1 Objetivos Generales.....	32
4.2 Objetivos Específicos.....	32
5 Hipótesis.....	32
6 Metodología.....	32
6.1 Modelos Utilizados.....	34
6.1.1 Predicción del Tiempo de Retención Adimensional en HIC.....	34
6.1.2 Predicción del Coeficiente de Partición en ATPS.....	37
6.2 Escalas de Propiedades Aminoacídicas.....	38
6.3 Evaluación de los Modelos.....	39
6.4 Extracción de Información Topológica Para Obtener Nuevos APVs.....	40
6.4.1 Un Problema de Optimización.....	40
6.4.2 <i>Growing Neuronal Gas</i> y <i>Growing Grid</i> .....	41
6.4.3 <i>Repeated Bisection</i> (gCluto).....	43
6.4.4 <i>Hierarchical Clustering</i> .....	44
6.4.5 <i>Genetic Algorithm</i> .....	47
6.4.6 <i>Restricted Neighborhood Search</i> (RNSC) y <i>Markov Clustering Algorithm</i> (MCL).....	49
6.5 Búsqueda de Escalas Precursoras.....	51
6.5.1 Metodología Base - Algoritmos de Redes Neuronales.....	52
6.5.2 Metodología Base - <i>Genetic Algorithm</i> .....	53
6.5.3 Síntesis de Listados de Escalas Precursoras.....	54
6.5.4 Estudio de la Ubicación Espacial de los Mejores Resultados.....	57
6.5.5 Determinación Final de Escalas Precursoras.....	58

7	Resultados y Discusión.....	58
7.1	Escalas Asociadas a Mejores Modelos Predictivos Para HIC y ATPS .....	59
7.1.1	Análisis de Disparidad en el Número de Escalas Obtenidas por cada Algoritmo.....	59
7.1.2	Tiempo Requerido Para Ajuste de Modelos Utilizando Distintos Algoritmos.....	61
7.1.3	Resultados Para la Cromatografía de Interacción Hidrofóbica (HIC).....	62
7.1.4	Resultados Para Sistemas de Dos Fases Acuosa (ATPS) .....	68
7.1.5	Análisis Comparativo Para los Modelos HIC y ATPS .....	76
7.2	Análisis de las Escalas Generadas que Permiten Obtener los Mejores Modelos .....	77
7.2.1	Análisis del Origen de las Mejores Escalas .....	78
7.2.2	Análisis Matemáticos y Bibliográficos de los APVs Más Cercanos.....	90
7.3	Análisis de la Propiedad de Hidrofobicidad a la Luz de los Resultados Obtenidos.....	96
7.3.1	El concepto de hidrofobicidad.....	96
7.3.2	Aproximación a la Hidrofobicidad a Partir de la Estructura de las Proteínas.....	97
7.3.3	Aproximación a la Hidrofobicidad a Partir de Procesos de Purificación de Proteínas.....	100
7.4	Representación de la Hidrofobicidad Mediante Los Modelos Desarrollados .....	104
7.5	Análisis de las Características de Sobre Ajuste de los Modelos.....	105
7.5.1	Modelos utilizados .....	105
7.5.2	Consecuencia del uso de la escala de propiedades aminoacídicas como variable 106	
7.5.3	Estudio del Comportamiento Hidrofóbico de las Proteínas.....	109
7.6	Información Aportada a los Modelos por los APVs Más Relevantes .....	111
8	Conclusiones .....	112
9	Glosario.....	115
10	Bibliografía.....	117
	Anexo A: definiciones.....	127
	Anexo B: ejemplos de ligandos en HIC .....	129
	Anexo C: gráfico efecto de una sal en capacidad en HIC .....	130
	Anexo D: series de Hofmeister [35, 36].....	131
	Anexo E: ejemplo de información topológica.....	132
	Anexo F: pseudocódigo GA.....	133
	Anexo G: pseudocódigo GNG .....	134
	Anexo H: pseudocódigo GG.....	135
	Anexo I: DRT proteínas .....	136
	Anexo J: coeficiente de partición de proteínas utilizadas .....	137
	Anexo K: listado APVs, clasificación y referencia.....	138
	Anexo L: ajuste visual GNG .....	141
	Anexo M: multiplicidad de resultados .....	143
	Anexo N: N° de Lisas APVs .....	148
	Anexo O: escalas mejores resultados ATPS.....	149
	Anexo P: parámetros iniciales y finales algoritmos .....	153
	Anexo Q: lista estudios mejores APVs .....	154
	Anexo R: resumen estudios mejores APVs.....	155

Anexo S: área superficial accesible de proteínas con DRT conocido .....	158
Anexo T: hidrofobicidad de las proteínas con DRT conocido al modificar la escala de Wertz y Scheraga [178] .....	161
Anexo U: tablas de resultados obtenidas a partir de un análisis de Anova .....	162
Anexo V: tablas de resultados obtenidas a partir de un análisis de Anova	163
Anexo Y: tablas de resultados obtenidas a partir de un análisis de Anova .....	164
Anexo Z: estudio del área Superficial accesible (ASA) de las proteínas con DRT conocido.....	167
Anexo AA: escala de Wertz y Scheraga.....	169



## Índice de Tablas

Tabla 1: Ejemplos de modelos propuestos en la literatura para ATPS. ....	12
Tabla 2: Definición matemática de las funciones criterio de similitud entre clusters. ....	24
Tabla 3: Algoritmos escogidos para el análisis de los 74 APVs. ....	31
Tabla 4: Nomenclatura utilizada para distinguir los 12 sistemas ATPS utilizados.....	38
Tabla 5: Conjunto inicial de parámetros para el algoritmo Growing Neuronal Gas. ....	41
Tabla 6: Conjunto inicial de parámetros para el algoritmo Growing Grid. ....	42
Tabla 7: Parámetros por defecto del algoritmo <i>Repeated Bisection</i> .....	43
Tabla 8: Parámetros del <i>Genetic Algorithm</i> con el valor predeterminado cambiado. ....	48
Tabla 9: Lista de parámetros del que depende el algoritmo MCL. ....	51
Tabla 10: Lista de parámetros del que depende el algoritmo RNSC.....	51
Tabla 11: Estructura de la información definida para sintetizar los 78 listados de escalas precursoras.....	55
Tabla 12: Cantidad de vectores generados por cada uno de los algoritmos utilizados. .	59
Tabla 13: Algoritmos de clustering y optimización y tiempos que se utilizó para generar nuevas escalas y ajustar los modelos para la el predicción del DRT y el coeficiente de partición K.....	62
Tabla 14: Atributos de los modelos de HIC obtenidos con APVs con base en la literatura.....	63
Tabla 15: Escalas utilizadas para obtención de mejores modelos con APVs con base en la literatura. ....	63
Tabla 16: Error de Jacknife de modelos con mejores resultados por algoritmo. ....	64
Tabla 17: Característica de los modelos con mejores niveles de ajuste de los DRT en HIC.....	65
Tabla 18: Coeficientes de los mejores modelos del DRT en HIC.....	66
Tabla 19: Mejores Escalas obtenidas con algoritmos que utilizan la topología de los APVs.....	67
Tabla 20: Mejores escalas obtenidas con el <i>Genetic Algorithm</i> . ....	67
Tabla 21: Valores base para la comparación de los resultados obtenidos con los APVs con base en la literatura.....	68
Tabla 22: Escalas con las cuales se construyeron los modelos tipo 3D utilizados como caso base. ....	69
Tabla 23: Escalas utilizadas en la elaboración de los modelos del tipo lineal, utilizados como caso base. Se incluye una breve descripción del tipo de escala al que corresponden.....	70
Tabla 24: Mejores resultados por algoritmo para cada uno de los 12 sistemas de dos fases acuosas, según los valores del Error de Jacknife ( $MSE_{jk}$ ). ....	71
Tabla 25: Mejores resultados por algoritmo para cada uno de los 12 sistemas de dos fases acuosas, como porcentajes de reducción del Error de Jacknife respecto a la línea base ( $\%MSE_{jk}$ ). ....	71

Tabla 26: Parámetro estadísticos asociados a los mejores modelos obtenidos con algoritmos de análisis topológico.....	73
Tabla 27: Indicadores del resultado para los mejores modelos obtenidos con algoritmos de análisis topológico.....	73
Tabla 28: Indicadores del resultado para los mejores modelos obtenidos con el <i>Genetic Algorithm</i> .....	74
Tabla 29: Indicadores del resultado para los mejores modelos obtenidos con el <i>Genetic Algorithm</i> .....	74
Tabla 30: Coeficientes de mejores modelos del coeficiente de partición obtenidos con los algoritmos de análisis topológico .....	75
Tabla 31: Coeficientes de mejores modelos del coeficiente de partición obtenidos con optimización genética. ....	75
Tabla 32: Desglose de los modelos generados por tipo.....	76
Tabla 33: APVs seleccionados para análisis teórico (ver referencias de cada APV en Anexo Q).....	90
Tabla 34: Variación del coeficiente del logaritmo entre modelos tipo 3D y lineales.. ....	93
Tabla 35: Características de las cromatografías realizadas por Lienqueo <i>et. al.</i> [54], Parker <i>et. al.</i> [144] y Meek [130]. ....	101
Tabla 36: Detalle componentes de la estructura del péptido de Parker [144] (Ac-Gly-X-X-(Leu) <sub>3</sub> -Lys <sub>2</sub> -Amine).....	102
Tabla 37: Características de los ATPS realizadas por Andrews <i>et. al.</i> [22] y Fauchère <i>et. al.</i> [143]. La sigla PEG corresponde a Polietilenglicol.....	104
Tabla 38: Tiempos de retención dimensional obtenidos por Lienqueo [54].....	136
Tabla 39: Coeficientes de partición obtenidos por Schmidt normalizados [22, 24].....	137
Tabla 40: Escalas de propiedades aminoacídicas utilizadas (APVs) con su numeración y clasificación.. ....	138
Tabla 41: Escalas generadas a partir de algoritmos de análisis topológico, utilizadas para la construcción de los mejores modelos tipo 3D para los doce sistemas ATPS .....	149
Tabla 42: Escalas generadas a partir de algoritmos de análisis topológico, utilizadas para la construcción de los mejores modelos tipo lineal para los doce sistemas ATPS .....	150
Tabla 43: Escalas generadas a partir de algoritmos de optimización genética, utilizadas para la construcción de los mejores modelos tipo 3D para los doce sistemas ATPS .....	151
Tabla 44: Escalas generadas a partir de algoritmos de optimización genética, utilizadas para la construcción de los mejores modelos tipo lineal par los doce sistemas ATPS .....	152
Tabla 45: Listado de valores de los parámetros iniciales y finales utilizados con los algoritmos <i>Growing Neuronal Gas</i> .....	153
Tabla 46: Listado de valores de los parámetros iniciales y finales utilizados con los algoritmos <i>Growing Grid</i> . ....	153

Tabla 47: Listado de valores de los parámetros iniciales y finales utilizados con los algoritmos <i>Bisection Algorithm</i> .....	153
Tabla 48: Identificación de <i>trabajos asociados a</i> escalas más relevantes para la generación de los mejores modelos. ....	154
Tabla 49: Resumen de trabajos asociados a escalas más relevantes para la generación de los mejores modelos. ....	155
Tabla 50: Área superficial accesible (ASA) de los aminoácidos hidrofóbicos y neutros (GLY), según la clasificación de Rose et. al. [145], para las proteínas con DRT conocido (ver anexo I). ....	158
Tabla 51: Área superficial accesible (ASA) de los aminoácidos medianamente polares, según la clasificación de Rose et. al. [145], para las proteínas con DRT conocido (ver anexo I).....	159
Tabla 52: Área superficial accesible (ASA) de los aminoácidos polares, según la clasificación de Rose et. al. [145], para las proteínas con DRT conocido (ver anexo I). ....	160
Tabla 53: Hidrofobicidad de las proteínas utilizadas para HIC (ver proteínas en Anexo I), obtenidas a partir de las escalas generadas al variar el valor de hidrofobicidad de un aminoácido sobre la escala de Wertz y Scheraga [178].. ....	161
Tabla 54: Tabla resultado del análisis de varianza (ANOVA) de la variación de la hidrofobicidad de las proteínas utilizadas para HIC, obtenida a partir de las escalas generadas al variar el valor de hidrofobicidad de un aminoácido sobre la escala de Wertz y Scheraga [178] (se utiliza la matriz traspuesta de la Tabla 54). ....	162
Tabla 55: Tabla resultado del análisis de varianza (ANOVA) de la variación de la hidrofobicidad de las proteínas utilizadas para HIC, obtenida a partir de las escalas generadas al variar el valor de hidrofobicidad de un aminoácido sobre la escala de Wertz y Scheraga [178] (ver Tabla 54). ....	162
Tabla 56: Tabla resultado del análisis de varianza (ANOVA) de la variación de la hidrofobicidad de las proteínas utilizadas para HIC [ref], obtenida a partir de las escalas generadas al variar el valor de hidrofobicidad de un aminoácido sobre la escala de Wertz y Scheraga [178] (ver Tabla 54).....	162
Tabla 57: Variación de la hidrofobicidad de las proteínas utilizadas para HIC [ref], obtenida a partir de las escalas generadas al variar el valor de hidrofobicidad de un aminoácido sobre la escala de Wertz y Scheraga [178]. ....	163
Tabla 58: Porcentaje del área superficial accesible según la clasificación propuesta por Rose et. al. [145], e incluyendo el aminoácido Gly como un aminoácido neutro. ...	167
Tabla 59: Se muestran los porcentaje de ASA de cada proteína, según las clasificaciones: hidrofóbico-neutro y medianamente polar-polar. ....	168
Tabla 60: Escala de hidrofobicidad de Wertz y Scheraga [178]. ....	169

## Índice de Figuras

Figura 1: Diagrama de un sistema cromatográfico.....	5
Figura 2: Esquema de dos fases acuosas .....	9
Figura 3: Representación de datos en un plano y agrupación mediante un proceso intuitivo de <i>clustering</i> . .....	14
Figura 4: Diagrama del proceso de <i>clustering</i> . .....	15
Figura 5: Esquema de algoritmos de clustering aglomerativos y de partición. ....	17
Figura 6: Dendrograma. Una forma común de mostrar los resultados de un algoritmo jerárquico. ....	19
Figura 7: Esquema de un grafo del tipo <i>scale-free</i> . ....	20
Figura 8: Ejemplo de cluster globulares en el plano. ....	20
Figura 9: Ejemplo de clusters transitivos .....	21
Figura 10: Clasificación propuesta para los algoritmos de <i>clustering</i> . ....	21
Figura 11: Ejemplos de <i>clusters</i> elongados (A) y compactos (B). ....	23
Figura 12: Ejemplo de <i>clusters</i> difíciles de distinguir. ....	23
Figura 13: Esquema del proceso de <i>crossover</i> . ....	26
Figura 14: Estructura de la función de una neurona en una red neuronal.....	27
Figura 15: Esquema de una red neuronal. Se aprecia la existencia de una capa input, una output y capas ocultas ( <i>hidden</i> ). ....	28
Figura 16: Esquema general de la metodología para la construcción de modelos. ....	34
Figura 17: Esquema de análisis de sensibilidad para el ajuste de parámetros. ....	42
Figura 18: Esquema dendrograma e identificación de posibles <i>clusters</i> con iguales y distintos niveles de similitud o jerarquía.....	46
Figura 19: Interfaz gráfica del <i>Markov Clustering Algorithm</i> .....	50
Figura 20: Esquema del proceso de inteligencia animal simplificado en una relación entre un conjunto de señales, una o más capas intermedias ( <i>hidden</i> ) y la capa que entrega el resultado o acción (denotada como salida). ....	52
Figura 21: Esquema de metodologías utilizadas para la obtención de las escalas precursoras para los mejores resultados obtenidos con algoritmos de análisis topológico. ....	53
Figura 22: Esquema de metodologías utilizadas para la obtención de las escalas precursoras para los mejores resultados obtenidos con el <i>Genetic Algorithm</i> .....	54
Figura 23: Esquema metodología de síntesis de listados de escalas precursoras (APVS). En la lista de los APVs por frecuencia se tiene que: $f_i > f_{i+1}$ . ....	56
Figura 24: Esquema metodología de síntesis de listados de escalas precursoras (APVS). En la lista de los APVs por frecuencia se tiene que: $f_i > f_{i+1}$ . ....	56
Figura 25: Esquema de la generación de gráficos para evaluar la composición de un..	58
Figura 26: Síntesis de los APVs más influyentes y cercanos a las mejores soluciones.	80

Figura 27: Composición global de APVs para ATPS y HIC. En azul se representa el porcentaje de APVs hidrofóbicos, en rojo los de tipo conformacional y en verde los del tipo estadístico. ....	81
Figura 28: Composición de APVs más influyentes y cercanos de las mejores soluciones encontradas con los algoritmos de redes neuronales.....	83
Figura 29: Composición de APVs más influyentes y cercanos de las mejores soluciones encontradas con el <i>Genetic Algorithm</i> . ....	84
Figura 30: Ejemplo de distribución en el plano de nuevas escalas (puntos negros), y APVs hidrofóbicos (celeste), conformacionales (rojo), y estadísticos (verdes).....	86
Figura 31: Esquema de concordancia entre el gráfico n°4 de la Figura 28 y el ejemplo de la Figura 30, de escalas y APVs en el espacio. ....	87
Figura 32: Esquema de concordancia entre el gráfico n°2 de la Figura 28 y el ejemplo de la Figura 30, de escalas y APVs en el espacio. ....	88
Figura 33: Listas de APVs más relevantes (influyentes y cercanos) para mejores escalas generadas con algoritmos de redes neuronales. ....	90
Figura 34: Esquema de contenido y orden esperado de las listas de APVs de la Figura 33.....	91
Figura 35: Evaluación de Concordancia Tipo I y Concordancia Tipo II. ....	92
Figura 36: Composición de APVs más influyentes y cercanos de las mejores soluciones encontradas con algoritmos de análisis topológico (redes neuronales).....	95
Figura 37: Ejemplos de distintos tipos de ligandos utilizados en una Cromatografía de Interacción Hidrofóbica [29]. ....	102
Figura 38: Esquema del efecto de un cambio de escala de hidrofobicidad sobre la ubicación espacial de una proteína en la gráfica del DRT en función de la hidrofobicidad.. ....	106
Figura 39: DRT en función de la hidrofobicidad de las 12 proteínas utilizadas para la generación de los nuevos modelos.. ....	107
Figura 40: DRT en función de la hidrofobicidad de las 12 proteínas utilizadas para la generación de los nuevos modelos.. ....	108
Figura 41: DRT en función de la hidrofobicidad de las 12 proteínas utilizadas para la generación de los nuevos modelos.. ....	109
Figura 42: Esquema de relación entre la variación de un componente de una escala de hidrofobicidad, y la hidrofobicidad de dos proteínas. ....	111
Figura 43: Distintos tipos de ligando en una cromatografía de interacción hidrofóbica [29].....	129
Figura 44: Gráfico de la relación entre la capacidad de una matriz en función de la concentración la sal sulfato de amonio, $(\text{NH}_4)_2\text{SO}_4$ , para $\alpha$ -chymotrypsinogen y RNase [29]. ....	130
Figura 45: Diagrama de las series de Hofmeister – <i>efecto salting out</i> y <i>salting in</i> .....	131
Figura 46: Diagrama de las series de Hofmeister – contribución de sales a la tensión superficial.....	131
Figura 47: Diagrama del Metro de Madrid: Ejemplo de Información Topológica [199].	132

Figura 48: Diagrama pseudo código algoritmos evolutivos.....	133
Figura 49: Solución obtenida con el algoritmo <i>Growing Grid</i> al analizar un conjunto de datos que representan de forma discreta una circunferencia.....	141
Figura 50: Representación de una solución obtenida por el algoritmo <i>Growing Grid</i> , mostrando los datos iniciales en azul, las neuronas en rojo y las conexiones de la grilla de neuronas con líneas negras.....	142
Figura 51: Convergencia de la varianza con el aumento del tamaño de la muestra o ejecuciones del algoritmo GNG para un set determinado de parámetros.....	146
Figura 52: Convergencia de la varianza con el aumento del tamaño de la muestra o ejecuciones del algoritmo GNG para un set determinado de parámetros.....	147
Figura 53: Variación de la hidrofobicidad de las proteínas utilizadas para generar los modelos del DRT en HIC, obtenida al modificar la hidrofobicidad de cada uno de los aminoácidos hidrofóbicos y neutros (según la clasificación de Rose et. al. [145]) por separado en un valor fijo e igual a la media de la escala original sobre la escala de Wertz y Scheraga [178]. .....	164
Figura 54: Variación de la hidrofobicidad de las proteínas utilizadas para generar los modelos del DRT en HIC, obtenida al modificar la hidrofobicidad de cada uno de los aminoácidos medianamente polares (según la clasificación de Rose et. al. [145]) por separado en un valor fijo e igual a la media de la escala original sobre la escala de Wertz y Scheraga [178]. .....	164
Figura 55: Variación de la hidrofobicidad de las proteínas utilizadas para generar los modelos del DRT en HIC, obtenida al modificar la hidrofobicidad de cada uno de los aminoácidos polares (según la clasificación de Rose et. al. [145]) por separado en un valor fijo e igual a la media de la escala original sobre la escala de Wertz y Scheraga [178]. .....	165
Figura 56: Gráfico del DRT-Hidrofobicidad de las proteínas utilizadas para generar los modelos del DRT en HIC, obtenida al modificar la hidrofobicidad de cada uno de los aminoácidos hidrofóbicos y neutros (según la clasificación de Rose et. al. [145]) por separado en un valor fijo e igual a la media de la escala original sobre la escala de Wertz y Scheraga [178]. .....	165
Figura 57: Gráfico del DRT-Hidrofobicidad de las proteínas utilizadas para generar los modelos del DRT en HIC, obtenida al modificar la hidrofobicidad de cada uno de los aminoácidos medianamente polares (según la clasificación de Rose et. al. [145]) por separado en un valor fijo e igual a la media de la escala original sobre la escala de Wertz y Scheraga [178]. .....	166
Figura 58: Gráfico del DRT-Hidrofobicidad de las proteínas utilizadas para generar los modelos del DRT en HIC, obtenida al modificar la hidrofobicidad de cada uno de los aminoácidos polares (según la clasificación de Rose et. al. [145]) por separado en un valor fijo e igual a la media de la escala original sobre la escala de Wertz y Scheraga [178]. .....	166

# 1 Introducción

El principal coste en la industria biotecnológica se produce en I+D, alcanzando un 53% de los ingresos en USA y 63% en Europa (1997-1999) [1]. Esto se explica por la complejidad de las técnicas utilizadas, como en algunos procesos de separación y purificación de proteínas.

En la investigación y la industria, los procesos de separación y purificación de proteínas son indispensables para la generación de productos libres de contaminantes. La disponibilidad de procesos eficientes y de fácil escalamiento es fundamental para la producción industrial de estas moléculas.

Dentro de las limitantes para el uso exitoso y masivo de una técnica de separación en el mercado, se puede mencionar la comprensión del proceso de separación y una fácil predicción de los valores óptimos de las variables de proceso a partir de la información disponible. Esta capacidad de predicción permite que se reduzca el tiempo y costos asociados a ajustar un proceso de separación específico, cuando no se cuenta con información previa sobre el comportamiento de las moléculas que se quieren separar en el sistema de separación a utilizar.

Dos técnicas de separación y purificación de proteínas son la cromatografía de interacción hidrofóbica (HIC) y los sistemas de dos fases acuosas (ATPS). En la primera una mezcla de proteínas se introduce por una columna con una matriz hidrofóbica. Los ligandos no polares de la matriz interaccionan con las zonas no polares de la estructura tridimensional de las proteínas, de esta manera, el grado de interacción dependerá de la cantidad de zonas hidrofóbicas y cómo éstas están distribuidas en la superficie de cada proteína [2, 3]. Por otra parte, un sistema ATPS consiste en dos fases basadas en agua y algún compuesto orgánico o sal (por ejemplo polietilenglicol y dextrano), en las cuales las proteínas se transfieren preferentemente hacia una de las dos fases. En el primer caso el mecanismo se ha estudiado extensamente y existen muchos modelos con buena capacidad predictiva, sin embargo en ATPS el tipo de mecanismo bajo el cual se produce la separación no está completamente identificado.

Ambas técnicas son utilizadas en procesos río abajo para la separación de proteínas en industrias químicas y biotecnológicas. Además, ambas técnicas tienen aplicaciones a escala industrial y de laboratorio.

Los sistemas HIC dependen de las siguientes variables: tipo y densidad de los ligandos en la matriz, tamaño e hidrofobicidad de las proteínas, y concentración y tipo de sal empleada [4]. La formación de las dos fases en ATPS depende del pH, temperatura, fuerza iónica, concentración de los compuestos orgánicos (o sales), y peso molecular [5]. El coeficiente de partición de proteínas en ATPS depende de su tamaño, carga neta e hidrofobicidad [5].

El número de variables y la complejidad de los fenómenos fisicoquímicos que ocurren en estos procesos de separación no son despreciables. Esto explica la necesidad de modelos que permitan una reducción de tiempo y costos asociados a la puesta en marcha y afinamiento de las técnicas con un propósito puntual.

Algunos modelos que se han utilizado para predecir el coeficiente de partición en HIC son: el modelo del Área Hidrofóbica de Contacto (HCA) [3] y el modelo de Docking Molecular Local por Hidrofobicidad [6]; sin embargo, una debilidad de estos modelos es: el primero requiere ser calculado de forma experimental para cada proteína, y el segundo requiere una simulación para cada proteína y tipo de ligando a utilizar.

Por otro lado, algunos modelos utilizados para predecir el coeficiente de partición en ATPS son: el modelo en base a diferencia de potencial eléctrico de Albertsson [7] y el modelo de Eiteman *et. al.* [8], que utiliza la hidrofobicidad de la proteína y la diferencia de concentración de un compuesto específico entre las fases para realizar la predicción. Una debilidad de estos modelos es: el primero se basa únicamente en la diferencia de potencial eléctrico entre las fases, y el segundo requiere demasiada información (diferencia de concentración del componente *i* entre las fases) considerando que es necesario ajustar el modelo para cada sistema.

Tanto el tiempo de retención (DRT) para HIC como el coeficiente de partición (K) para ATPS se han logrado estimar de manera aceptable con el modelo de la hidrofobicidad superficial media (ASH) de las proteínas [4, 5], lo cual presenta una ventaja sobre los modelos mencionados con anterioridad. La ASH se obtiene considerando la contribución relativa de cada aminoácido que está presente en la superficie [9]. Para el cálculo de la ASH de una proteína se requiere conocer su estructura tridimensional, y la hidrofobicidad de cada aminoácido [5, 10]. La estructura tridimensional de una proteína se puede obtener mediante metodologías experimentales y predictivas [11], pero ninguna de las dos son triviales, involucran altos costos y/o horas hombre. Hoy en día esta información solamente está disponible para 57.298 proteínas (2010), es decir, en muchos casos no está disponible [11, 12]. Por otro lado, no existe una escala estándar de hidrofobicidad, por el contrario, existen muchas escalas tanto de origen teórico como experimental [9].

En esta misma línea, Salgado *et. al.* [13] ha desarrollado una metodología que permite la predicción de la ASH utilizando únicamente la composición aminoacídica de las proteínas. La aplicación de esta ASH ha permitido en algunos casos mejorar la predicción del DRT obtenido por medio de metodologías que utilizan la estructura tridimensional de las proteínas [13]. La metodología de Salgado *et. al.* también ha demostrado servir en la predicción del coeficiente de partición en ATPS, aunque la calidad de la predicción depende de cada ATPS y las condiciones de operación [4].

La segunda limitante para el cálculo de la ASH se mantiene presente en la metodología de Salgado, esto es, la dependencia de la escala de hidrofobicidad escogida (o en general, la escala de propiedad aminoacídica - APV), o de las escalas generadas a partir de las conocidas con base en la literatura.

Para generar nuevas escalas, Salgado *et. al.* utilizó dos algoritmos de clasificación no supervisada (*K-mean* y *SOM*) sobre un conjunto de 74 APV [13, 14]. Estos algoritmos permiten analizar el conjunto de APV para generar grupos relacionados en base a la topología inherente, o para generar directamente nuevas escalas. Luego, las nuevas escalas son una "mezcla" de las originales en un sentido multidimensional. Con las escalas generadas se logró mejorar la predicción del DRT en HIC con respecto a las 74 escalas originales.



Por otro lado, las técnicas para la búsqueda de *clusters* de vectores son muy variadas, así como los tipos de resultados que se obtiene a partir de éstos [15]. Algunas de estas técnicas incluso trascienden la búsqueda de *clusters*, y generan directamente nuevas escalas manteniendo las propiedades topológicas más relevantes de las escalas originales [16]. En este último caso la metodología utilizada es una simplificación del procesamiento de información que realiza el cerebro (inteligencia artificial).

Dado la falta de un criterio válido para la identificación de la técnica idónea, es necesario escoger un conjunto de técnicas de *clustering* lo más exhaustiva posible, que represente el estado del arte en el área de la clasificación no supervisada.

El objetivo de esta tesis es realizar un análisis de *clustering* robusto que permita seleccionar grupos de APV y generar nuevas escalas, a través de las cuales sea posible mejorar el poder predictivo de los modelos utilizados por Salgado *et. al.* en HIC [9] y doce ATPS [11].

Los modelos a obtener tienen el potencial de reducir tiempo y costos asociados a la puesta en marcha y ajuste de HIC y los doce ATPS estudiados, tanto a nivel de laboratorio como a nivel industrial. Además, la metodología utilizada puede ser replicada para otros sistemas de separación del mismo tipo.

Finalmente, el presente estudio tiene el potencial de determinar cómo la información provista por los algoritmos de *clustering* es capaz de describir los distintos aspectos en la complejidad de las interacciones hidrofobicidad, complementando la información particular asociada a cada sistema experimental.

## 2 Motivación

El gran interés en el uso de proteínas es consecuencia de la diversidad de propiedades útiles, y la eficiencia en la función que desempeñan, en general, muy superior a otro tipo de moléculas. Algunos ejemplos son:

- Su capacidad catalítica en reacciones a muy baja concentración (enzimas). Esto permite realizar reacciones en condiciones que normalmente no es posible, disminuyendo los costos de operación en muchos procesos. Por ejemplo, disminuyendo la temperatura requerida para alcanzar la energía de activación de una reacción.
- Su elasticidad, como del de la Resilina, tiene un gran potencial de uso en telas, zapatos para atletas, medicina, entre otras muchas aplicaciones. Las moléculas elásticas artificiales (no proteicas) no logran la misma eficiencia o durabilidad [17, 18].
- Su adhesión, como el de la caseína (presente en la leche), que ha sido utilizada para pegar madera desde 1800. En la actualidad la caseína también es utilizada en la remineralización de los dientes [19].
- Sus múltiples funciones en el cuerpo, que a partir del cual poseen el potencial de ser utilizados en aplicaciones terapéuticas, o como parte de medicamentos. Por ejemplo, pueden participar como señales específicas que permiten al cuerpo el transporte de fármacos a lugares específicos de éste. Otro ejemplo es la ingesta de lactasa artificial, que ayuda a la digestión de alimentos que contienen lactosa, en personas que no la toleran.

- La capacidad anticongelante de algunas proteínas, que en vez de actuar modificando las propiedades coligativas del sistema, como otros químicos, inhiben la formación de cristales, lo que se traduce en concentraciones entre 300 y 500 veces menores [20] [21].
- El poder nutricional que aportan es vital para los animales. Las proteínas no solo se utilizan por el cuerpo para la generación de energía (proceso en que predomina el uso de los Glúcidos y Lípidos), sino que también se utilizan para formar las estructuras que permiten al cuerpo y las células mantener un adecuado funcionamiento. Esto se debe principalmente a que, a diferencia de las bacterias y hongos, los animales no son capaces de sintetizar todos los aminoácidos que necesitan, por lo que éstos se deben obtener, por ejemplo, a través de la ingesta en complementos alimenticios en cápsulas.

Por lo tanto, las propiedades de las proteínas que presentan utilidad para el hombre son muy variadas, y tienen aplicaciones prácticas en la industria y en la investigación. En la medida que se descubran cuáles son las proteínas responsables de efectos observables en la naturaleza, y que son de gran interés, es posible lograr avances y aplicaciones tecnológicas relevantes. Sin embargo, un limitante para el estudio de las proteínas y su aplicación industrial es el grado de pureza requerido.

Distintos tipos de aplicaciones pueden requerir distintos niveles de pureza. En general la pureza de proteínas requeridas en la investigación puede llegar a niveles mayores del 99,0%. Todo esto ha llevado al desarrollo de diversas técnicas de separación y purificación de proteínas. Por otro lado, el uso a nivel industrial introduce nuevas restricciones, como es la aplicabilidad a gran escala, manteniendo los costos a niveles que permitan la competitividad de los productos elaborados.

El punto de partida para el diseño de un proceso de separación o purificación es la identificación de las propiedades a partir de las cuales se produce la separación, como el tamaño, carga, hidrofobicidad, y actividad física, entre otras. Algunas técnicas desarrolladas son: la ultra centrifugación, la precipitación, la electroforesis, la cromatografía y los sistemas de dos fases acuosas.

Dentro de los factores que pueden tener incidencia en la aplicabilidad de una técnica, se encuentra el conocimiento sobre las leyes físico-químicas que la gobiernan. En técnicas como la cromatografía y los sistemas de dos fases acuosas, el trabajo experimental requerido para ajustar un protocolo para una separación específica es muy laborioso, razón por la cual los modelos de predicción del comportamiento de proteínas son una muy buena ayuda, ahorrando tiempo y costos. Entre más sencillo el modelo y más básica la información requerida, más ayuda puede presentar en etapas tempranas de la investigación.

A la fecha se ha realizado esfuerzos por desarrollar modelos sencillos con poder predictivo razonable para la aplicación puntual [4, 5, 9, 13, 22-24]. Sin embargo, estos modelos tienen debilidades que presentan una oportunidad para generar nuevos modelos, o nuevas metodologías para construir dichos modelos, con un mayor poder predictivo, punto de partida para el presente trabajo.

### 3 Antecedentes Generales

#### 3.1 Cromatografía de Interacción Hidrofóbica (HIC)

Una cromatografía consiste en una columna con una matriz, por la cual se hace pasar una solución con distintos compuestos o moléculas. En la Figura 1 se puede ver una columna donde se identifica la fase móvil y la estacionaria. La fase móvil se carga por la parte superior, y la separación de las moléculas se produce por la interacción diferenciada de cada una de ellas con la matriz, de manera que eluyen a distintos tiempos.

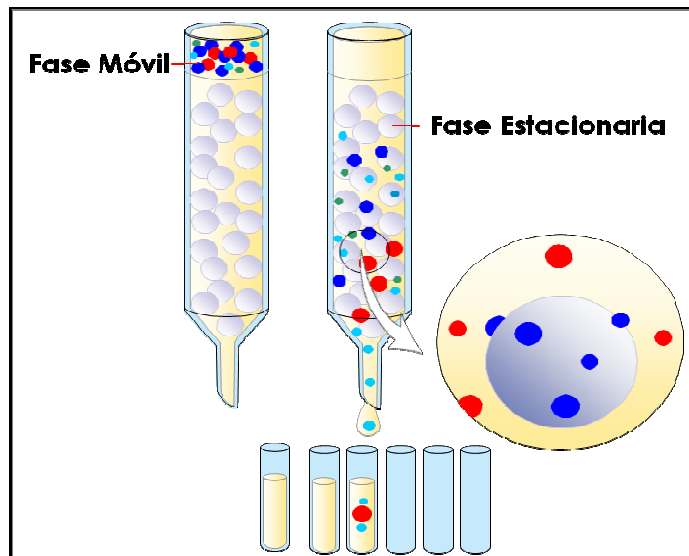


Figura 1: Diagrama de un sistema cromatográfico.

La matriz se escoge para explotar distintas propiedades de las proteínas, como: la carga iónica, la hidrofobicidad, el tamaño, la afinidad específica a algún ligando, entre otras.

Por ejemplo, si se desea separar por tamaño, la matriz normalmente está compuesta de esferas que poseen canales interiores, de manera que las moléculas de tamaño más pequeño pueden atravesar estos canales y las moléculas de mayor tamaño utilizan los espacios entre esferas. Los canales interiores generan un trayecto más largo, por lo que las moléculas de mayor tamaño eluyen por el final de la columna antes, y las de menor tamaño eluyen al último, lo que permite que sean recolectadas en distintos recipientes.

En la cromatografía de interacción hidrofóbica la separación se produce por una interacción diferencial con una matriz hidrofóbica. El término cromatografía de interacción hidrofóbica (*hydrophobic Interaction chromatography*) fue inicialmente propuesto por Hostee [25] y Shaltiel *et. al.* [26] bajo el supuesto que el modo de interacción entre proteínas y los ligandos hidrofóbicos inmovilizados es similar a la asociación que ocurre entre pequeñas moléculas alifáticas en agua. Por otro lado, Porath *et. al.* [27] sugieren que este efecto de interacción es incrementado por la presencia de sal. Él indica que la fuerza motriz es la entropía obtenida a partir de cambios estructurales en el agua alrededor de los grupos hidrofóbicos interactuantes.

Este concepto luego fue extendido y formalizado por Hjertén [28], quien basó su teoría en la relación termodinámica de Gibbs:

$$\Delta G = \Delta H - T\Delta S \quad (1)$$

Él propuso que el desplazamiento de moléculas de agua, ordenadas en torno a las proteínas y los ligandos, producen un aumento de la entropía ( $\Delta S$ ) y una disminución de la energía libre  $\Delta G$ , lo que implica que es termodinámicamente favorable. Otras teorías han ayudado a la comprensión de la relación entre la concentración de sal y la hidrofobicidad.

Los parámetros de una cromatografía de interacción hidrofóbica que son relevantes en la optimización de la separación, son: el tipo y densidad del ligando, el tipo de matriz base, el tipo y concentración de sal, el pH, la temperatura y los aditivos [29].

### **3.1.1 Principales Variables Fijas de Operación de HIC**

#### ***Condiciones de la Fase Estacionaria***

##### *Ligando*

La dependencia de una cromatografía con el tipo de ligando se mide en base a la capacidad de la matriz de retener proteínas en unidades de [mg proteína/ml gel]. Esta medida depende tanto del tipo de ligando como de la densidad o grado de sustitución del ligando ([ $\mu\text{mol}$  de ligando/ml gel]) en el gel. En ambos casos, siempre que el tipo de ligando sea lineal y varíe solo en el tamaño de la cadena alifática, la dependencia es logarítmica, aunque la fuerza de atracción sigue aumentando en función del grado de sustitución [29-33], aún cuando no aumenta la capacidad de la matriz (ver Anexo B).

##### *Tipo de Matriz Base*

Los dos tipos de matriz más utilizadas son las basadas en agarosa o un copolímero sintético. Normalmente en ambos casos las matrices son fuertemente hidrofílicas, y la selectividad que se obtiene con una matriz puede variar respecto a otra, siendo necesario ajustar algunas de las condiciones de elución [29].

#### ***Condiciones de la Fase Móvil***

##### *Tipo y Concentración de Sal*

El aumento de la concentración de sal, o de compuestos que forman sales, favorece la interacción entre los ligandos y las proteínas [27, 34-36]. Esta relación es lineal hasta una concentración específica, punto en que la relación se vuelve exponencial (ver Anexo C).

Respecto al tipo de sal, se ha encontrado que algunas contribuyen más que otras a la precipitación de las proteínas a una misma concentración (*salting out effect*). A partir de este fenómeno se generaron las series de Hofmeister [35, 36], que es una clasificación de las sales (ver Anexo D).

### *Efecto del pH*

En general, se ha encontrado que en la separación de proteínas un aumento del pH se traduce en una disminución aparente de la hidrofobicidad, producto del aumento de grupos cargados. Por otro lado, una disminución del pH resulta en un aumento aparente de la hidrofobicidad [37].

La forma de la relación anterior es compleja y se ha detectado que, en general, en el rango de pH 5-8,5 el efecto es menor que fuera de dicho rango [37].

### *Efecto de la Temperatura*

El efecto de la temperatura en la hidrofobicidad en HIC es compleja, y se ha reportado evidencias que indican que aumentos en la temperatura producen aumentos en la hidrofobicidad [28, 38] y en las fuerzas de van der Waals [39]; sin embargo, también se ha reportado el efecto contrario [40]. Esto se cree puede depender de la relación de la estructura y la solubilidad de las proteínas con la temperatura [29].

### **3.1.2 Modelación de HIC**

Para representar un sistema HIC se puede distinguir dos tipos de modelos: los que utilizan propiedades de las proteínas para determinar un tiempo de retención de éstas en HIC; y los modelos que utilizan los parámetros de HIC para determinar el efecto que producen sobre el sistema y la hidrofobicidad.

Dentro de los parámetros uno de los más estudiados es la concentración de sal. Para explicar el efecto de una sal en un sistema HIC se ha propuesto dos tipos de teorías, las solvofóbicas (*Solvophobic Theory*) y las de interacción preferencial (*Preferential Interaction Theory*) [41].

La teoría solvofóbica se basa en la asociación y solvatación de las moléculas participantes [35]. Este modelo asume que la variación en el tiempo de retención producto de la concentración de la sal, es proporcional a la tensión superficial molal generada por el incremento de la concentración de sal. Sin embargo, esta teoría no es válida para sales que interactúan fuertemente con las proteínas.

La teoría de interacción preferencial se basa en la interacción entre una sal y las proteínas [42-50], y es válida en un amplio rango de concentración de sal [51].

La principal propiedad físico-química de las proteínas que determina su comportamiento en HIC es la hidrofobicidad. Aunque no existe acuerdo universal sobre una única medida para la hidrofobicidad de las proteínas, si hay consenso respecto a que ella está determinada por la contribución hidrofóbica de los residuos de sus aminoácidos [52, 53]. Existen tres métodos prometedores para la predicción de la interacción entre una proteína y una resina, basados en la hidrofobicidad. Estos son:

#### ***Área Hidrofóbica Superficial, $\phi_{surface}$***

En esta metodología se utiliza la hidrofobicidad superficial media, bajo el supuesto que cada aminoácido en la superficie tiene una contribución a la hidrofobicidad proporcional a su área accesible por el solvente [6].

$$\phi_{\text{surface}} = \frac{\sum s_{\text{aai}} \phi_{\text{aai}}}{s_p} \quad (2)$$

Donde  $\phi_{\text{surface}}$  es la hidrofobicidad superficial media de la proteína,  $i$  ( $i = 1, \dots, 20$ ) representa a cada uno de los aminoácidos estándar,  $s_{\text{aai}}$  es el área superficial accesible al solvente del aminoácido  $i$ ,  $\phi_{\text{aai}}$  es el valor de hidrofobicidad asignado al aminoácido  $i$  por una escala de hidrofobicidad, y  $s_p$  es el área total accesible de la proteína accesible al solvente. La escala de hidrofobicidad se debe escalar de manera que  $\phi_{\text{aai}} \in [0,1], \forall i$ . Finalmente, se utiliza un modelo cuadrático simple para estimar el tiempo de retención adimensional (DRT) [54].

$$\text{DRT} = b_0 + b_1 \phi_{\text{surface}} + b_2 \phi_{\text{surface}}^2 \quad (3)$$

### **Área Hidrofóbica de Contacto (HCA)**

El área hidrofóbica de contacto es el área de una proteína que está en contacto con los ligandos de la matriz de HIC. El HCA se calcula experimentalmente mediante la expresión:

$$\text{Log } k' = A + C m_s \quad (4)$$

Donde  $k'$  es el factor de retención isocrático,  $m_s$  es la concentración molal de sal, y  $A$  y  $C$  son constantes del modelo. Luego, el HCA se calcula mediante la siguiente expresión:

$$C = \frac{HCA \sigma_s}{2,3RT} \quad (5)$$

Donde HCA es el área de contacto entre una proteína y los ligandos de la matriz del sistema HIC,  $\sigma_s$  es una propiedad de la sal medida como el incremento en la tensión superficial debido a la adición de una sal neutra.  $R$  es la constante universal de los gases y  $T$  es la temperatura absoluta [3].

El cálculo del tiempo de retención adimensional se realiza con la siguiente expresión [3]:

$$\text{DRT} = b_0 + b_1(\text{HCA}) \quad (6)$$

### **Docking Molecular Local por Hidrofobicidad**

Esta metodología se basa en la identificación de las zonas en la superficie de una proteína con mayor probabilidad de interacción con la matriz hidrofóbica del sistema HIC (área de acoplamiento). A partir del conocimiento de la estructura tridimensional de proteínas se determina la hidrofobicidad local (LH), la cual se ha encontrado que tiene una alta correlación con el área hidrofóbica accesible y el tiempo de retención adimensional (DRT) de distintas proteínas [6].

Para determinar las áreas de acoplamiento se requiere una simulación basada en métodos como el *Genetic Algorithm* o algoritmos tipo Montecarlo, y una función para evaluar el ajuste, como la función de energía libre.

Una vez que los complejos (proteína-ligando) más probables se encuentran, la zona de interacción se define como un radio de 5 Å desde el centro hasta la zona identificada.

Luego se determina el área accesible al solvente, del área interactuante ( $s_{IZ}$ ); y el área parcial accesible correspondiente a cada residuo ( $s_{aai}$ ). Finalmente, mediante la siguiente ecuación se obtiene el área hidrofóbica local (LH).

$$LH = \frac{\sum s_{aai} \phi_{aai}}{s_{IZ}} \quad (7)$$

Donde  $\phi_{aai}$  corresponde a los valores de una escala de hidrofobicidad [6]. La correlación entre el LH y el tiempo de retención se puede obtener mediante la siguiente expresión [54].

$$DRT = b_0 + b_1 LH + b_2 LH^2 \quad (8)$$

### 3.2 Sistemas de Dos Fases Acuosa (ATPS)

Los sistemas de dos fases acuosas son básicamente una solución acuosa donde se genera dos fases en equilibrio, las cuales tienen distintas composiciones y propiedades. Éstas se producen al mezclar abundante agua, un polímero y una sal; o bien dos polímeros inmiscibles en abundante agua. A partir de un valor crítico en la concentración de los polímeros (o del polímero y la sal) se forman las dos fases [55], ver la Figura 2 a continuación.

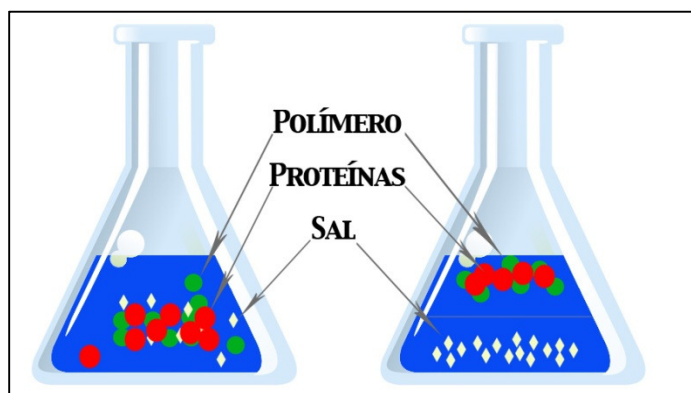


Figura 2: Esquema de dos fases acuosas

En este sistema se puede incorporar proteínas, RNA u organelos celulares; los que migrarán preferentemente a una de las dos fases. Por ejemplo, en un sistema compuesto por polietilenglicol (PEG) y dextrano, la fase rica en proteínas hidrofóbicas, RNA u organelos; es la fase superior [24, 56]. En general, en estos sistemas la fase superior es rica en PEG y la inferior en el otro polímero, o en la sal utilizada [24, 56].

La distribución selectiva de las sustancias entre las fases se expresa a través del coeficiente de partición  $K$ , el cual se calcula mediante la siguiente relación:

$$K = \frac{C_t}{C_b} \quad (9)$$

Donde  $C_t$  y  $C_b$  corresponden a la concentración molar de la sustancia particionada en [mol/L] de la fase superior (top) e inferior (*bottom*), respectivamente. Idealmente el coeficiente de partición es independiente de la concentración total y de la razón entre los volúmenes de las fases [56].

En general, un sistema ATPS no es tan versátil como una cromatografía, desde el punto de vista de la amplia variedad de propiedades que puede utilizar esta última como fuerza motriz de la separación, aunque tanto en ATPS como HIC se pueden agregar ligandos en los polímeros para aumentar la afinidad de una proteína específica. Por otro lado, un sistema ATPS permite la presencia de elementos viscosos como RNA, y es un proceso más simple y menos costoso, en comparación a una cromatografía, si se desea realizar un escalamiento para producción masiva.

Se ha encontrado que las principales propiedades de las proteínas bajo las cuales se produce la separación son: el tamaño [56], la carga total [57-60], y la hidrofobicidad [23, 61-65]. La evidencia indica que la más relevante de éstas es la hidrofobicidad, en base a la cual se ha desarrollado modelos matemáticos, los que han sido de utilidad para la predicción del coeficiente de partición [22, 23].

Los principales parámetros de que depende ATPS son: pH, buffer, tipo y concentración de sal, fuerza iónica, concentración del o los polímeros, peso molecular del o los polímeros, y temperatura [4].

### **3.2.1 Principales Variables Fijas de Operación**

#### ***Efecto de Temperatura***

Existe evidencia que indica que el aumento de la temperatura aumenta el coeficiente de partición de algunas proteínas como la  $\beta$ -glucosidasa en un sistema de PEG/fosfato, pese a que no afecta significativamente al coeficiente de otras, como es el caso de la mioglobina en un sistema PEG/dextrano [55, 66, 67].

#### ***Efecto del pH***

El pH produce un doble efecto, aumentando el coeficiente de partición de moléculas con carga negativa, y disminuyéndolo en moléculas con carga positiva [67-69]. Por otro lado, el pH modifica la carga de las proteínas; un pH inferior al punto isoeléctrico (pH al que se obtiene una carga neutra) de ésta genera una carga neta positiva, mientras que un pH superior genera una carga neta negativa [55, 66, 67].

#### ***Efecto del Peso Molecular del Polímero***

El aumento del peso molecular disminuye la capacidad de interacción del polímero con la proteína, debido a que aumenta su densidad [70, 71]. Por lo tanto, al aumentar el peso molecular del polímero de la fase superior disminuye el coeficiente de partición, y si aumenta el peso molecular del polímero de la fase inferior, entonces aumenta el coeficiente de partición.

#### ***Efecto de la Concentración de Polímero***

Al aumentar la concentración de polímero aumenta la diferencia entre las composiciones de ambas fases, alejando el coeficiente de partición de la unidad. Sin embargo, después de un valor crítico se ha encontrado que  $K$  comienza a acercarse a la unidad nuevamente [71, 72].



### **Efecto de la Concentración de Sal**

El aumento de la concentración de sal tiene dos efectos distintos, primero el efecto *salting out*, que es la disminución de la solubilidad de las proteínas producto del aumento de la tensión superficial; y por otro lado, aumenta el potencial eléctrico entre ambas fases [7].

#### **3.2.2 Modelación de ATPS**

A partir de la termodinámica se puede calcular la relación entre el coeficiente de partición y la energía ( $\Delta E$ ) necesaria para transferir una partícula desde la fase superior a la inferior, según la siguiente expresión [7]:

$$\frac{c_t}{c_b} = e^{\Delta E/K_B T} \quad (10)$$

Donde  $K_B$  es la constante de Boltzmann y T es la temperatura absoluta en [°K]. Esta expresión no depende directamente de ningún parámetro de la proteína ni del sistema, los cuales están indirectamente incluidos en el término  $\Delta E$ . Considerando esto, se aprecia que existe una relación exponencial entre el coeficiente de partición y la propiedad de la proteína mediante la cual se lleva a cabo la separación, lo que concuerda con la evidencia experimental que indica que estos sistemas tienen una gran resolución.

En 1986 Albertson propone el modelo de aproximación de contribución grupal modificada (*modified group contribution approach*) [7], el que se muestra a continuación.

$$\ln(K) = \ln(K^0) + \ln(K_{el}) + \ln(K_{hfob}) + \ln(K_{biosp}) + \ln(K_{size}) + \ln(K_{conf}) \quad (11)$$

Donde los subíndices *el*, *hfob*, *biosp*, *size* y *conf* indican las contribuciones de los factores electro-químicos, hidrofóbicos, bio-específicos, de tamaño y configuración conformacional de la partícula, respectivamente. El término  $K^0$  engloba otros fenómenos no abordados de forma específica.

A continuación, en la Tabla 1, se muestra un resumen de modelos más específicos, propuestos para predecir el coeficiente de partición a partir de una o más variables.

**Tabla 1: Ejemplos de modelos propuestos en la literatura para ATPS.**

Autor	Modelo	Nomenclatura	
1. Albertsson [7]	$\ln K = \ln K_0 + \frac{FZ_p}{RT} \Delta\varphi$	$\Delta\varphi$	Diferencia del potencial eléctrico entre las fases.
		$K_0$	Coefficiente de partición en ausencia de potencial.
		$Z_p$	Carga neta de la proteína.
		$T$	Temperatura.
2. Asenjo <i>et. al.</i> [23]	$\log K = R(\log P - \log P_0)$	$\log P$	Hidrofobicidad proteína.
Hachem <i>et. al.</i> [62]		$\log P_0$	Hidrofobicidad intrínseca
		$R$	Resolución hidrofóbica.
3. Eiteman <i>et. al.</i> [8]	$\ln K = D\Delta w_i \log \frac{P}{P_0}$	$\Delta w_i$	Diferencia de concentración del componente $i$ entre las fases.
		$\log P$	Hidrofobicidad proteína.
		$\log P_0$	Hidrofobicidad intrínseca.
		$D$	Factor de discriminación.
4. Eiteman [68]	$\frac{K}{K_0} = \frac{1 + \sum \Lambda'_{i+} \frac{a_{H+}^i}{\prod K_{bl}} + \sum \Lambda'_{j-} \frac{\prod K_{cl}}{a_{H+}^{j'}}}{1 + \sum \Lambda''_{i+} \frac{a_{H+}^{i'}}{\prod K_{bl}} + \sum \Lambda''_{j-} \frac{\prod K_{cl}}{a_{H+}^{j'}}$	$\Lambda$	Tasa entre los coeficiente de actividad de una especie neutra y cargada.
		$a_x$	Actividad de la especie $x$ .
		$K_0$	Coefficiente de partición proteínas neutras.
		$K_{lb}$	Constante de equilibrio para solutos negativos.
		$K_{cl}$	Constante de equilibrio para solutos positivos.
5. Olivera-Nappa <i>et. al.</i> [73]	$\ln K = A(\delta E)^a (M_p)^b (L)^c + B(\delta E)^d (M_p)^e (L)^f + C$	$\delta E$	Diferencia energía electrostática de solvatación.
		$M_p$	Peso molecular de la proteína.
		$L$	Factor de esfericidad de la proteína.
6. Olivera-Nappa <i>et. al.</i> [73]	$\ln K = A' \frac{(L)^{a'}}{(pH)^{b'} (\delta E)^{c'} (M_p)^{d'}} - B' \frac{(\delta E)^{e'} (pH)^{f'}}{(L)^{g'} (M_p)^{h'}} + C'$	$\delta E$	Diferencia de energía electrostática de solvatación.
		$M_p$	Peso molecular de la proteína.
		$L$	Factor de esfericidad de la proteína.

1. Modelo basado en la diferencia de potencial eléctrico entre las fases, y la carga neta de la proteína.
2. Modelo basado en la capacidad de separación del sistema en base a la hidrofobicidad de una proteína.
3. Modelo basado en la capacidad de separación del sistema en base a la hidrofobicidad de una proteína, y la diferencia en la concentración de un compuesto  $i$  entre las dos fases.
4. Modelo para la predicción en un sistema con compuestos cargados. Basado en el coeficiente de partición con las especies neutras, y el equilibrio fisicoquímico del sistema cargado, medido a través de la actividad de las especies cargadas y neutras, y las constantes de equilibrio de solutos negativos y positivos.
5. Modelo basado en la capacidad de separación del sistema en base a la diferencia de energía electrostática entre las fases. Considera el peso molecular y el factor de esfericidad de las proteínas.
6. Modelo basado en la capacidad de separación del sistema en base a la diferencia de energía electrostática entre las fases. Considera el pH del medio, y el peso molecular y el factor de esfericidad de las proteínas.

### 3.3 Métodos de Clustering

El análisis de información es una ciencia ancestral que precede a los computadores, y que junto con el progreso industrial ha adquirido un nivel creciente de sofisticación.

Dentro de las herramientas de análisis más conocidas se encuentran las derivadas de la ciencia estadística, las cuales permiten por ejemplo: utilizar distribuciones de probabilidad para aprobar o rechazar una hipótesis en base a un nivel de confianza determinado; o ajustar modelos minimizando el error cuadrático y entregando indicadores de la calidad del éste. Las herramientas estadísticas han permitido que las industrias generen productos a gran escala utilizando un mínimo de recursos para asegurar su calidad, y se utilizan constantemente en la investigación de nuevas tecnologías, entre otras muchas aplicaciones.

En la actualidad se genera constantemente una cantidad de información que no es trivial analizar, tanto por su volumen como por su complejidad. Un ejemplo es la información obtenida con los experimentos de amplio espectro (*large-scale*) en genómica funcional como las interacciones proteína-proteína, las interacciones proteína-ADN y los estudios globales de expresión genética [88].

A raíz de la dificultad para analizar grandes volúmenes de información o datos complejos, se ha generado nuevas metodologías que utilizan el gran poder de cálculo de los computadores para extraer información no trivial. A las propiedades que permiten obtener al menos una parte de la información implícita en los datos se le conoce como propiedades topológicas.

Un ejemplo concreto de información topológica es el plano del metro de Madrid (ver Anexo E), o en general de cualquier metro. En ellos se representan las estaciones y las líneas que las unen, pero no siempre son geoméricamente exactos: la curvatura de las líneas de metro no coincide, ni su longitud está a escala, ni la posición relativa de las estaciones, entre otros aspectos; sin embargo, aun así es un plano perfectamente útil. Esto se debe a que este plano representa fielmente cierto tipo de información, la única que necesitamos para decidir nuestro camino por la red de metro: información topológica.

Un grupo importante de herramientas basadas en la topología son los algoritmos de *clustering* o clasificación no supervisada, los que permiten clasificar un conjunto de objetos en grupos (*clusters*) significativos o relevantes sin ninguna información más que los objetos a clasificar, es decir, los grupos son un ordenamiento natural y no arbitrario de los objetos. Un ejemplo de esto se puede ver en la Figura 3, donde en el lado A se muestran datos representados como puntos en un plano, los cuales pueden ser separados de forma intuitiva en los *clusters* del lado B, en donde se diferencian por colores.

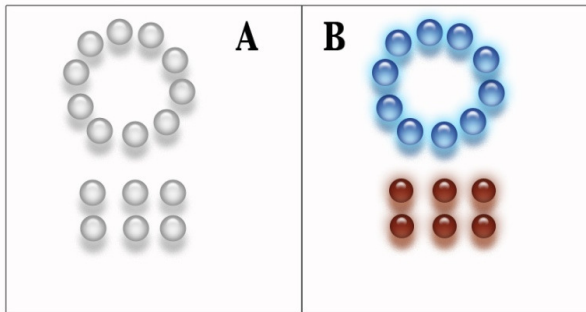


Figura 3: Representación de datos en un plano y agrupación mediante un proceso intuitivo de *clustering*.

Algunos ejemplos de la aplicabilidad de los algoritmos de *clustering* son: el reconocimiento de patrones [74], el procesamiento de imagen [75] y extracción de información (*information retrieval*) [76]. Otras áreas donde han tenido una extensa aplicación son la biología, la arqueología, la geología, la geografía, y el marketing, entre otros [77].

Si bien la evidencia demuestra que estas herramientas tienen un gran potencial, su uso no es trivial. Escoger un método de *clustering* es un paso importante y complejo, ya que tienen distintos enfoques, así como diferencias importantes en sus resultados [15, 78]. Además, cada algoritmo de *clustering* depende de un conjunto de parámetros cuya elección no es directa, frente a los cuales pueden ser muy sensibles. Más aún, el conjunto de parámetros no es universal para cualquier tipo de datos, sino que deben ser ajustados caso a caso. El riesgo de escoger de forma incorrecta el algoritmo y/o los valores adecuados para sus parámetros, está en no identificar algunas estructuras importantes en los datos analizados [15].

Por lo tanto, al trabajar sobre un conjunto de datos para el cual no hay antecedentes suficientes sobre su topología, o los algoritmos más apropiados para su análisis, se requiere utilizar un conjunto de algoritmos que se diferencien en sus características y capacidades para asegurar la extracción de los *clusters*. Para escoger los algoritmos de *clustering* que permitan resolver óptimamente el problema objeto de esta tesis, es necesario entender cómo funcionan estos algoritmos, así como sus principales diferencias. Con este propósito, se presenta a continuación una revisión de las etapas involucradas en la clasificación no supervisada. Luego se describen las principales características de este tipo de algoritmos, en base a la cual se presenta una clasificación. Finalmente, se presenta una descripción general de cada clase y los algoritmos a utilizar en esta tesis.

A través del uso de estas herramientas se espera poder determinar grupos de APVs que compartan características, así como APVs complementarios. Además se espera poder generar nuevas escalas que conserven las propiedades topológicas, algunas de las cuales sirvan para construir modelos con un poder predictivo superior a los reportados a la fecha, respecto a los mismos tipos de modelos.

### 3.3.1 Etapas Involucradas en la Clasificación no Supervisada (*Clustering*)

El proceso de *clustering* se puede dividir en cinco etapas, de las cuales las tres primeras son las necesarias para llevar a cabo el proceso de *clustering*, estas son: representación de objetos, definición de similitud y agrupamiento o *clustering* (ver Figura 4). Las últimas dos etapas son: abstracción de datos y evaluación del resultado. Estas últimas dos etapas pueden ser prescindibles en algunos casos, ya que corresponden a procesos de validación y procesos necesarios para utilizar los resultados en otros tipos de análisis.

A continuación, en la Figura 4, se muestran las etapas involucradas en el proceso de *clustering*, las que son secuenciales con un *feedback*. Esto se debe a que la clasificación no supervisada requiere ajuste, y es un proceso iterativo complejo [15, 78].

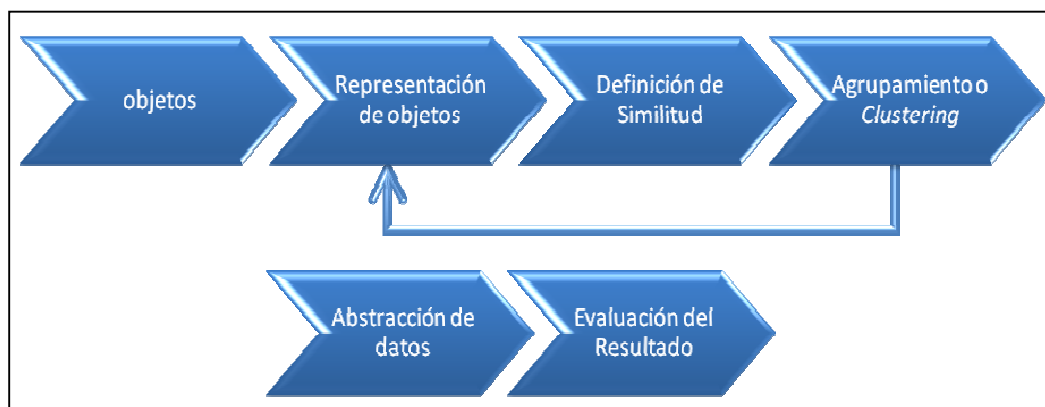


Figura 4: Diagrama del proceso de *clustering*.

#### **Representación de los Objetos**

Una de las estructuras más comunes para representar los objetos son los vectores, donde cada componente o dimensión corresponde a una característica de un objeto, por lo tanto el valor que haya en dicha componente es una evaluación respecto a la característica correspondiente. Por ejemplo, un vector puede representar una posición en el espacio, en cuyo caso es un arreglo con tres componentes, en que cada uno representa una distancia medida en uno de los tres ejes espaciales  $\hat{i}$ ,  $\hat{j}$  y  $\hat{k}$ . Otras alternativas para representar un conjunto de puntos en el espacio son el uso de coordenadas esféricas, o el uso de un grafo.

En un grafo un objeto está representado mediante un nodo y sus propiedades dependen de sus conexiones con otros objetos. Un grafo está constituido por la lista de nodos y la lista de interacciones entre ellos, la que puede incluir un coeficiente de similitud (ver sección 3.3.2). Mediante grafos se puede representar prácticamente cualquier problema.

Además de la estructura utilizada para representar los datos, también es importante determinar las propiedades de los objetos que se representarán, es decir, aquellas que se consideran más relevantes.

## **Definición de una Medida de Similitud**

La similitud entre dos objetos se entiende como el nivel de cercanía entre ellos como un todo, o el nivel de cercanía entre las distintas propiedades de éstos, por lo tanto, para la clasificación de los objetos en *clusters* se utiliza una medida de proximidad que permita maximizar la similitud entre los objetos de un mismo *cluster* y/o minimizar la similitud entre objetos de distintos *clusters* [78]. Existen diversas formas de medir la proximidad, dentro de las cuales se pueden distinguir dos clases: medición de proximidad entre objetos individuales, por ejemplo, a través de la distancia euclidiana; y medición entre grupos de objetos de forma simultánea. En este último caso la proximidad se puede medir: entre los objetos de un mismo *cluster*, por ejemplo, midiendo el error cuadrático con respecto al centroide; o entre los objetos de dos *clusters* diferentes, por ejemplo, midiendo el error cuadrático de una clase respecto al centroide de otra clase, y sumando el error cuadrático recíproco.

En general, las distintas definiciones de distancia permiten obtener distintos tipos de *clusters*, por ejemplo: a través de la Distancia Euclidiana se ha extraído *clusters* hiper-esféricos; y utilizando la Distancia Regularizada de Mahalanobis se ha extraído *clusters* hiper-elipsoidales [79]. La distancia de Mahalanobis está definida de la siguiente manera:

$$d_M(X_i, X_j) = (X_i - X_j) \Sigma^{-1} (X_i - X_j)^T \quad (12)$$

Donde los objetos  $X_i$  y  $X_j$  se asumen que son vectores columnas y  $\Sigma^{-1}$  es la matriz de covarianza de los objetos muestreados, o la matriz de covarianza conocida del proceso de generación de objetos.  $d_M(\cdot, \cdot)$  asigna diferentes pesos a los diferentes objetos basado en sus varianzas y correlación lineal entre pares [15].

En algunos casos la medida de proximidad puede tomar formas que derivan del trabajo de los objetos como grafos, como el Coeficiente de Clustering [80] o la Entropía en Grafos [81].

## **Agrupamiento ó Clustering**

Las metodologías de *clustering* son diversas, por ejemplo, algunas se basan en sucesivas particiones del conjunto inicial de datos; otras en aglomeraciones sucesivas; y otras abordan el problema a través de aprendizaje topológico. Además, existen métodos determinísticos y estocásticos, y algunos utilizan algoritmos de optimización [15]. Estos métodos, sus características y clasificación se presentan en las secciones siguientes.

## **Abstracción de los Datos**

Una vez terminado el proceso de *clustering*, puede ser necesario representar los datos para su análisis visual; o transformar los datos para ser utilizados en otros procesos de análisis o aplicaciones, automatizando el proceso global.

Un ejemplo para el primer tipo de abstracción es la reducción del número de dimensiones de los datos para que se puedan representar gráficamente (2D o 3D); y un ejemplo del segundo tipo de abstracción es el uso del centroide de cada *cluster*.

## Evaluación del Resultado

La evaluación de los *clusters* obtenidos por un algoritmo no es sencilla, y el proceso de discriminación entre un “buen” y un “mal” *cluster* puede ser bastante subjetivo. Sin perjuicio de esta subjetividad, existen tres formas para validar los clusters. Estas son:

- Una validación independiente, que compara las estructuras obtenidas con estructuras conocidas de antemano.
- Una validación interna, para determinar si las estructuras encontradas son intrínsecamente apropiadas a los datos. Una forma de realizar lo anterior es transformar los datos a un grafo y analizar el Coeficiente de Clustering [80] de la red (para encontrar más información respecto a otras metodologías sobre este punto, revisar [82, 83]).
- Un test relativo, en el cual se comparan dos estructuras y se miden sus méritos relativos. Por ejemplo, al comparar los clusters obtenidos con un algoritmo de clustering con otros obtenidos al azar (algunos índices utilizados para estas comparaciones se discuten en detalle en [77, 82]).

### 3.3.2 Características de los Algoritmos de Clustering

A continuación se presenta una breve descripción de las principales características que afectan el desempeño de los algoritmos de *clustering*.

#### Mecanismo de Aglomeración o Partición

Esta característica define el estado de partida y la forma en que opera un algoritmo. En la Figura 5 se muestra como, a partir de un estado inicial en que cada objeto es a la vez un *cluster* diferente, los algoritmos aglomerativos utilizan una heurística para agrupar los *clusters* de forma iterativa hasta que se cumpla algún criterio de parada. Por otro lado, en los algoritmos de partición inicialmente todos los objetos pertenecen a un mismo *cluster*, y la heurística del algoritmo define la forma en que este *cluster* se divide de forma sucesiva en *clusters* hasta que se satisface un criterio de detención. Esta forma de división sucesiva normalmente está dada por la optimización de una función objetivo.

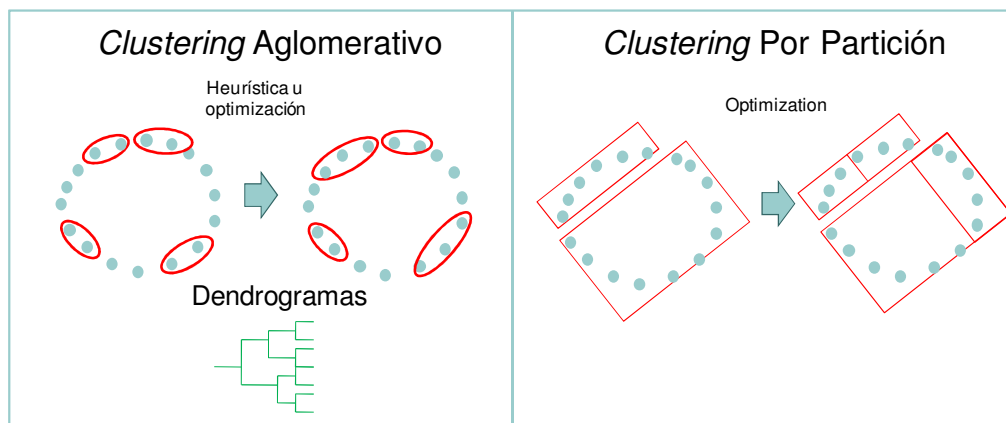


Figura 5: Esquema de algoritmos de clustering aglomerativos y de partición.

## **Metodología de Uso de las Propiedades de los Objetos**

Las propiedades de los objetos se pueden emplear de muchas maneras, pero se puede distinguir entre dos metodologías básicas: secuencial (*Monotetic Algorithm*) o simultánea (*Politetic Algorithm*) [15]. En la primera, un algoritmo utiliza de forma separada las características de los objetos en el proceso de *clustering*, en cambio, las metodologías simultáneas utilizan todas las dimensiones o propiedades de los objetos a través de alguna función de distancia. El problema con los algoritmos del tipo secuencial es que se genera un alto número de *clusters*, lo que puede generar pequeños fragmentos de *clusters* que no resultan útiles [15].

### **Definición de Clusters Utilizada**

La definición de un *cluster* se puede realizar de manera que cada objeto pertenezca a un único *cluster* (*Hard Algorithm*); o de forma que cada objeto tenga un grado de pertenencia a cada *cluster* (*Fuzzy Algorithm*) [15, 84].

### **Determinístico v/s Estocástico**

Este aspecto es relevante para los algoritmos de partición, en los cuales se optimiza alguna función de similitud, entre otros. La clasificación de los algoritmos como determinísticos o estocásticos depende de si utilizan el azar para explorar las distintas soluciones o agrupaciones, dentro del proceso de optimización. A pesar de esta definición, existen algoritmos determinísticos que utilizan soluciones iniciales determinadas al azar.

### **Uso de Recursos**

Los distintos algoritmos de *clustering* y sus implementaciones se pueden diferenciar por la cantidad de memoria y tiempo de ejecución requerida. Algunos algoritmos están diseñados para examinar un número reducido de objetos, o bien para disminuir el tamaño de las estructuras de datos utilizadas en las operaciones del algoritmo. Sin embargo, las especificaciones de un algoritmo de *clustering* normalmente dejan mucha flexibilidad para su implementación [15].

### **Representación de Resultados**

Un algoritmo se puede clasificar como jerárquico dependiendo de la estructura de la solución que genera. Los algoritmos jerárquicos no producen particiones simples de los datos, sino que estructuras llamadas dendrogramas, a partir de los cuales se pueden generar distintos grupos de *clusters*. En los dendrogramas se reflejan los distintos niveles de similitud entre los objetos. En la Figura 6 se muestra un dendrograma en el cual se aprecia que, al mover el nivel de corte (línea entrecortada), se pueden generar distintos grupos. Por ejemplo, en la posición de corte actual se tienen 3 grupos: (A, B, C); (D, E) y (F, G).



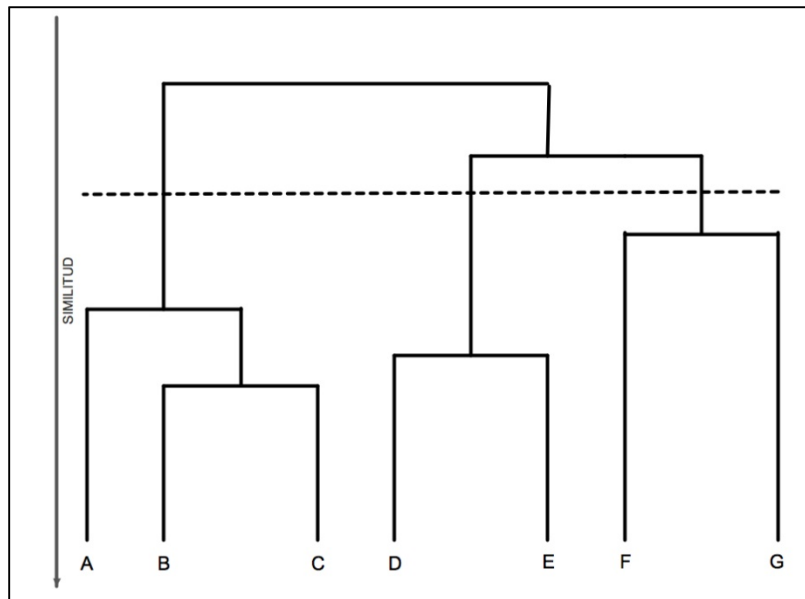


Figura 6: Dendrograma. Una forma común de mostrar los resultados de un algoritmo jerárquico.

### **Estructura de los Objetos a Clasificar**

Anteriormente en esta sección se mencionó que existen distintas estructuras de representación de la información, siendo las más utilizadas los vectores y los grafos.

El trabajo con grafos ha sido muy exitoso para diversas aplicaciones, como el descubrimiento de los principios de la organización que gobierna la formación y evolución de varias tecnologías complejas y redes sociales [80, 85]. Mediante grafos se puede representar prácticamente cualquier problema. La teoría de grafos es muy compleja e igualmente rica, y los algoritmos de *clustering* para grafos se pueden utilizar en muchas clases de aplicaciones.

Una de las desventajas de los grafos es la dificultad para generarlos, que se requiere calcular las matrices de similitud ( $O(n^2)$ ) y solo se puede aplicar sobre unos cuantos miles de datos [15]. Sin embargo, en algunas aplicaciones la información está estructurada como grafo, como la generada por los *microarrays*, que permite la evaluación simultánea del estado de componentes celulares [85]. En la Figura 7 se muestra una representación de un grafo, el cual es del tipo *scale-free*, una clase muy frecuente como es el caso de la internet, las redes sociales, las redes de interacción proteína – proteína, etc.; el cual se caracteriza por una ley de distribución de potencia, esto es, la probabilidad de que un nodo tenga  $k$  conexiones es  $P(k) \sim k^{-\gamma}$ , donde  $\gamma$  es el grado del exponente [85].

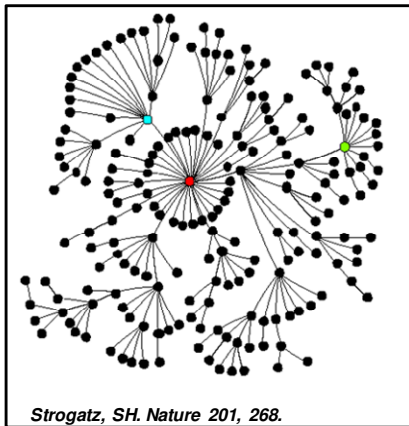


Figura 7: Esquema de un grafo del tipo *scale-free*.

### Características de los Clusters

Hay dos tipos generales de *clusters* que se diferencian en el tipo de relaciones entre los objetos y las dimensiones de las características que comparten, los globulares y los transitivos [15]. Estos términos nacen de la forma característica de estos *clusters* en un grafo, donde es más sencillo comprender sus diferencias.

En un *cluster* globular todos los pares de objetos componentes del *cluster* tienen arcos con valores de similitud muy grandes, por lo tanto los objetos representados en un espacio multidimensional (como vectores) compartirán un sub-espacio (grupo de componentes de los vectores) en el cual forman un conjunto de objetos muy compacto (ver Figura 8). Las dimensiones de este sub-espacio, así como la densidad (que tan grande es la fracción de objetos que comparten las mismas dimensiones) puede ser diferente entre *clusters*, lo que puede dificultar su identificación.

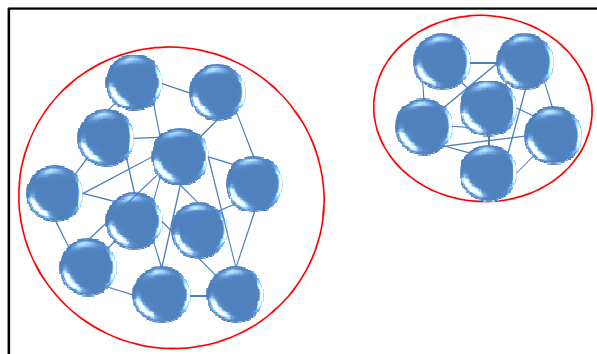


Figura 8: Ejemplo de cluster globulares en el plano.

En el caso de un *cluster* del tipo transitivo, puede haber valores de similitud bajos entre pares de objetos, pero éstos están conectados a su vez a otros objetos pertenecientes al *cluster* mediante conexiones con altos valores de similitud (ver Figura 9). En términos de vectores esto significa que dentro del *cluster* puede haber un sub-espacio, en el cual los grupos de objetos del *cluster* comparten muy pocas dimensiones (esto es, que existen pocas dimensiones en que los objetos representados como vectores poseen componentes dentro de un rango en que se pueden considerar cercanos o que denotan una similitud), pero en el cual hay un *strong path* entre los *clusters* que los conectará. Por *strong path* se entiende que si  $A$  y  $B$  son dos *sub-clusters* que comparten solo unas

pocas dimensiones, entonces habrá otro conjunto de *sub-clusters*  $x_1, x_2, \dots, x_k$ ; tal que cada uno de los pares de *sub-clusters*  $(A, x_1), (x_1, x_2), \dots, (x_k, B)$  compartirán varias de las dimensiones del sub-espacio de  $A$  y  $B$  [86]. Lo que complica el descubrimiento de estos clusters es que las conexiones entre *sub-clusters* tienden a ser de distinta fuerza.

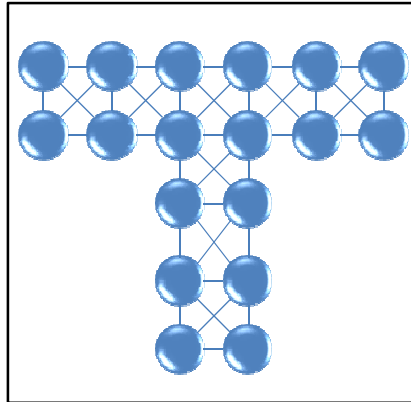


Figura 9: Ejemplo de clusters transitivos

Al decidir entre un algoritmo capaz de encontrar *clusters* globulares y transitivos, es necesario comprender qué tipo de *clusters* tiene mayor significancia en la aplicación puntual.

### 3.3.3 Clasificación de los Algoritmos de Clustering

En base a las características descritas en la sección 3.3.2 es posible clasificar los algoritmos de *clustering* de distintas maneras. Para efectos de este trabajo, en la Figura 10 se propone una modificación de la clasificación propuesta por Jain y Dubes [77].

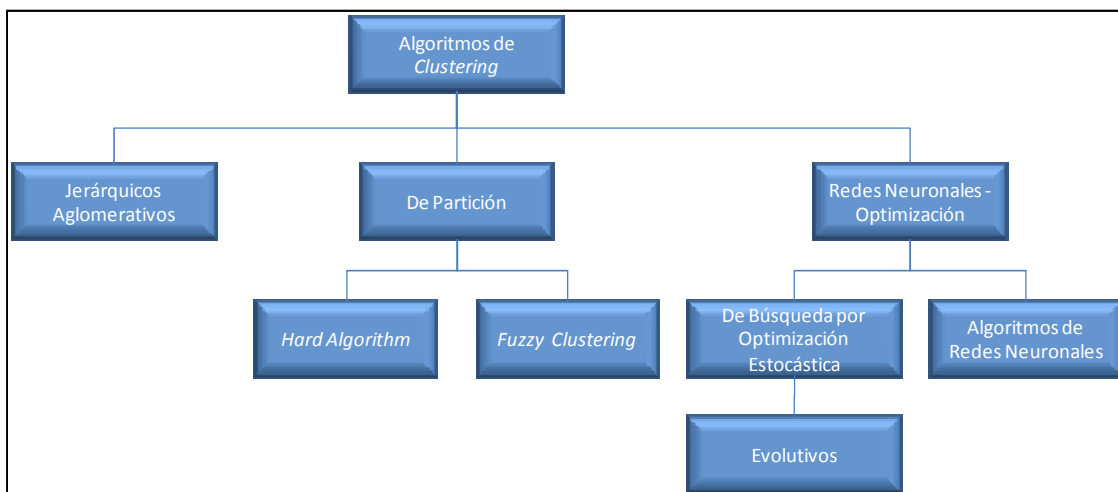


Figura 10: Clasificación propuesta para los algoritmos de *clustering*.

En la clasificación propuesta por Jain y Dubes (referente utilizado), se definen dos principales clases de algoritmos; sin embargo, en la clasificación propuesta en este trabajo se definen tres clases de algoritmos. En la clase agregada, denominada Redes Neuronales-Optimización se incluyen los Algoritmos de Búsqueda por Optimización Estocástica y los Algoritmos de Redes Neuronales.

Además, Jain y Dubes incluyen una clase especial para los algoritmos que utilizan grafos; sin embargo, en general los algoritmos pueden utilizar una representación de los datos mediante vectores, o grafos, y en ambos casos se utiliza (con algunas excepciones) el mismo algoritmo de *clustering*.

El primer nivel de clasificación propuesto se basa en el tipo de metodología: aglomerativas, de partición, y de búsqueda por Optimización o de Redes Neuronales (ANNs). Dentro de los Algoritmos Aglomerativos (que también son del tipo jerárquicos) nace una división natural por la metodología específica utilizada para aglomerar los objetos. Dentro de los algoritmos de partición encontramos los algoritmos que se basan en criterios de similitud tipo error cuadrático y los algoritmos difusos (*Fuzzy Algorithm*). Dentro de los Algoritmos de Redes Neuronales-Optimización se encuentran los Algoritmos de Búsqueda por Optimización Estocástica y los Algoritmos de Redes Neuronales. Dentro de los Algoritmos de Búsqueda por Optimización Estocástica se destacan en la Figura 10 los Algoritmos Evolutivos, ya que se utilizó un algoritmo de optimización perteneciente a esta clase.

A continuación se hace una breve descripción de las categorías de clasificación y los algoritmos más importantes en éstas.

### 3.3.4 Descripción de las Distintas Clases de Algoritmos

#### ***Aglomerativos***

La mayoría de los algoritmos de *clustering* aglomerativos, base de los Algoritmos Jerárquicos, son variantes de los algoritmos *single-link* [87], *complete-link* [87], y *minimum-variance* [88, 89]. De éstos, *single-link* y *complete-link* son los más populares, y se diferencian en la medida de similitud entre pares de *clusters*, a partir de la cual se define si un par de *clusters* se fusionan o no.

En el método *single-link* la distancia entre dos clusters es la distancia mínima entre todos los pares de objetos al escoger cada vez uno del primer *cluster* y otro del segundo. Este método sufre de un efecto en cadena a raíz del cual se genera una tendencia a producir *clusters* que son elongados [90] (ver Figura 11), pero es versátil, ya que es capaz de separar los *clusters* de la Figura 12.

El algoritmo *complete-link* utiliza la máxima distancia entre los mismos pares de objetos de los dos *clusters*. Al comparar ambos algoritmos se aprecia que el *complete-link* produce *clusters* más compactos y estrechamente unidos [91].

En la Figura 11-A (*clusters* generados con el algoritmo *single-link*) y Figura 11-B (*clusters* generados con el algoritmo *complete-link*) se aprecia dos *clusters* separados por un conjunto de objetos con la intención de generar ruido (esferas en rojo). En la práctica se ha demostrado en varias aplicaciones que las jerarquías generadas por el algoritmo *complete-link* son de mayor utilidad [77].

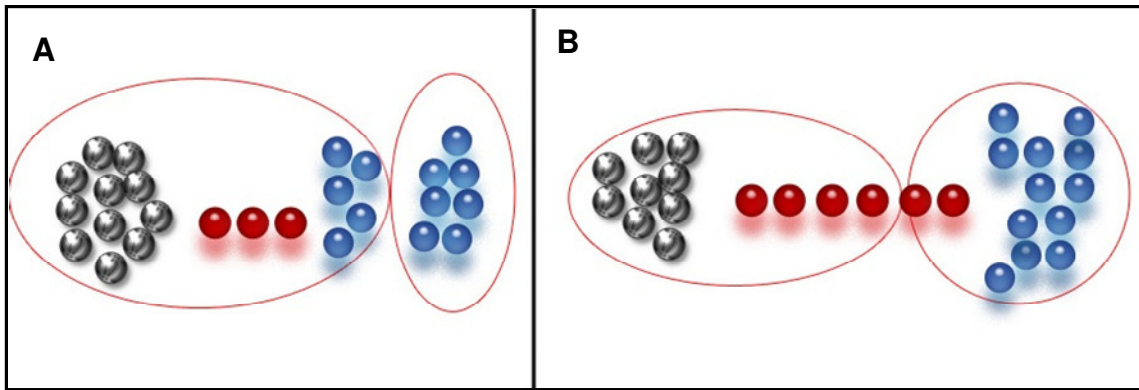


Figura 11: Ejemplos de *clusters* elongados (A) y compactos (B).

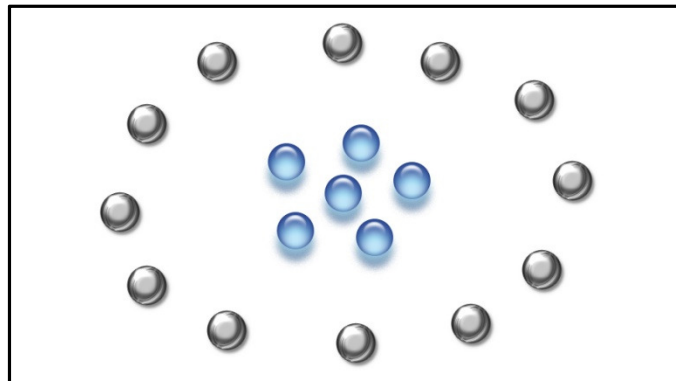


Figura 12: Ejemplo de *clusters* difíciles de distinguir.

En esta tesis se realiza una preselección, donde se evalúan los métodos *single-link* y *complete-link*, entre otros; pero los mejores resultados se obtienen con el *average-link*, el cual utiliza la distancia media entre todos los objetos de dos *clusters* para evaluar la similitud.

### **Algoritmos de Partición**

#### *Hard Algorithm*

Los algoritmos de partición permiten trabajar sobre conjuntos de datos para los cuales la creación de un dendrograma tiene un requerimiento computacional prohibitivo, sin embargo, están limitados debido a que no son capaces de obtener el número de *clusters* que representan la mejor partición de los datos, lo que debe ser ingresado como un parámetro más del algoritmo (para ver más sobre la elección del número de *cluster* revisar [82]).

El criterio de similitud entre grupos de objetos más intuitivo y frecuentemente utilizado en las técnicas de *clustering* es el error cuadrático (ver Tabla 2, criterio MSE), el cual se ha encontrado que en general funciona muy bien con *clusters* aislados y compactos [79]. Sin embargo, existen muchos criterios de similitud, como por ejemplo los que se muestran a continuación en la Tabla 2. Estos criterios se puede utilizar para caracterizar un *cluster* al comparar los objetos que contiene, o para caracterizar de forma relativa al comparar objetos entre *clusters*.

Tabla 2: Definición matemática de las funciones criterio de similitud entre clusters. La notación es la siguiente:  $k$  es el número total de clusters,  $S$  es el total de objetos a clasificar,  $S_i$  es el conjunto de objetos asignados al cluster  $i$ ,  $n_i$  es el número de objetos en el cluster  $i$ ,  $v$  y  $u$  son dos objetos, y  $\text{Sim}(v, u)$  es la similitud entre los objetos.

Criterio de Similitud	de Función para Optimización
MSE	$\text{Min} \sum_{i=1}^k \sum_{u,v \in S_i} \frac{\text{Sim}(v,u)^2}{n_i^2}$
$J_1$	$\text{Max} \sum_{i=1}^k \frac{1}{n_i} (\sum_{u,v \in S_i} \text{Sim}(v, u))$
$J_2$	$\text{Max} \sum_{i=1}^k \sqrt{\sum_{u,v \in S_i} \text{Sim}(v, u)}$
$\varepsilon_1$	$\text{Min} \sum_{i=1}^k n_i \frac{\sum_{v \in S_i, u \in S} \text{Sim}(v,u)}{\sqrt{\sum_{u,v \in S_i} \text{Sim}(v,u)}}$
$\mathcal{G}_1$	$\text{Min} \sum_{i=1}^k \frac{\sum_{v \in S_i, u \in S} \text{Sim}(v,u)}{\sum_{u,v \in S_i} \text{Sim}(v,u)}$
$\mathcal{G}'_1$	$\text{Min} \sum_{i=1}^k n_i^2 \frac{\sum_{v \in S_i, u \in S} \text{Sim}(v,u)}{\sum_{u,v \in S_i} \text{Sim}(v,u)}$
$\mathcal{H}_1$	$\text{Max } J_1 / \varepsilon_1$
$\mathcal{H}_2$	$\text{Max } J_2 / \varepsilon_1$

El algoritmo *K-means* es el más simple y comúnmente utilizado de los algoritmos que emplean el criterio del error cuadrático [92]. Éste parte con una solución inicial al azar y luego agrupa objetos a los *clusters* basado en la similitud entre éstos y el centro del *cluster*, hasta que se alcanza un criterio de convergencia. Este algoritmo es popular por ser de fácil implementación, y porque su complejidad es del orden del número de objetos,  $\mathcal{O}(n)$ . El mayor problema con este algoritmo es que es muy sensible a la partición inicial seleccionada, y puede converger fácilmente a un mínimo local.

Existen distintas variaciones de este algoritmo, algunas de las cuales incluyen metodologías para escoger una buena partición inicial [74]. Otros permiten fusión y división de los *clusters* en base a valores críticos de la distancia entre los centroides y la varianza respectivamente.

En esta tesis se utiliza el algoritmo *K-means* y el *Bisection* [78]. Este último es similar al *K-means*, con la diferencia que siempre realiza sucesivas particiones de un grupo de datos, en otros dos.

### *Fuzzy Clustering (FC)*

Los algoritmos del tipo *Fuzzy clustering* son una generalización de los algoritmos de partición, en la cual el concepto de pertenencia a un *cluster* se abre a un grado o nivel de pertenencia. Por lo tanto, el problema más importante es definir la función de pertenencia.

Un par de ejemplos son la generalización del algoritmo *fuzzy c-means* que fue presentado por Bezdek [93], cuya convergencia es al mínimo del criterio del error cuadrático; y el algoritmo *fuzzy c-shell* presentado por Dave [94] que es capaz de detectar *clusters* circulares y elípticos.

## **Búsqueda por Optimización Estocástica**

El concepto de búsqueda normalmente se asocia al conocimiento del objeto que se desea obtener, por lo tanto se busca dentro de una lista o un vector; por otro lado, la optimización global comprende todas las técnicas que pueden ser utilizadas para encontrar el mejor elemento  $x_i$  en el dominio  $\mathbb{X}$  con respecto al criterio  $f \in F$ . Como se señala en esta sección, los algoritmos de partición también utilizan la optimización, pero dentro de un contexto específico. Cuando hablamos de búsqueda por optimización estocástica, el elemento  $x_i$  corresponde a una configuración específica de datos agrupados en *clusters*, y el dominio  $\mathbb{X}$  son todos los estados o configuraciones posibles al agrupar  $n$  datos en  $m$  *clusters* de tamaño promedio  $s$  datos, que da un dominio del orden de  $s^m$  (si consideramos  $s = 5$  y  $m = 14$ , entonces  $n = 70$  y  $s^m \approx 3 \cdot 10^{10}$ ). Como el dominio es extremadamente grande es posible utilizar estrategias de reducción de éste, por ejemplo, buscando un *cluster* a la vez.

Para que esta búsqueda sea efectiva se requiere un algoritmo rápido y de búsqueda global, razón por la cual se utilizan algoritmos estocásticos, los que obtienen soluciones cercanas a la óptima en tiempos muy inferiores a los determinísticos [84].

Algunos ejemplos de métodos estocásticos son: *Evolutionary Algorithm* [84, 95-97], *Hill Climbing* [84, 98], *Random Optimization* [84, 99], *Simulated Annealing* [84, 100] y *Tabu Search* [84, 101-105].

Junto con lo anterior, para realizar una búsqueda en todo el espacio es necesario conocer propiedades específicas que caractericen los *clusters* que se desea encontrar, y funciones que utilicen la información topológica para reconocer estas propiedades en los datos sobre los cuales se trabaja. Como este no es el caso de los APVs, porque se desconoce esta información, no se espera tener buenos resultados con esta clase de algoritmos, aunque de todas formas se utilizó el algoritmo *Restricted Neighbour Search* [106, 107], con el objeto de no descartar a priori ninguna clase de algoritmo.

## **Algoritmos Evolutivos**

Este tipo de algoritmos de *optimización* nace motivado por la evolución natural según los principios descritos por Charles Darwin, con los cuales a partir de un primer organismo (o unos cuantos) se habría creado toda la diversidad de especies que ha existido en el planeta [108].

Los aspectos más básicos e importantes de la teoría de Darwin que son emulados por algoritmos evolutivos, como el *Genetic Algorithm* por David Goldberg [109], y que está influenciado por los trabajos de Henry Holland [110, 111], son:

- Los organismos padres pueden generar nuevos organismos hijos que no son idénticos a los padres.
- Los individuos están sometidos a una presión selectiva natural, frente a la cual los distintos organismos tienen un desempeño diferenciado, ya que algunos se adaptan mejor al medio. Las nuevas características que adquieren los hijos pueden mejorar o empeorar su adaptación al medio respecto al desempeño de los padres.

En esta clase de algoritmos, los candidatos a solución de un problema determinado juegan el rol de individuos. La adaptación al medio se evalúa a través de las funciones objetivos a optimizar, las que direccionan la evolución en ciertos sentidos. Para cada problema que se quiera resolver se pueden especificar múltiples funciones objetivos, donde cada una representa una característica en la cual se está interesado.

Para generar los nuevos candidatos a solución (nuevos individuos) se emula el proceso de recombinación del material genético de la reproducción de mamíferos. A través de este proceso es posible cambiar en un paso de una solución a otra muy distinta, lo que en general no son capaces de realizar otros algoritmos como: *K-means*, *Fuzzy Clustering Algorithms*, ANNs (ver sección 3.3.5), varios esquemas de *Annealing*, y *Tabu Search*; por ser todas técnicas de búsqueda localizada [84]. También se emula el proceso de mutación genética, el cual modifica las soluciones al azar, asegurando que la búsqueda de nuevas soluciones se realice en todo el espacio.

En la implementación del *Genetic Algorithm* [84, 112-115] típicamente se utiliza una representación como cadena de caracteres (*string*) binaria. A continuación se utiliza esta representación para esquematizar cómo funciona el proceso de *crossover* y el de mutación [84].

Sean dos soluciones (padres) representadas como las cadenas de caracteres [1 0 1 1 0 1 0 1] y [1 1 0 0 1 1 1 0]. Como se muestra en la Figura 13, para realizar el proceso de *crossover* se alinean ambas soluciones y se escoge un punto (punto de *crossover*) que da lugar a dos sub-cadenas para cada padre. La primera de las dos nuevas soluciones será la secuencia constituida por la sub-cadena izquierda del primer padre, concatenada a la derecha por la sub-cadena izquierda del segundo padre. La segunda solución es la concatenación de las otras dos sub-cadenas en el orden correspondiente.

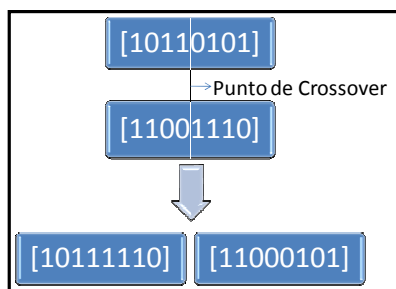


Figura 13: Esquema del proceso de *crossover*.

La mutación por otro lado toma solo una solución, por ejemplo [1 1 1 1 1 1 1 0], y se escoge una posición al azar a la cual se le cambia el valor. Si el resultado de la selección de una posición al azar fuera 2, entonces la nueva solución tendría la forma [1 0 1 1 1 1 1 0] (para ver el procedimiento completo que utiliza este algoritmo para realizar la optimización ver el Anexo F).

La ventaja del *Genetic Algorithm*, comparado con otros métodos de optimización, es que requiere muy pocas hipótesis sobre la metodología empleada para medir la adaptación al medio, y por lo tanto, mantiene un buen desempeño en muchos problemas de distinta índole [15]. Por otra parte, la desventaja de dicho algoritmo es que requiere ajustar varios parámetros frente a los cuales es muy sensible [15], y que



el costo computacional también depende del costo de la función objetivo, ya que ésta se evalúa para cada individuo u objeto generado por el algoritmo.

### **Algoritmos de Redes Neuronales (ANNs)**

Estos algoritmos, cuya sigla en inglés es ANNs (*Artificial Neural Networks*), son modelos basados en redes neuronales biológicas [116]. Desde un punto de vista práctico se puede decir que las redes neuronales artificiales son una herramienta de estadística no lineal para la modelación de datos [117].

Los algoritmos de esta clase se originan a través de la simulación del proceso de aprendizaje que tiene el cerebro humano. El cerebro humano consiste de un gran número (más de mil millones) de células neuronales que procesan información. Cada célula funciona como un procesador simple, y solo la masiva interacción entre ellas y los procesos paralelos que ejecutan hacen posible las habilidades del cerebro. La información pasa a través de las neuronas como impulsos eléctricos, los que son transmitidos de una neurona a otra siempre que la estimulación provocada exceda un valor crítico; si pasa dicho valor, se dice que la neurona está activa. Las conexiones entre neuronas se adaptan con el tiempo, por lo tanto la estructura que generan las conexiones es dinámica, lo cual es la base del aprendizaje [16].

Las limitaciones computacionales, así como la complejidad de los procesos físico-químicos involucrados en la transmisión de información a través de las neuronas, han llevado a una simulación simplificada. En este modelo, una red neuronal consiste en neuronas y conexiones entre ellas (ver Figura 15). Como se muestra en la Figura 14, una neurona es capaz de recibir información (*input*) de muchas neuronas, evaluar su activación, y en caso de activarse podrá transportar la información a otras neuronas con las que está conectada (*output*). Estas conexiones tienen una ponderación o peso relativo, lo que permite simular la información eléctrica y además permite simular el cambio estructural a través del cambio del valor de estos pesos. Cada neurona recibe muchos *inputs* simultáneos ponderados con distintos pesos, y mediante una función de propagación se transforman todos éstos a un único valor, que se compara con el valor crítico de la neurona, el cual a su vez es calculado con una función de activación.

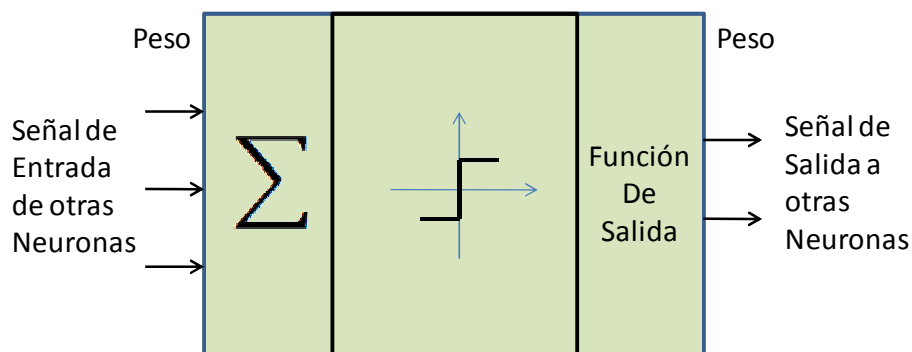


Figura 14: Estructura de la función de una neurona en una red neuronal.

En una red neuronal las distintas neuronas son agrupadas en capas neuronales. Usualmente cada neurona de una capa está conectada con todas las neuronas de la capa siguiente, excepto la capa final (*output*).

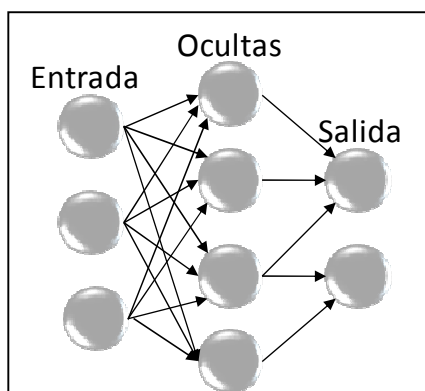


Figura 15: Esquema de una red neuronal. Se aprecia la existencia de una capa input, una output y capas ocultas (*hidden*).

Los algoritmos de redes neuronales (ANNs) han sido utilizados extensivamente durante las últimas tres a cuatro décadas para la clasificación y *clustering* [117, 118]. Probablemente una de las características más importantes de los ANNs durante el proceso de *clustering* de objetos es la capacidad de modificar el peso de las interconexiones de los nodos de forma adaptativa [119, 120]. Más específicamente, pueden actuar como normalizadores y selectores de las características de los objetos mediante la selección apropiada de los pesos de las conexiones.

Algunos de los últimos algoritmos hechos en base a ANNs han incorporado el objetivo de la reducción dimensional. Para una proyección lineal de los datos se puede utilizar el análisis de componentes principales o algunos modelos de redes en forma iterativa [120, 121]. El *Self-Organizing Maps* (SOM) [122] y el *Growing Cell Structures* [123] permiten la proyección no lineal, aunque se deben escoger las dimensiones a priori.

Otros algoritmos han logrado un aprendizaje topológico: dada una distribución  $P(\xi)$  de datos con un gran número de dimensiones, son capaces de encontrar estructuras de menor dimensión que preservan la topología de los datos. Algunos ejemplos de estos algoritmos son *Competitive Hebbian Learning* [124], *Neural Gas* [125], *Growing Neuronal Gas* (GNG) [16], y *Growing Grid* (GG) [126] (para ver el pseudo código de los algoritmos GNG y GG, que son utilizados en esta tesis, ir al Anexo G y Anexo H, respectivamente).

### 3.3.5 Comparación Entre Algoritmos de Clustering

En la sección anterior se examinaron distintas clases de algoritmos, los cuales utilizan metodologías muy diversas con la intención de encontrar *clusters* significativos. Las principales clases de algoritmos definidas son: Algoritmos Aglomerativos, Algoritmos de Partición, y Algoritmos de Redes Neuronales-Optimización, dentro de los cuales se encuentran los Algoritmos de Búsqueda por Optimización y los Algoritmos de Redes Neuronales (ANNs). En esta sección se realiza una comparación de estos algoritmos en base a estudios empíricos de comparación existentes en la literatura.

Según los estudios revisados los *clusters* obtenidos por Algoritmos Aglomerativos son más versátiles que los encontrados por los algoritmos basados en el error cuadrático (una clase de algoritmos de partición). Esto se debe a que los *clusters* seleccionados por estos últimos métodos normalmente son hiper esféricos [84].

Por otro lado, en general los algoritmos de optimización estocástica realizan una búsqueda global, mientras que los Jerárquicos (aglomerativos) y los de Partición son, en general, enfoques de búsqueda localizada. Se ha mostrado que el Algoritmo *K-means* converge rápidamente a la solución óptima local, por lo que si se pudiera obtener rápidamente una buena partición de partida con cualquiera de los otros algoritmos, *K-means* funcionaría bien aún con grandes conjuntos de datos [127].

Un estudio empírico hecho por Mishra y Raghavan [127] compara los algoritmos de optimización GA, SA, TS e *Hybrid Search* (HS) [128]; y concluye que, en general, ningún método supera en el rendimiento a los otros por un margen significativo, al trabajar con datos unidimensionales o multidimensionales.

Otro estudio empírico sobre los algoritmos *K-means*, SA, TS y GA fue presentado por Al-Sultan y Khan [129]. TS, GA y SA se evaluaron como comparables en términos de la calidad de la solución, y en todos los casos fueron superiores al algoritmo *K-means*. Sin embargo, el algoritmo *K-means* es más rápido por factores de 500 a 2500 (en una partición de 60 datos en 5 *clusters*). GA es el más rápido en encontrar la mejor solución, y le siguen TS y SA. Por su parte, SA es el más rápido en converger, le siguen TS y GA.

El enfoque evolutivo sobresale debido a que puede encontrar soluciones de funciones objetivos discontinuas. Por otro lado, los algoritmos de redes neuronales (ANNs) no son Algoritmos de búsqueda, sino que de aprendizaje topológico. Los algoritmos ANNs y el *Genetic Algorithm* (GA) son inherentemente paralelizables, por lo que utilizando un *clúster* de computadores se puede mejorar su velocidad.

En los dos estudios mencionados anteriormente, la cantidad de datos a analizar fue pequeña, menor a 200 objetos. En general los enfoques evolutivos son buenos solo si la cantidad de datos es inferior a 1000 y con objetos con pocas dimensiones. Solo *K-means* y su ANN equivalente, *Kohonen Maps* [122], han sido aplicados a grandes conjuntos de datos. Esto se debe a que es difícil obtener parámetros adecuados para ANNs, GAs, TS y SA para grandes conjuntos de datos, debido a que los tiempos de ejecución son elevados.

### **3.4 Enfoque Escogido para la Predicción**

En las secciones anteriores se ha introducido los conceptos de cromatografía de interacción hidrofóbica y los sistemas de dos fases acuosas, y se ha dado algunos ejemplos de modelos propuestos para la predicción del comportamiento de proteínas en dichas técnicas de separación. A partir del resumen, es posible ver que para ambos casos existe más de un enfoque posible para mejorar las predicciones obtenidas hasta ahora. Con el objetivo de mejorar el poder predictivo, a partir de un mínimo de información, se puede utilizar el concepto de hidrofobicidad superficial media [6] o las aproximaciones propuestas por Salgado *et. al.* [13]; y los modelos cuadráticos y logarítmicos propuestos por Lienqueo *et. al.* [54] y Asenjo *et. al.* [23] respectivamente.

En todos estos modelos, una de las principales limitantes es la escala de hidrofobicidad o, en términos generales, la escala o vector de propiedad aminoacídica (APV). A la fecha se ha reportado estudios basados en 74 APVs para HIC [9] y 36 [4] para ATPS, obteniéndose los mejores resultados con distintas escalas. La existencia de más de una

escala de hidrofobicidad se debe a que no hay una definición universalmente aceptada y única de hidrofobicidad; así como tampoco existe una metodología estándar y universal para la medición de ésta. Lo anterior se debe a que la hidrofobicidad no es un fenómeno fisicoquímico claro, específico, ni medible de una molécula, y depende fuertemente de su interacción con el medio y de las características del medio mismo. Esto explica que el valor de hidrofobicidad de una molécula varíe cuando varían las condiciones del sistema.

A raíz de lo explicado, se ha generado una gran cantidad de estudios independientes que obtienen escalas de hidrofobicidad, utilizando distintas metodologías. En general, las escalas de origen experimental permiten mejores predicciones al caracterizar el comportamiento de aminoácidos, péptidos y proteínas; en los procesos físico-químicos y condiciones específicas mediante los cuales fueron generados (si así fuera el caso); pero en las predicciones obtenidas a partir de modelos que las utilizan aumentan las desviaciones (respecto al valor real) cuando cambian las condiciones del sistema [130]. Además, no todas las escalas se han generado mediante mediciones en sistemas físico-químicos, algunas de éstas se basan en: estudios de la configuración espacial de los aminoácidos en proteínas; la pérdida de acceso a solvente durante el plegamiento de proteínas; estudios que combinan estadísticas con experimentación; o estudios netamente estadísticos.

Debido a la gran diversidad de escalas, en la presente tesis se emplea el grupo de 74 escalas de propiedades aminoacídicas (APVs) utilizados anteriormente por Salgado *et al.* [9]. Las escalas de propiedades aminoacídicas incluyen escalas hidrofóbicas, conformacionales y estadísticas. Sobre los APVs se hace un extenso análisis utilizando un grupo de herramientas de *clustering* o clasificación no supervisada, representativo del estado del arte en esta materia, con el objetivo de generar nuevas escalas capaces de mejorar significativamente los modelos existentes en la actualidad, de manera de generar un precedente del poder de estas herramientas, así como algunas indicaciones para un uso posterior efectivo y eficiente en recursos y tiempo.

### **3.5 Elección de los Algoritmos a Utilizar**

En el estudio realizado previamente en la sección 3.3, se encontró que todos los algoritmos de *clustering* tienen ventajas y desventajas, las que se reflejan en distintos(as) tiempos de ejecución, requerimientos computacionales, capacidad de convergencia a mínimos locales o globales, tipo de *clusters* que son capaces de encontrar, capacidades para utilizar y representar información topológica, niveles de complejidad en su implementación (programación), y niveles de complejidad en el ajuste de sus parámetros, entre otros.

Por otra parte, el conjunto de datos que se utiliza en esta tesis consta de 74 vectores de propiedades aminoacídicas (APVs), para los cuales se desconoce su topología y las clases de *clusters* que poseen. Con estos dos antecedentes, se concluye que es necesario utilizar un conjunto de algoritmos de *clustering* que sea representativo del estado del arte, para así evitar eventuales sesgos.

Desde un punto de vista de la capacidad de representar la topología, probablemente los más avanzados son los Algoritmos de Redes Neuronales (ANNs), aunque son

complejos de implementar y utilizar. Además, los *clusters* se crean sobre las neuronas, y no directamente sobre los datos, lo que hace difícil rastrear los *clusters* originales. Sin embargo, en esta tesis las neuronas pueden ser utilizadas directamente sobre la generación de nuevos modelos, ya que corresponden a una recombinación de la información topológica de los 74 APVs [16]. Como antecedente, Salgado *et. al.* [9] utilizó los algoritmos *K-means* y *Self-Organizing Maps* (SOM) para crear nuevas escalas que conducen a mejores modelos, siendo el algoritmo SOM el que obtuvo modelos con menor Error de Jackknife.

El listado de algoritmos escogidos es el siguiente:

**Tabla 3: Algoritmos escogidos para el análisis de los 74 APVs.**

Algoritmo	Sigla	Clase	Tipo de Implementación
Growing Neuronal Gas	GNG	Redes Neuronales	Propia en base a pseudocódigo [16]
Growing Grid	GG	Redes Neuronales	Propia en base a pseudocódigo [126]
Self-Organizing Maps	SOM	Redes Neuronales	Resultados obtenidos de [9, 122]
K-means	-	De Partición	Resultados obtenidos de [9]
Bisection Algorithm	-	De Partición	Software gCluto [78]
Hierarchical Algorithm	HA	Aglomerativo –Jerárquico	Biblioteca de Matlab [131]
Markov Clustering Algorithm	MCL	Partición en Grafo	Aplicación web [107, 132-135]
Restricted Neighbourhood Search	RNS	Búsqueda Optimización Global en Grafo	Aplicación web [106, 107, 134, 135]
Genetic Algorithm	GA	Optimización	Biblioteca de Matlab [131]

En la Tabla 3 se puede ver que se escoge algoritmos de todas las clases propuestas, según se puede ver en la sección Clasificación de los Algoritmos de *Clustering* (3.3.3). Además de la diversidad de clases de algoritmos de *clustering*, se utilizan distintos tipos de implementaciones, las que incluyen: implementaciones propias, uso de algoritmos que forman parte de softwares altamente especializados y de índole comercial como Matlab [131], uso de *freeware* altamente desarrollados con interfaz gráfica como gCluto [78], y finalmente aplicaciones web de libre acceso [135].

Adicionalmente, se incluye el uso del *Genetic Algorithm* (GA) como un algoritmo puramente de optimización del poder predictivo de los modelos de HIC y ATPS, y no como un algoritmo de *clustering*. Esto tiene como objetivo permitir una comparación de los resultados y sus propiedades.

Finalmente se puede decir que el conjunto de algoritmos escogidos es representativo del estado del arte de la clasificación no supervisada, o algoritmos de *clustering*.

## 4 Objetivos

### 4.1 *Objetivos Generales*

El objetivo de esta tesis es realizar un análisis robusto, mediante un conjunto de técnicas de *clustering*, para generar nuevos vectores de hidrofobicidad, o vectores de propiedades aminoacídicas (APVs), que permitan mejorar modelos del tiempo de retención adimensional (DRT) para HIC y del coeficiente de partición (K) para doce sistemas ATPS.

### 4.2 *Objetivos Específicos*

- Seleccionar un grupo de algoritmos de clustering que representen el estado del arte en el área de clasificación no supervisada, y que utilicen distintas metodologías, para así poder aprovechar las ventajas de cada una.
- Seleccionar un conjunto de implementaciones de los algoritmos de clustering a través de la búsqueda de herramientas disponibles, y generar otras implementaciones propias, en los casos necesarios.
- Generar modelos para el cálculo del área superficial hidrofóbica (ASH), a partir de un análisis de clustering de un conjunto de 74 vectores de propiedades aminoacídicas (APVs), utilizando solamente la composición aminoacídica de las proteínas.
- Generar modelos para el cálculo del DRT en HIC y del coeficiente de partición de proteínas en ATPS, utilizando los modelos para el cálculo de ASH.
- Analizar el fenómeno de hidrofobicidad, desde el punto de vista de los fenómenos fisicoquímicos y propiedades de las proteínas que más aportan para mejorar las predicciones del comportamiento de proteínas en HIC y ATPS, a partir de los estudios que generan los APVs asociados a las mejores escalas.
- Determinar cómo la información provista por los algoritmos de clustering es capaz de describir los distintos aspectos en la complejidad del fenómeno de hidrofobicidad, complementando la información particular asociada a cada sistema experimental.

## 5 Hipótesis

Los modelos existentes para la predicción del tiempo de retención adimensional en HIC y el coeficiente de partición en ATPS pueden ser mejorados utilizando la información de alto nivel subyacente en la estructura topológica de las escalas de propiedades aminoacídicas disponible, la cual puede ser recuperada mediante herramientas de *clustering*.

## 6 Metodología

Como se vio en el capítulo 3, el tipo de modelo propuesto por Lienqueo *et. al.* tiene tres debilidades importantes: requiere la estructura tridimensional de las proteínas, no

considera la distribución de la hidrofobicidad en la superficie, y depende de la elección de una escala de hidrofobicidad, sobre las cuales aún no hay consenso.

Mahn *et. al.*, entre otros, han trabajado para incluir la distribución de la hidrofobicidad en las predicciones [5, 35, 136]. Salgado *et. al.*, a través de sus modelos eliminan la dependencia de la estructura tridimensional con buenos resultados, y también da un primer paso en el problema de elección de una escala de hidrofobicidad en una HIC, al analizar 74 escalas/vectores de propiedades aminoacídicas (APVs) mediante dos técnicas de *clustering* (SOM y *K-means*) [9].

En este capítulo se describe las metodologías utilizadas para extender el trabajo de Salgado *et. al.* a los sistemas ATPS, e incorporar el uso de 6 algoritmos de *clustering* y uno de optimización para los dos tipos de sistemas (conjunto de técnicas representativas del estado del arte en la clasificación no supervisada de datos, según se vio en la sección 3.3).

Las metodologías descritas a continuación se dividen en cinco secciones principales:

- Modelos utilizados (sección 6.1)
- Escalas de propiedades aminoacídicas (sección 6.2)
- Evaluación de los modelos (sección 6.3)
- Extracción de información topológica para obtener nuevos APVs (sección 6.4)
- Búsqueda de escalas precursoras (de las nuevas generadas, sección 6.5)

A continuación en la Figura 16 se presenta un diagrama general de la metodología empleada para generar los modelos predictivos del coeficiente de partición y el tiempo de retención, la que es detallada en las secciones siguientes.

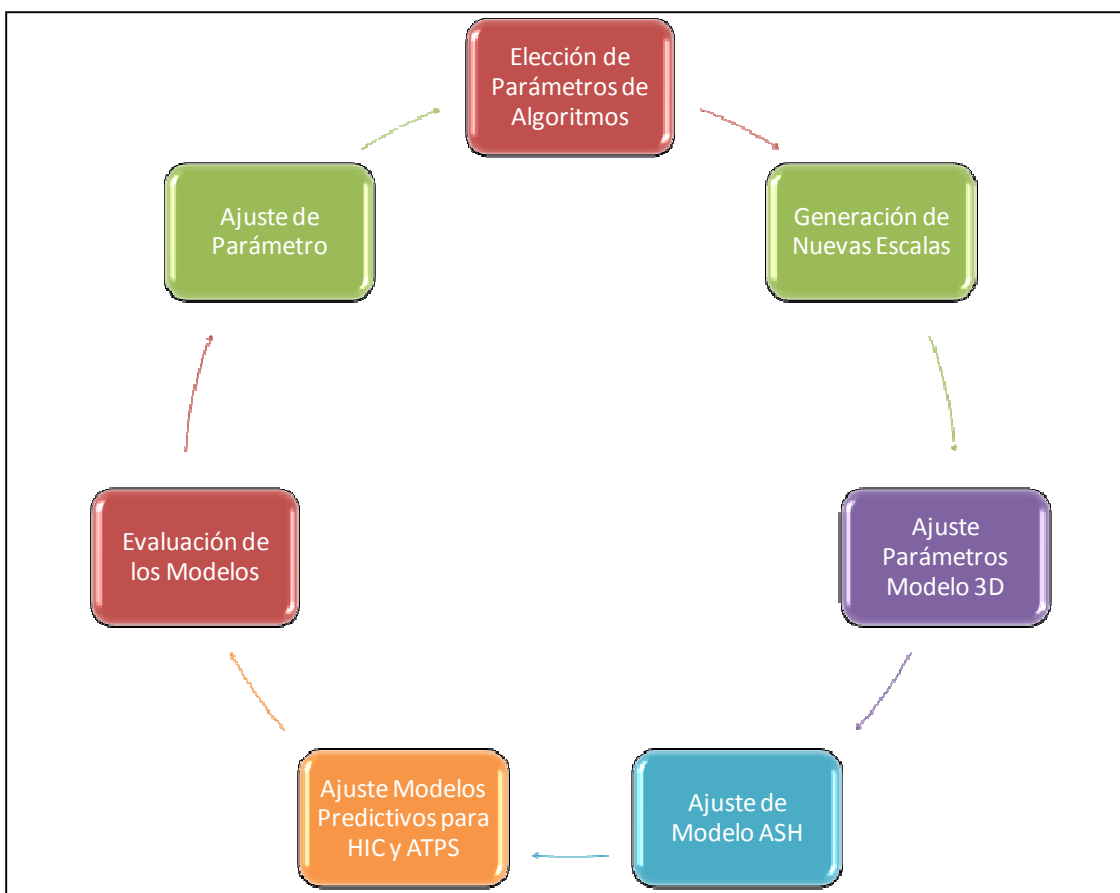


Figura 16: Esquema general de la metodología para la construcción de modelos.

## 6.1 Modelos Utilizados

### 6.1.1 Predicción del Tiempo de Retención Adimensional en HIC

#### Proteínas

Se utilizó la información del mismo conjunto de veinte proteínas utilizado anteriormente por Lienqueo *et. al.* y Salgado *et. al.* [9, 13, 54], para las cuales se conoce la estructura tridimensional y el tiempo de retención [54] en un mismo sistema cromatográfico que explota las diferencias de hidrofobicidad para obtener la separación. Éstas son: Citocromo C (1HRC), Ribonucleasa A (1AFU), Mioglobina (1YMB), Conalbúmina (1OVT), Ovoalbúmina (1OVA), Lisozima (2LYM), Taumatina (1THV), Quimotripsinógeno A (2CHA),  $\beta$ -Lactoglobulina (1CJ5),  $\alpha$ -Amilasa (1BLI),  $\alpha$ -Quimotripsina (4CHA),  $\alpha$ -Lactalbúmina (1A4V).

#### **Cromatografía de Interacción Hidrofóbica (HIC) y Tiempo de Retención Adimensional**

Los tiempos de retención son los determinados por Lienqueo *et. al.* [54] en HIC, con una columna de 1 ml *Phenyl-Sepharose Fast Flow* con sulfato de amonio 2 M en buffer 20 mM Bis-Tris pH 7.0 con 2 M sulfato de amonio/ 20 mM Bis-Tris pH 7.0 como eluyente (ver Anexo I).



Para cada proteína, el tiempo de retención se transformó para obtener el tiempo de retención adimensional, mediante la siguiente expresión:

$$DRT = \frac{t_R - t_0}{t_f - t_0} \quad (13)$$

Donde  $t_R$  corresponde al tiempo para el cual se obtiene el *peak* de la cromatografía,  $t_0$  es el tiempo en el que comienza el gradiente de sal, y  $t_f$  es el tiempo en el que termina el gradiente de sal.

### **Modelo DRT 0**

Según Lienqueo *et. al.* [54], el tiempo de retención adimensional (DRT) de una proteína corresponde a un polinomio de orden 2 con respecto al área superficial media asociada a una propiedad (ASP), por ejemplo, la hidrofobicidad superficial media (ASH), según se muestra en la siguiente ecuación.

$$DRT = b_0 + b_1\Gamma + b_2\Gamma^2 \quad (14)$$

Donde  $\Gamma$  es el ASP de la proteínas y  $b_i$  son los coeficientes del modelo cuadrático que se ajustan mediante el método de los mínimos cuadrados [54]. La  $\Gamma$  se calcula asumiendo que cada aminoácido en la superficie de la proteína contribuye a la propiedad de forma proporcional a su abundancia en la superficie [53]. El cálculo se realiza mediante la ecuación:

$$\Gamma = \sum_{i \in A} \hat{r}_i \varphi_i \quad (15)$$

Donde  $A$  es el conjunto de 20 aminoácidos posibles,  $\varphi_i$  es el componente número  $i$  de un vector de propiedad aminoacídica (APV),  $\hat{r}_i$  representa la fracción de la superficie proteica ocupado por el aminoácido de clase  $i$ , el que se calcula según la siguiente expresión:

$$\hat{r}_i = \frac{S_i}{\sum_{j \in A} S_j} \quad (16)$$

Donde  $S_i$  es la suma del área superficial accesible (ASA) para todo aminoácido de clase  $i$ . Los valores del ASA fueron calculados previamente por Salgado *et. al.* utilizando el software STRIDE a partir de la estructura tridimensional de las proteínas [13, 137].

### **Modelo DRT I**

En este caso el modelo no se basa en la estructura tridimensional de la proteína cuyo DRT se quiere predecir, sino que sólo se requiere su composición aminoacídica. La propiedad superficial media (ASP o  $\Gamma$ ) se estima mediante la siguiente expresión:

$$\Gamma^I = c_0 + \sum_{i=1}^{20} c_i \hat{a}_i^I + c_{21} \hat{l} \quad (17)$$

Donde  $c_i$  corresponde a los parámetros del modelo lineal obtenidos por el procedimiento de mínimos cuadrados,  $\hat{l}$  es la razón entre la longitud de la proteína y el máximo largo observado en la base de datos de trabajo, y  $\hat{a}_i^I$  es la fracción del área correspondiente al aminoácido de tipo  $i$  dentro de la máxima área superficial accesible

(ASA), cuando los aminoácidos están totalmente expuestos [9, 13]. El término  $\hat{a}_i^I$  se obtiene por medio de la siguiente ecuación:

$$\hat{a}_i^I = \frac{n_i S_{max,i}}{\sum_{j \in A} n_j S_{max,j}} \quad (18)$$

Donde  $n_i$  es el número de aminoácido de clase  $i$  en la proteína y  $S_{max,i}$  es la máxima área superficial media posible (ASA), obtenida cuando los aminoácidos se disponen en un tripéptido del tipo G-X-G [138]. Los valores de  $S_{max}$  en  $\text{\AA}^2$  son: 113 (Ala), 241 (Arg), 158 (Asn), 151 (Asp), 140 (Cys), 189 (Gln), 183 (Glu), 85 (Gly), 194 (His), 182 (Ile), 180 (Leu), 211 (Lys), 204 (Met), 218 (Phe), 143 (Pro), 122 (Ser), 146 (Thr), 259 (Trp), 229 (Tyr), 160 (Val) [9, 13].

### **Modelo DRT II**

A diferencia del DRT I, este modelo incorpora un factor de corrección que considera la tendencia general de cada aminoácido a estar expuesto al solvente. Este factor  $\alpha$  se calcula como la probabilidad de que un aminoácido tenga un RASA superior a  $\mu=0,6$ . Donde el RASA de un aminoácido  $k$  se define como la razón entre su área superficial media (ASA o  $S_k$ ) y su máximo ASA ( $S_{max,k}$ ). Entonces las nuevas ecuaciones del modelo DRT II son:

$$\Gamma^{II} = c_0 + \sum_{i=1}^{20} c_i \hat{a}_i^{II} + c_{21} \hat{l} \quad (19)$$

Donde  $\hat{a}_i^{II}$  se calcula como:

$$\hat{a}_i^{II} = \frac{n_i S_{max,i} \alpha_i}{\sum_{j \in A} n_j S_{max,j} \alpha_j} \quad (20)$$

Donde  $\alpha_i$  es el factor de exposición al solvente de cada aminoácido de clase  $i$  [13].

### **Modelo DRT III**

Finalmente, en el modelo DRT III se establece una relación lineal entre el ASA ( $S_i$ ) para todos los aminoácidos de clase  $i$ , y el área superficial posible definido por  $n_i S_{max,i}$ . En este caso las nuevas ecuaciones son:

$$\Gamma^{III} = c_0 + \sum_{i=1}^{20} c_i \hat{a}_i^{III} + c_{21} \hat{l} \quad (21)$$

Donde  $\hat{a}_i^{III}$  se calcula según la siguiente expresión:

$$\hat{a}_i^{III} = \frac{n_i S_{max,i} \beta_i + \eta_i}{\sum_{j \in A} (n_j S_{max,j} \beta_j + \eta_j)} \quad (22)$$

Donde  $\beta_i$  y  $\eta_i$  son los coeficientes del modelo lineal entre el  $S_i$  y  $S_{max,i}$  calculados para cada uno de los aminoácidos de clase  $i$  presentes en la base de datos utilizando el procedimiento de los mínimos cuadrados [13].

## **Ajuste de la ASP y Base de Datos Utilizada**

En los modelos DRT I, DRT II y DRT III, la suma de los coeficientes  $\hat{a}_i$  es por definición uno, por lo que forman un sistema lineal dependiente. Esto permite no considerar la información de los aminoácidos con  $\varphi_i$  igual a cero.

Para determinar los coeficientes  $c_i$  de las ecuaciones 16, 18 y 20 se utilizó el procedimiento de los mínimos cuadrados sobre una base de datos de 1982 proteínas con estructura tridimensional conocida, publicada por Hohhm *et. al* [139], y utilizada previamente por Salgado *et. al.* [13]. Esta base de datos se obtuvo a partir de la selección de proteínas no redundantes (*identity cut-off* de 25%) y eliminando las proteínas de membrana, y La estructura tridimensional se obtuvo a partir de la base de datos *Protein Data Base* (PDB) [140, 141]. Los  $c_i$  se calcularon como el promedio al realizar 100 repeticiones, utilizando distintos subconjuntos de proteínas determinados al azar.

### **6.1.2 Predicción del Coeficiente de Partición en ATPS**

Se utilizó un conjunto de 11 proteínas con coeficiente de partición conocido [22, 24] (ver Anexo J). Este grupo de proteínas ha sido utilizado previamente por Salgado *et. al.* para modelar el coeficiente de partición a partir de la composición aminoacídica de 37 escalas de hidrofobicidad [4]. Las proteínas seleccionadas son aquellas con estructura conocida y disponible en la base de datos PDB [140, 141] o que comparten una identidad de secuencia mayor al 93% con una proteína en la base de datos PDB. Las proteínas escogidas son:  $\alpha$ -Amilasa (1E40),  $\alpha$ -Quimotripsinógeno A (2CGA),  $\alpha$ -Lactalbúmina (1F6S), Amiloglucosidasa (3GLY), Conalbúmina (1OVT), Lisozima (2LYM), Ovalbúmina (1OVA), Subtilisina (1SBC), Taumatina (1THV) e Inhibidor de Tripsina (1AVU). También se consideró Albúmina de suero de bovino, cuya estructura tridimensional no estaba disponible en PDB, por lo que se utilizó la estructura facilitada por Carredano *et. al.* [142].

### **Sistemas de Dos Fases Acuosa y Coeficientes de Partición**

Los Coeficientes de Partición utilizados en esta tesis son los reportados por Schmidt [22, 24] para cuatro sistemas de dos fases acuosas compuestas por polietilenglicol (PEG) 4000: fosfato, sulfato, citrato o dextrano. Además para cada sistema se consideró tres niveles de concentración de NaCl: sin sal (0,0% p/p), baja concentración de sal (0,6% p/p), y alta concentración de sal (8,8% p/p). Los Coeficientes de Partición se muestran en el Anexo J.

La solución de fosfato es una mezcla de  $K_2HPO_4$  y  $NaH_2PO_4$  a pH 7. En el resto de los sistemas se controló el pH con ácido cítrico (sistema citrato) e hidróxido de sodio (sistemas PEG+Sulfato y PEG+Dextrano). La temperatura durante la preparación y partición de los sistemas se mantuvo controlada a 20 [°C].

La distancia entre el punto que representa la composición del sistema y la curva binoidal es idéntica para todos los sistemas, y el cociente de volúmenes entre la fase superior e inferior utilizado es 1. Las proteínas se agregaron a los sistemas de manera que la concentración final fuera siempre 1 [g/l] [24].

En esta tesis se utiliza la nomenclatura descrita en la Tabla 4 para presentar los resultados asociados a cada uno de estos sistemas:

**Tabla 4: Nomenclatura utilizada para distinguir los 12 sistemas ATPS utilizados.**

Sistema	NaCl 0,0%	NaCl 0,6%	NaCl 8,8%
<b>PEG-Fosfato</b>	PEG-F 0,0%	PEG-F 0,6%	PEG-F 8,8%
<b>PEG-Sulfato</b>	PEG-S 0,0%	PEG-S 0,6%	PEG-S 8,8%
<b>PEG-Citrato</b>	PEG-C 0,0%	PEG-C 0,6%	PEG-C 8,8%
<b>PEG-Dextrano</b>	PEG-D 0,0%	PEG-D 0,6%	PEG-D 8,8%

### **Modelos Utilizados**

El modelo utilizado para predecir el coeficiente de partición es el de Asenjo *et. al.* [23], y se calcula según la siguiente ecuación:

$$\log K = R \log \Gamma - R \log \Gamma_0 \quad (23)$$

Donde  $K$  y  $\Gamma$  son el coeficiente de partición del sistema ATPS y el área superficial media asociada a una propiedad puntual (por ejemplo hidrofobicidad), respectivamente. El término  $R$  es la resolución hidrofóbica del sistema, que se entiende como la habilidad de éste para discriminar entre proteínas con distinta hidrofobicidad; y  $\log \Gamma_0$  es una propiedad intrínseca del sistema.

Al igual que con HIC, se incluyen tres modelos para el cálculo de los coeficientes de partición a partir de la composición aminoacídica:

$$\log K^I = R \log \Gamma^I - R \log \Gamma_0 \quad (24)$$

$$\log K^{II} = R \log \Gamma^{II} - R \log \Gamma_0 \quad (25)$$

$$\log K^{III} = R \log \Gamma^{III} - R \log \Gamma_0 \quad (26)$$

Donde  $\Gamma^I$ ,  $\Gamma^{II}$  y  $\Gamma^{III}$  se calculan con las ecuaciones 5, 7 y 9 respectivamente [4].

## **6.2 Escalas de Propiedades Aminoacídicas**

Se utilizó un conjunto de 74 vectores de propiedades aminoacídicas reportados en la literatura. Este conjunto incluye un amplio espectro de características físicas, químicas y biológicas de los aminoácidos, entre ellas: peso molecular, volumen molecular, energía libre de transferencia de aminoácidos en octanol/agua, accesibilidad media a solvente, preferencias por estructuras secundarias, tiempo de retención en HIC, coeficiente de partición en ATPS, número de codones, entre otros [47, 130, 143-182].

Todos los APVs utilizados fueron normalizados y escalados al intervalo [0,1]. La normalización permite expresar los coeficientes como la variación sobre la media y respecto a la desviación estándar de cada escala. Por otro lado, el escalamiento

permite que los coeficientes tengan valores en el intervalo [0,1]. Las escalas de hidrofiliidad se transformaron a escalas de hidrofobicidad sustrayendo a 1 el valor de la escala correspondiente a cada aminoácido.

En este estudio se utilizó la misma clasificación de las escalas utilizada por Salgado *et. al.* [13], que distingue entre escalas de origen hidrofóbico, conformacional y estadístico; dependiendo del tipo de estudio donde se generó cada una (ver Anexo K).

### 6.3 Evaluación de los Modelos

El desempeño de los modelos generados se evaluó a través de tres parámetros: la Validación Cruzada de Jacknife, el Coeficiente de Pearson, y el Error Cuadrático Medio. El parámetro que se utilizó para la búsqueda de los mejores modelos es el error de predicción estimado por el método de Jacknife, en adelante Error de Jacknife. Los otros parámetros se calculan solo para los modelos que presentan una reducción del Error de Jacknife, respecto a los modelos de comparación [4, 9]. El Error de Jacknife es una metodología ampliamente conocida [183], y es considerada en la actualidad como una herramienta poderosa para evaluar el poder predictivo de un modelo, para modelos construidos en base a conjuntos de entrenamiento de tamaño reducido [184, 185].

El Error Cuadrático Medio se calculó según la siguiente expresión:

$$MSE_D = \frac{1}{N} \sum_{k=1}^N (D_k - \widehat{D}_k)^2 \quad (27)$$

Donde  $D_k$  y  $\widehat{D}_k$  son el valor experimental y el predicho para la proteína  $k$ ,  $N$  es el número de proteínas con valor experimental conocido. En el caso de HIC,  $D$  corresponde al tiempo de retención adimensional ( $DRT$ ), y en el caso de ATPS es el coeficiente de partición ( $K$ ).

El Coeficiente de Pearson se calculó según la siguiente expresión:

$$Pearson_D = \frac{N \sum_{k=1}^N (D_k \times \widehat{D}_k) - \sum_{k=1}^N D_k \sum_{k=1}^N \widehat{D}_k}{\sqrt{N \sum_{k=1}^N (D_k)^2 - (\sum_{k=1}^N D_k)^2} \sqrt{N \sum_{k=1}^N (\widehat{D}_k)^2 - (\sum_{k=1}^N \widehat{D}_k)^2}} \quad (28)$$

Este parámetro permite determinar la calidad del ajuste de un modelo lineal a los datos, entregando un valor en el intervalo [0,1].

El Error de Jacknife se obtuvo al ajustar el modelo un número de veces igual al número de datos de ajuste, dejando en cada caso un dato fuera. La ecuación utilizada es la siguiente:

$$MSE_{JK} = \frac{1}{N} \sum_{k=1}^N (D_k - \widehat{D}_k^{-k})^2 \quad (29)$$

Donde,  $\widehat{D}_k^{-k}$  es el valor predicho para el dato número  $k$ , utilizando el modelo construido con el sub conjunto de datos que no incluye el dato número  $k$ .

Como se puede ver, la estructura de la ecuación es similar a la del Error Cuadrático Medio, con la diferencia que a través de éste se calcula un error asociado al ajuste del

modelo a los datos, mientras que a través del Error de Jackknife se estima un error de predicción del modelo.

## 6.4 Extracción de Información Topológica Para Obtener Nuevos APVs

Para realizar la extracción de información topológica se utilizó 9 algoritmos distintos: *Growing Neuronal Gas*, *Growing Grid*, *Self-Organizing Maps*, *K-mean*, *Bisection Algorithm*, *Hierarchical Algorithm*, *Markov Clustering Algorithm*, *Restricted Neighborhood Search*, y *Genetic Algorithm*. Dos de estos algoritmos fueron utilizados previamente por Salgado *et. al.* [9]. Para los otros 7, sus descripciones generales se encuentran en la sección 3.3.

### 6.4.1 Un Problema de Optimización

Los modelos que se utilizaron en esta tesis, descritos en las ecuaciones 13, 16, 18, 20, y 22 a 25; dependen de la escala aminoacídica utilizada, por lo tanto, el poder predictivo depende de la escala generada por los algoritmos de *clustering*. La escala obtenida, depende de los parámetros utilizados durante el proceso de extracción (de los APVs) y síntesis de información (formación de la escala) desarrollada por estos algoritmos, por lo tanto, la búsqueda de los parámetros que permite minimizar el error predictivo es un problema de optimización. Por ejemplo, el *Growing Neuronal Gas* depende de 8 parámetros ( $P$ ) dentro de los que se encuentra el máximo número de nodos y arcos. El ajuste de estos parámetros genera un problema de optimización que tiene un espacio de soluciones dado por las posibles combinaciones de los distintos valores que pueden tomar los parámetros  $P$ , y en el cual la función objetivo a minimizar es el Error de Jackknife del modelo (ver sección 6.3).

En el capítulo 3.3 se reportó que no existe una metodología que permita escoger *a priori* el conjunto de parámetros óptimos de un algoritmo, en función de las características de un conjunto de datos. Obtener dichos parámetros es un proceso sumamente complejo que usualmente se resuelve mediante prueba y error, por lo que puede ser un proceso intensivo en el uso de recurso computacional y/o requerir altos tiempos de ejecución. Lo anterior se ilustra en el siguiente ejemplo:

- Consideramos un algoritmo con ocho parámetros,  $P = 8$ . Si se realiza todas las búsquedas que resulten de las distintas combinaciones posibles, al considerar tan solo 10 valores ( $V$ ) para cada parámetro, es necesario realizar un total de cien millones de ejecuciones del algoritmo correspondiente ( $B = 10^8$ ,  $B = V^P$ ).
- Para ilustrar qué significa realizar  $10^8$  ejecuciones, a continuación transforma esta cifra a una relacionada con el tiempo. En un año hay aproximadamente  $T = 525.600$  minutos, por lo tanto, si se quiere terminar en un plazo de un año la búsqueda de los parámetros óptimos de un algoritmo que tiene 8 parámetros y 10 valores posibles para cada uno, cada búsqueda no puede tardar más de  $T/B = 5,25 \cdot 10^{-3}$  minutos, o  $3,15 \cdot 10^{-1}$  segundos.

En la práctica, algunas metodologías de búsqueda de nuevas escalas son relativamente rápidas, como el *K-mean*, y otras pueden tardar varios minutos por ejecución, como el *Growing Grid* [129] (se utiliza un procesador Intel® Core™ 2 Duo CPU T7250 @ 2 GHz - 2 GHz, 2 GB de memoria RAM, bajo un sistema operativo Windows Vista de 32 bits).

Por lo tanto no se puede utilizar ocho parámetros con diez valores cada uno, sobre todo teniendo en cuenta que se buscarán escalas con 7 algoritmos distintos. Además, para cada nueva escala generada es necesario ajustar 52 modelos matemáticos (considerando los diferentes modelos y sistemas experimentales), lo que consume una cantidad significativa de tiempo adicional.

Considerando esta limitante, a continuación se explica cómo se obtuvo nuevas escalas utilizando los algoritmos seleccionados en el capítulo 3.3, y cómo se escogió el conjunto de parámetros para cada caso.

#### 6.4.2 *Growing Neuronal Gas* y *Growing Grid*

Para utilizar los algoritmos *Growing Neuronal Gas* (GNG) [16] y *Growing Grid* (GG) [126] se hizo implementaciones propias en Matlab [131], que fueron evaluadas mediante dos conjuntos de datos bidimensionales diseñados especialmente para ese efecto (ver Anexo L). Trabajar en dos dimensiones permite evaluar visualmente el funcionamiento de los algoritmos durante su ejecución, así como los resultados de éste. Además, el comportamiento visual se contrastó con el exhibido en demos de ejemplo de estos algoritmos [186].

Los algoritmos *Growing Neuronal Gas* y el *Growing Grid* crean neuronas en el hiperespacio que durante la ejecución adquieren la topología de los APVs (datos base). Para la construcción de los modelos descritos en las ecuaciones 13, 16, 18, 20, y 22 a 25; cada una de estas neuronas se consideró como una nueva escala aminoacídica.

El GNG depende de 8 parámetros y el GG depende de 6. En ambos casos existen valores sugeridos por los autores de los algoritmos [16, 126, 186]. Dado que el trabajo que se realizó con los datos bidimensionales no proporciona antecedentes que permitan modificar los parámetros sugeridos en la literatura, inicialmente se utilizó los recomendados por los autores, los que se muestran a continuación en la Tabla 5 y en la Tabla 6:

**Tabla 5: Conjunto inicial de parámetros para el algoritmo Growing Neuronal Gas.**

<b>Parámetro</b>	<b>Valor</b>
Lambda	600
Max. Edge Age	88
Epsilon Winner	0.05
Epsilon Neighbor	6.00E-04
Alpha	0.5
Beta	5.00E-04
Max. Nodes	100
Utility	3

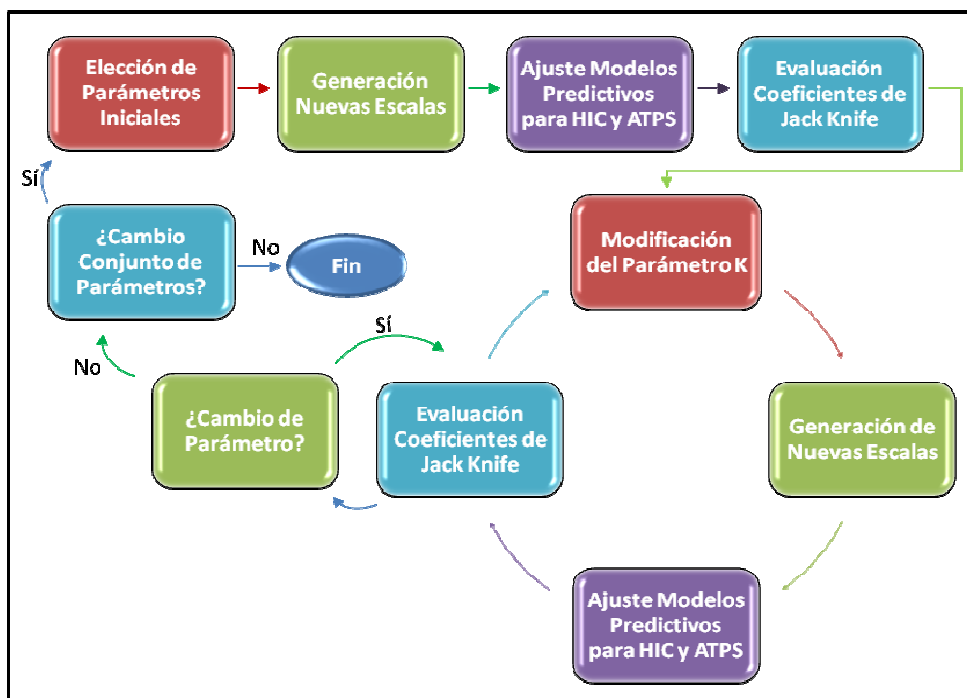
**Tabla 6: Conjunto inicial de parámetros para el algoritmo Growing Grid.**

Parámetro	Valor
Lambda_g	30
Lambda_f	100
Epsilon_i	0.1
Epsilon_f	0.005
Sigma	0.9
Max. Nodes	100

Utilizando los datos bidimensionales generados, se estudió visualmente el comportamiento de cada parámetro, y su influencia sobre el resultado final del algoritmo, sin que se observaran tendencias.

Finalmente, como no es posible probar una gran cantidad de distintas combinaciones de valores de parámetros, el estudio del espacio de optimización de los valores de los parámetros se llevó a cabo mediante un análisis de sensibilidad.

Dado un sistema (físico, químico, de optimización, etc.) que depende de múltiples parámetros, un análisis de sensibilidad consiste básicamente en hacer variar un parámetro del sistema manteniendo los demás fijos, y evaluar cómo se comporta el sistema frente a dichas variaciones. En este caso, la evaluación se realizó sobre la variación del Error de Jackknife (función objetivo) generado con la mejor solución obtenida a partir de los vectores generados por el algoritmo.



**Figura 17: Esquema de análisis de sensibilidad para el ajuste de parámetros.**

En concreto, el procedimiento escogido para la búsqueda de los mejores valores de los parámetros es un proceso iterativo, el que se ilustra en la Figura 17. Durante la primera



iteración se utilizó los valores de la Tabla 5 y la Tabla 6, y se hizo variar un parámetro a la vez en torno al conjunto de valores iniciales. Luego, el conjunto de mejores valores obtenidos para cada parámetro se utilizaron como los valores iniciales para una siguiente iteración. Dado que los algoritmos GNG y GG son estocásticos, la ejecución de los mismos se repitió 5 y 3 veces respectivamente (ver Anexo M). Por otro lado, debido a que cada iteración demora varias semanas para cada modelo, no se procede más allá de la segunda iteración.

### 6.4.3 Repeated Bisection (gCluto)

La implementación del algoritmo *Repeated Bisection* utilizada se encuentra dentro del software gCluto [78]. Esta herramienta está hecha para facilitar el análisis de datos de distinta índole, al reunir distintos tipos de algoritmos de *clustering* y visualización de resultados.

El algoritmo *Repeated Bisection* realiza sucesivas particiones del conjunto de APVs entregando como resultado un conjunto de *clusters*, que son conjuntos de APVs que están relacionados topológicamente [15], ya que son la agrupación que minimiza la proximidad entre APVs pertenecientes a un mismo *cluster* y maximiza la proximidad entre *clusters*. Para utilizar la información topológica en la creación de nuevas escalas se calculó el centroide matemática de cada *cluster*, según la siguiente ecuación.

$$\vec{C} = \frac{\sum_i^n \vec{v}_i}{n} \quad (30)$$

Este algoritmo tiene seis parámetros disponibles, de los cuales cuatro corresponden a coeficientes, y dos a funciones (cambiando de por sí el algoritmo de tipo *Repeated Bisection*). Los parámetros establecidos por defecto por la herramienta son:

Tabla 7: Parámetros por defecto del algoritmo *Repeated Bisection*.

Parámetros	R.B.
Crfun	l2
Simfun	Cosin
Colpurne	1
Clusters	10
Trials	10
Iterations	100

En la Tabla 7 se puede apreciar que este algoritmo tiene un menor número de parámetros que los algoritmos GNG y GG. Aunque este algoritmo es más rápido que el GNG y el GG, esta herramienta no permite automatizar la ejecución del algoritmo para distintos parámetros, por lo que se debió realizar de forma manual. Debido a esto, el proceso de búsqueda del conjunto de parámetros óptimo se realizó mediante un análisis de sensibilidad, según se describe en la sección 6.4.2. En este caso se realizó una sola iteración, debido a que se necesitó grandes variaciones del valor de cada

parámetro para generar una solución diferente; es decir, se obtuvo una gran parte de las soluciones que se pueden generar con este algoritmo (en otras palabras, se detecta una menor capacidad de generar distintas soluciones en comparación con el algoritmo GG y con el algoritmo GNG).

Dentro de las opciones de *clustering* se utilizó todos los criterios de similitud que permite el programa gCluto, las que se pueden ver en la Tabla 2, en la sección 3.3.4.

La función de similitud utilizada es el coseno del ángulo entre vectores, que se calculó según la siguiente expresión:

$$\cos(\vec{u}, \vec{v}) = \frac{\sum_i^n v_i u_i}{\|v\|_2 \|u\|_2} \quad (31)$$

Donde  $n$  es el número de componentes, y  $\vec{u}$  y  $\vec{v}$  son los vectores que representan los objetos que se desea comparar.

#### 6.4.4 Hierarchical Clustering

La implementación del algoritmo utilizado se encuentra dentro del paquete de algoritmos para el análisis de datos del software Matlab [131].

Este algoritmo se basa en la aglomeración sucesiva de los distintos APVs en *clusters*, proceso basado en una medida de proximidad, y por ende en la topología de los APVs. Junto al proceso de aglomeración se generó un dendrograma, que contiene un índice de proximidad entre los APVs de cada *cluster*, a cada nivel jerárquico. Para utilizar la información topológica en la creación de nuevas escalas se calculó el centroide matemática de cada *cluster* generado (ver ecuación número 30).

Existen diversas formas de utilizar este algoritmo:

- A través del *Statistics Toolbox Functions*, una herramienta orientada a usuario que permite realizar el *clustering* con gran facilidad, pero que tiene las mismas limitaciones que gCluto, ya que no se puede automatizar.
- A través de la función  $T = clusterdata(X, cutoff)$ , que permite sintetizar todos los pasos del algoritmo. La función requiere como entradas los datos  $X$  y el *cutoff* (ver sección 3.3.4).
- Finalmente, existe la opción de desagregar la función *clusterdata* en funciones más básicas, que permiten escoger con mayor detalle las especificaciones con que se quiere realizar el *clustering*, estas son:  $Y = pdist(X, metric)$ ,  $Z = linkage(Y, method)$ ,  $Y = inconsistent(Z)$ ,  $T = cluster(Z, 'cutoff', c)$ ,  $c = cophenet(Z, Y)$ . Esta última es la metodología empleada.

#### Ajuste de los Parámetros y Condiciones Iniciales

Las cinco funciones de Matlab utilizadas tienen en total tres parámetros a considerar: *metric*, *method*, y *'cutoff'*. *Metric* corresponde a la métrica de proximidad entre APVs, y existen once opciones distintas, de las cuales la distancia euclidiana es la predefinida;

sin embargo, se encontró en base al *Cophenetic Correlation Coefficient* (ver siguiente sección) que la mejor métrica para los datos a analizar es la distancia *City Block*. A continuación se describe ambas métricas.

Dada una matriz de datos  $X_{n \times m}$ , que se trata como un vector fila ( $1 \times m$ )  $x_1, x_2, \dots, x_m$ ; las distintas distancias entre el vector  $x_r$  y  $x_s$  están definidas como sigue:

- Distancia Euclidiana:  $d_{rs}^2 = (x_r - x_s)(x_r - x_s)'$  (32). La comilla denota el complejo conjugado.
- Distancia *City Block*:  $d_{rs} = \sum_{j=1}^n |x_{rj} - x_{sj}|$  (33).

Existen siete criterios disponibles dentro del parámetro *methods*, que definen distintas formas de medir la distancia entre dos *clusters* (grupo de APVs). El criterio predefinido es el *Single Linkage*, pero en base a las recomendaciones encontradas en la literatura se decidió utilizar el criterio *Average* [86]. A continuación se define las dos metodologías nombradas: Si  $n_r$  es el número de objetos en el *cluster*  $\mathcal{R}$ ,  $n_s$  es el número de objetos en el *cluster*  $\mathcal{S}$ , y  $x_{ri}$  es el objeto número  $i$  en el *cluster*  $\mathcal{R}$ , entonces los métodos se pueden definir de la siguiente manera:

- *Single linkage* (nearest neighbor): utiliza la menor distancia entre los objetos de dos *clusters*.

$$d(r, s) = \min \left( \text{dist}(x_{ri}, x_{sj}) \right), i \in (1, \dots, n_r), j \in (1, \dots, n_s) \quad (34)$$

- *Average linkage*: utiliza la distancia media entre todos los pares de objetos del *cluster*  $r$  y el *cluster*  $s$ .

$$d(r, s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} \text{dist}(x_{ri}, x_{sj}) \quad (35)$$

El parámetro '*cutoff*' indica que se utilizará un valor de inconsistencia corte para determinar los *clusters*. El parámetro  $c$  es el valor de inconsistencia utilizado para dividir  $Z$  en *clusters*. Cuando un link tiene un valor de inconsistencia igual o menor a  $c$ , el link se destruye y se forman dos *clusters*.

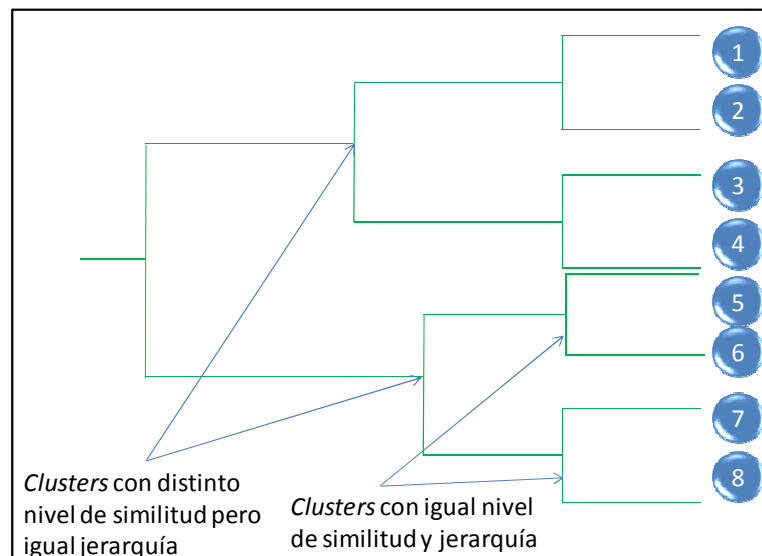
Como se explicó, los parámetros del algoritmo *Hierarchical Clustering* se han ajustado sin necesidad de un proceso de optimización o análisis de sensibilidad. La metodología utilizada requiere el uso coordinado de las cinco funciones:  $Y = \text{pdist}(X, \text{metric})$ ,  $Z = \text{linkage}(Y, \text{method})$ ,  $Y = \text{incosistent}(Z)$ ,  $T = \text{cluster}(Z, \text{'cutoff'}, c)$ ,  $c = \text{cophenet}(Z, Y)$ ; las que se explican con mayor detalle a continuación.

### **Funciones de Matlab Utilizadas**

- $Y = \text{pdist}(X, \text{metric})$ : función que calcula la matriz de distancia entre cada par de vectores en base a la métrica especificada. Se dispone de once métricas distintas para escoger, de las cuales la distancia euclidiana es la predefinida.
- $Z = \text{linkage}(Y, \text{method})$ : función que agrupa los distintos vectores mediante una heurística definida por la opción *method*, y que realiza varias iteraciones

en las que agrupa los datos hasta generar un único y gran grupo de vectores, lo que gráficamente se puede visualizar como un dendrograma (ver sección 3.3.4 o Figura 18). Esta función trabaja sobre una matriz de distancia dada a través de un vector  $\vec{Y}$  que guarda la información de distancias entre cada par de datos, sobre el cual la función *linkage* puede trabajar reconociendo a qué par de datos corresponde cada distancia. Existen siete criterios disponibles dentro del parámetro *methods*, que definen distintas formas de medir la distancia entre dos *clusters* (grupo de vectores).

- $Y = inconsistent(Z)$ : esta función calcula el Coeficiente de Inconsistencia para cada link en el árbol jerárquico de los *clusters*  $Z$ , donde  $Z$  es una matriz  $(m - 1) \cdot 3$  generada por la función *linkage*. El Coeficiente de Inconsistencia caracteriza cada link en el árbol, comparando su distancia con la distancia promedio de los otros links al mismo nivel de jerarquía. Entre mayor sea este coeficiente, existe menor similitud entre los objetos conectados por el link en cuestión. En la Figura 18 se aprecia cuatro niveles de similitud, pero solo tres niveles de jerarquía de aglomeración de objetos en *clusters*.



**Figura 18: Esquema dendrograma e identificación de posibles clusters con iguales y distintos niveles de similitud o jerarquía. Entre antes se produce una agrupación de derecha a izquierda, más similares son los objetos.**

- $T = cluster(Z, 'cutoff', c)$ : esta función construye los *clusters* a partir del árbol jerárquico de *clusters*  $Z$ , generado por la función *linkage*.  $Z$  es una matriz de orden  $(m - 1) \times 3$ , donde  $m$  es el número de datos originales. El parámetro *'cutoff'* indica que se utilizará un valor de inconsistencia corte para determinar los *clusters*. El parámetro  $c$  es el valor de inconsistencia utilizado para dividir  $Z$  en *clusters*. Cuando un link tiene un valor de inconsistencia igual o menor a  $c$ , el link se destruye y se forma dos *clusters*. El resultado de la función  $T$  es un vector de tamaño  $m$  que contiene el número del *cluster* al cual pertenece cada uno de los datos originales (cada APV).

- $c = cophenet(Z, Y)$ : esta función calcula el *Cophenetic Correlation Coefficient* para el árbol jerárquico de *clusters* representado por  $Z$ , que es el resultado de la función *linkage*.  $Y$  contiene las distancias o disimilitudes usadas para construir  $Z$ , como resultado de la función *pdist*.  $Z$  es una matriz de tamaño  $(m - 1) \times 3$ , con la información de las distancias en la tercera columna.  $Y$  es un vector de tamaño  $m \times (m - 1)/2$ .

La altura es la distancia entre los dos sub *clusters* que están unidos por el *link*. El valor resultado,  $c$ , es el *Cophenetic Correlation Coefficient*. La magnitud de este valor es muy cercana a la unidad para una solución de alta calidad. Esta medida puede ser utilizada para comparar soluciones alternativas obtenidas mediante distintos algoritmos.

La *Cophenetic Correlation Coefficient* entre  $Z(:,3)$  (tercera columna de  $Z$  en lenguaje de Matlab) e  $Y$  se define como:

$$c = \frac{\sum_{i < j} (Y_{ij} - \bar{y})(Z_{ij} - \bar{z})}{\sqrt{\sum_{i < j} (Y_{ij} - \bar{y})^2 \sum_{i < j} (Z_{ij} - \bar{z})^2}} \quad (36)$$

Donde:

$Y_{ij}$  es la distancia entre los objetos  $i$  y  $j$  en  $Y$ .

$Z_{ij}$  es la Cophenetic Distance entre los objetos  $i$  y  $j$ , obtenida de  $Z(:,3)$ .

$\bar{y}$  y  $\bar{z}$  son los promedios de  $Y$  y  $Z(:,3)$ , respectivamente.

#### 6.4.5 Genetic Algorithm

La implementación del *Genetic Algorithm* utilizado se encuentra dentro del paquete de algoritmos para el análisis de datos que posee el software Matlab [131].

Este algoritmo puede utilizar dos metodologías de recombinación de los APVs para generar otros nuevos (ver sección 3.3.4). A través de una función que simula el proceso de recombinación genética denominado *crossover*, utilizando dos APVs como genes, y a través de una función que simula una mutación genética en un APV (esta última no fue utilizada). A la fecha no se ha reportado que estas metodologías trabajen preservando información topológica [95-97, 109, 112, 118, 119].

Este algoritmo se implementó en un archivo con código ejecutable (*script*) utilizando la

La implementación del *Genetic Algorithm* se realizó a través de la siguiente función de Matlab:

$$[x \text{ fval}] = ga(@fitnessfun, nvars, A, b, Aeq, beq, LB, UB, nonlcon, options) \quad (37)$$

Donde:

*Fitnessfun* : Función objetivo.

*Nvars* : Dimensión de los vectores.

*Options* : Opciones escogidas utilizando *gaoptimset*.

*A* : Matriz *A* para restricciones del tipo inecuaciones.

- $B$  : Vector  $B$  para restricciones del tipo inecuaciones.
- $A_{eq}$  : Matriz  $A$  para restricciones del tipo ecuaciones.
- $b_{eq}$  : Vector  $B$  para restricciones del tipo ecuaciones.
- $LB$  : Límite inferior de  $x$ .
- $UB$  : Límite superior de  $x$ .
- $nonlcon$  : Función de restricciones no lineales.
- $randstate$  : Campo opcional para resetear la variable  $rand$ .
- $randnstate$  : Campo opcional para resetear la variable  $randn$ .

Para modificar las opciones se utilizó la función  $options = gaoptimset$ , y luego la estructura:

$$options.[parámetro] = [valor, condición inicia o función] \quad (38)$$

Por ejemplo, para modificar el parámetro  $CrossoverFraction$  de 0,8000 a 1, se procede de la siguiente manera.

$$options.CrossoverFraction = 1 \quad (39)$$

### Ajuste de los Parámetros y Condiciones Iniciales

Para lograr un adecuado funcionamiento del *Genetic Algorithm* se modificó los parámetros relacionados con el criterio de parada del algoritmo. Los parámetros modificados se encuentran resumidos en la Tabla 8.

Primero se definió la primera generación, que corresponde a los 74 APVs, cambiando el valor inicial de los parámetros:  $InitialPopulation$  e  $InitialScores$ .

Durante las primeras pruebas, el algoritmo calculó solamente dos a tres generaciones, pese a que el límite de generaciones utilizado era 100. A raíz de esto se modificó el resto de los parámetros relacionados con los criterios de parada del algoritmo, estos son:  $Generations$ ,  $StallGenLimit$  y  $StallTimeLimit$ . De esta manera se estableció la detención a través de un criterio de convergencia adecuado (ver Tabla 8).

**Tabla 8: Parámetros del *Genetic Algorithm* con el valor predeterminado cambiado.**

Parámetro	Descripción	Valor Predeterminado	Valor Usado
Generations	Número entero que especifica el máximo número de interacciones antes que el algoritmo se detenga.	100	50
StallGenLimit	Máximo número de generaciones consecutivas permitidas sin que haya una mejora en la función objetivo.	50	150
StallTimeLimit	Máximo tiempo (en segundos) permitido sin que haya una mejora en la función objetivo.	20	900
InitialPopulation	Población inicial usada como semilla (solución inicial) para el Genetic Algorithm.	[]	Datos
InitialScores	Valores iniciales que toma la función objetivo para cada integrante de la población inicial.	[]	$MSE_{jk}(datos)$
CrossoverFraction	Fracción de la población de la próxima generación, sin incluir a los hijos elite, es creada mediante la operación crossover.	0,8	1

Uno de los principales objetivos del presente trabajo es la búsqueda de nuevas escalas que optimicen el poder predictivo de los modelos, en cuanto al comportamiento de las proteínas en HIC y ATPS, mediante un análisis topológico de escalas preexistentes y con base en la literatura. Sin embargo, el *Genetic Algorithm* es un algoritmo global de optimización estocástica, y no se ha reportado que utilice la topología de los datos, a diferencia de los algoritmos de *clustering*.

Si consideramos el procedimiento utilizado por el *Genetic Algorithm* (3.3.4), se puede apreciar que hay un factor de azar en la operación mutación que permite recorrer todo el espacio de posibles soluciones. Para restringir el campo de acción del algoritmo a la recombinación de la información contenida en los 74 APVs, se modificó el valor del parámetro *crossover* de 0.8 a 1 (el 1 significa que el 100% de los padres “no elite” se generan con la función *crossover*).

De esta manera, las nuevas generaciones se crearon únicamente a través de dos procedimientos: la elección de dos vectores (APVs o nuevas escalas) de padres que corresponden a los de mejores resultados (elite); y mediante una recombinación de los padres (función *crossover*). De esta manera las componentes de las nuevas soluciones (nuevas escalas) obtenidas son siempre las mismas, y en términos generales el algoritmo sólo recombina la información entregada por los APVs originales.

#### **6.4.6 *Restricted Neighborhood Search (RNSC) y Markov Clustering Algorithm (MCL)***

Las implementaciones de los algoritmos *Restricted Neighbourhood Search* y *Markov Clustering* que se utilizó se encuentran en la plataforma *Network Analysis Tool* (NeAT) [134, 135]. Esta es una plataforma *on-line* que reúne algunas herramientas de análisis estadístico y de *clusters* para grafos. A modo de ejemplo, en la Figura 19 se muestra la interfaz para el método MCL (no se muestra la interfaz para el caso del RNSC, dado que es muy similar).

**NeA-tools - MCL**

Fast and scalable unsupervised cluster algorithm for graphs based on simulation of (stochastic) flow in graphs.  
 The MCL program was developed by Stijn van Dongen (Van Dongen, 2000, PhD thesis; Enright et al, 2002)  
 The stand-alone version of MCL is available at <http://micans.org/mcl/>

Input format: tab-delimited format

Graph

Upload graph from file :  Browse...

Column specification (only relevant for tab-delimited input)

Source node:

Target node:

Weight or label column:

Inflation value: 1.8

GO [RESET](#) [DEMO](#) [MANUAL](#) [TUTORIAL](#) [MAIL](#)

**Figura 19: Interfaz gráfica del Markov Clustering Algorithm**

El *Markov Clustering Algorithm* es una metodología basada en grafos (ver sección 3.1.1) que considera cada APV como un nodo, donde el grafo está determinado por la matriz de adyacencia. La matriz de adyacencia es aquella que contiene la distancia euclidiana entre cada par de APVs. A partir de esta matriz, el MCL computa otra matriz que contiene el número de caminos de largo  $k$  (para *weighted graphs*, calcula el *path capacities*) entre cada par de APVs [107]. El largo de un camino entre dos APVs se entiende como el número de otros APVs por los que hay que pasar para llegar de uno al otro en el grafo. Este proceso se lleva a cabo de forma iterativa calculando sucesivamente la matriz que contiene el número de camino de largo 1, 2, 3, etc.; hasta que no se producen variaciones en la matriz que contiene el número de caminos [107]. De esta manera, el MCL logra identificar los grupos de APVs topológicamente más relacionados en base al criterio del camino más largo [107].

El *Restricted Neighborhood Search Algorithm* es otra metodología basada en grafos (ver sección 3.3.1). El RNSC busca en el espacio de posibles clusters de APVs basado en una función de costo local a minimizar, que se calcula como la razón entre el número de conexiones entre pares de APVs *intra-cluster* y el número de conexiones entre pares de APVs *extra-cluster*. Luego, los *clusters* son el conjunto de APVs más fuertemente relacionados desde un punto de vista de las conexiones que los unen en un grafo.

En ambos casos el *input* es la matriz simétrica que contiene la distancia euclidiana entre todos los APVs.

A continuación, en la Tabla 9 y en la Tabla 10 se muestran los parámetros de que dependen el MCL y el RNSC, cuyos valores permitidos están acotados a tres posibilidades.



**Tabla 9: Lista de parámetros del que depende el algoritmo MCL.**

Parámetro	Descripción
<i>Source node</i>	Columna que contiene los nodos donde nace una unión (sólo para archivos delimitados por <i>tabs</i> )
<i>Target node</i>	Columna que contiene los nodos a los que llega una unión (sólo para archivos delimitados por <i>tabs</i> )
<i>Weight or label column</i>	Valor asociado a cada unión entre nodos (en el caso de esta tesis se utilizó la distancia euclidiana entre vectores).
<i>Inflation value</i>	Parámetro que actúa principalmente sobre el número de <i>clusters</i> a obtener. Incrementando este valor se obtiene un mayor número de <i>clusters</i> .

**Tabla 10: Lista de parámetros del que depende el algoritmo RNCS.**

Parámetros	Descripción
<i>Máximun number of clusters</i>	Especificar el número máximo de <i>clusters</i> , el cual puede ser como máximo el número de nodos (valor por defecto).
<i>Tabu length</i>	El valor por defecto es 1.
<i>Tabu list tolerance</i>	Es el número de veces que un nodo (APV) debe aparecer en la lista tabú antes que se prohíba su movimiento (de un <i>cluster</i> a otro).
<i>Naive stopping tolerance</i>	Número de pasos que el <i>naive scheme</i> continuará sin mejorar el mejor resultado. Valor por defecto de 5.
<i>Scaled stopping tolerance</i>	Número de pasos que el <i>scaled scheme</i> continuará sin mejorar el mejor resultado. Valor por defecto de 5. Si se define como cero, el algoritmo se salta el <i>scaled scheme</i> .
<i>Diversification frequency</i>	Sin esta opción, no se realiza diversificación. Si el <i>shuffling diversification lenght</i> también se utiliza, entonces el valor de este es también utilizado para el <i>diversification frequency</i> . Si la opción <i>shuffling diversification length</i> no es utilizada, entonces se utiliza el valor de <i>destructive diversification frequency</i> .
<i>Shuffling diversification length</i>	Si se fija en <i>num</i> , esto quiere decir que los últimos <i>num</i> movimientos en el período de diversificación, serán movimientos de diversificación. Este número no debe ser mayor que el de <i>diversification frequency</i> .
<i>Number of experiments</i>	Es el máximo número de experimentos o iteraciones.

En este caso escoger los parámetros no representó un problema, debido a que el MCL posee un solo parámetro, y el RNCS no obtiene soluciones distintas al modificar sus parámetros (en este caso, ver sección 7.1.1).

## 6.5 Búsqueda de Escalas Precursoras

Una vez que se realizó la búsqueda y generación de nuevas escalas con la metodología explicada en la sección 6.4, se obtuvo los modelos buscados según la metodología de la sección 6.1, y finalmente se evaluó éstos con el Error de Jackknife (sección 6.3). A partir del Error de Jackknife es posible determinar cuáles son las mejores escalas generadas, sin embargo, se desconoce el marco de aplicabilidad de los resultados obtenidos con los Algoritmos de Análisis Topológico y el *Genetic Algorithm*. Por un lado, esto se debe a que la información que alimenta los algoritmos es muy extensa y variada, y por el otro, a que la metodología de síntesis de nuevas escalas no es evidente (*straight forward*).

Por esto, en la presente sección se explica la metodología desarrollada para determinar los APVs precursores de las escalas que se obtuvo a través de Algoritmos de Análisis Topológico y el *Genetic Algorithm*, metodologías basadas en estas últimas, y que corresponden a un proceso inverso.

### 6.5.1 Metodología Base - Algoritmos de Redes Neuronales

En el capítulo de algoritmos de *clustering* (capítulo 3.3) se explicó la metodología básica de los algoritmos de redes neuronales, la que se basa en una simulación del comportamiento neuronal en la inteligencia animal. Brevemente, las neuronas se ordenan en distintas capas (ver Figura 20): la capa que recibe el estímulo externo, las capas intermedias u ocultas, y la capa que entrega el resultado o acción (por ejemplo, señal de movimiento de un músculo). El proceso de aprendizaje se produce gracias al ordenamiento de las neuronas, y al fortalecimiento de uniones entre ellas. En el caso de los algoritmos utilizados, el ordenamiento neuronal es un ordenamiento espacial, y se trata de un acercamiento a la señal. Una vez que la señal produce la activación de las neuronas de la primera capa, estas transmiten una señal a la segunda capa, y así sucesivamente.

En los algoritmos que se utilizaron, todas las neuronas pueden actuar como si estuvieran en la primera capa, al recibir directamente la señal, o como si estuvieran en una segunda o tercera capa, al ser una de las dos neuronas más cercanas a la que recibe la señal. Las neuronas que reciben la señal se acercan a ésta (APV) a un paso que disminuye según el número de capa en el que se encuentra la neurona.

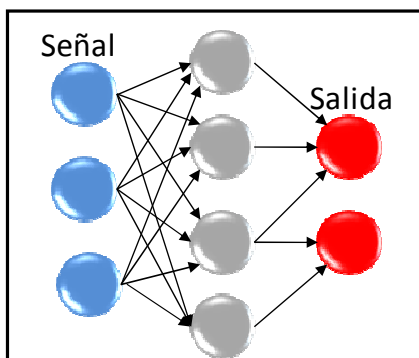


Figura 20: Esquema del proceso de inteligencia animal simplificado en una relación entre un conjunto de señales, una o más capas intermedias (hidden) y la capa que entrega el resultado o acción (denotada como salida).

En base a esta metodología se construyó de forma simplificada una metodología inversa, para determinar las escalas precursoras. La información base de esta metodología son los 74 APVs con base en la literatura (ver sección 6.2), y las escalas que se generaron simultáneamente con cada una de las mejores escalas. Lo anterior se esquematiza en la Figura 21, donde los 74 APVs con base en la literatura se representan por puntos de color azul (la señal), y las escalas creadas por el algoritmo se corresponden a las neuronas representadas mediante puntos rojos de mayor tamaño.

Según se muestra en la Figura 21, se utilizó dos metodologías, las que se explican a continuación:

## Metodología n°1

1. Se busca la neurona más cercana a cada APV (señal).
2. Se clasifica cada neurona (escala creada), según el APV asociada en el paso 1.
3. Se retiran las neuronas ya clasificadas.
4. Se repite el paso 1, 2 y 3; hasta que no quedan neuronas.

Al final, para cada APV (señal) se obtiene un conjunto de neuronas cuya ubicación espacial fue influenciada por el APV.

## Metodología n°2

1. Se busca la neurona más cercana a cada APV (señal).
2. Se clasifica cada neurona, según la escala/APV asociada en el paso 1.
3. Se retira las escalas (señales).
4. Las neuronas identificadas en el paso 1 ahora son consideradas como señales (toma el rol del APV).
5. Se repite el paso 1-4, hasta que no quedan más neuronas, pero en el paso 2 la neurona comparte la misma clasificación de la escala.

Al final, se obtuvo un registro de los APVs que dieron origen a cada neurona o escala creada, que genera al menos un mejor resultado al construir los modelos descritos en la sección 6.1.

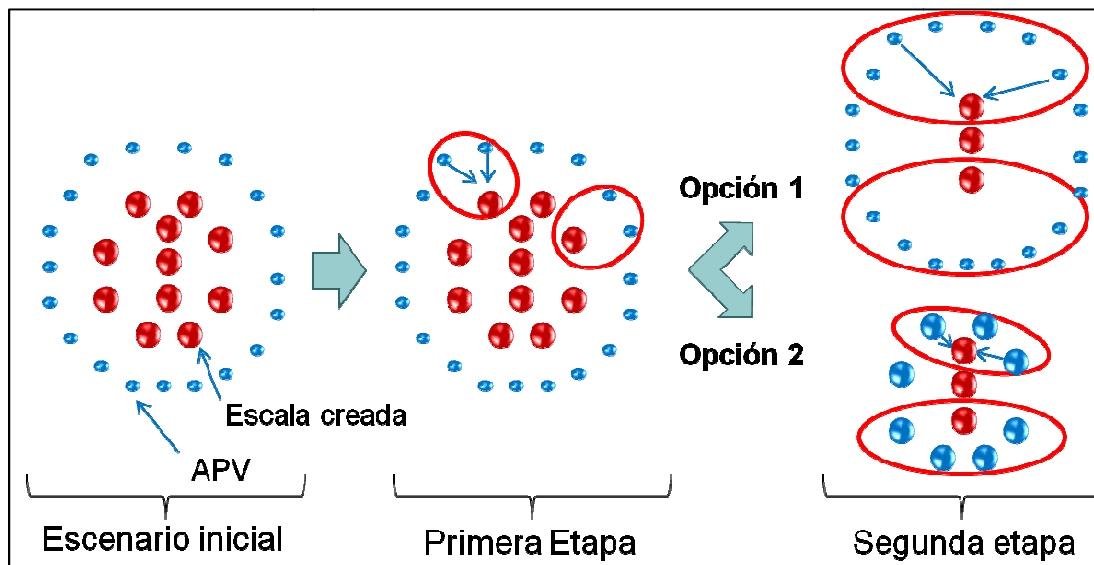


Figura 21: Esquema de metodologías utilizadas para la obtención de las escalas precursoras para los mejores resultados obtenidos con algoritmos de análisis topológico.

### 6.5.2 Metodología Base - *Genetic Algorithm*

La búsqueda de los APVs más influyentes en la ubicación espacial de las escalas creadas con el *Genetic Algorithm* es directo en algunos casos, y en otros existe más de una escala precursora posible.

Como se discutió en la sección 3.3.4, las escalas creadas con este algoritmo utilizan tres metodologías para generar la siguiente población o solución: selección de las

mejores soluciones de la generación anterior, mutación y *crossover*; pero en el presente trabajo no se utilizó la mutación.

La metodología empleada para buscar los APVs precursores de las mejores soluciones obtenidas con este algoritmo se esquematiza en la Figura 22. Para cada escala que da origen a uno de los mejores resultados, se identificó sus componentes (las escalas son vectores  $\vec{V}$ , con componentes  $v_i$ ). Luego, éstos se buscaron en los 74 APVs, y se registró para cada mejor resultado las coincidencias, esto es, que la posición y el valor coincidan entre el mejor resultado y un APV en al menos un componente.

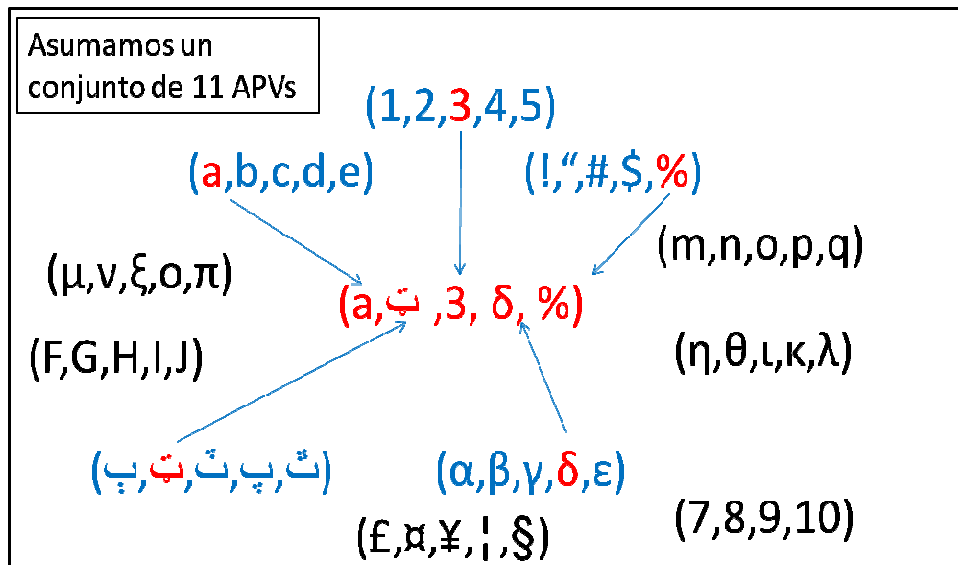


Figura 22: Esquema de metodologías utilizadas para la obtención de las escalas precursoras para los mejores resultados obtenidos con el *Genetic Algorithm*.

### 6.5.3 Síntesis de Listados de Escalas Precursoras

Una vez que se utiliza las metodologías de obtención de los APVs que originan las escalas asociados a los mejores modelos, se obtiene los listados de APVs. El objetivo es determinar los APVs más influyentes en la generación de las mejores escalas para su análisis, que corresponden a los APVs que contienen la mayor parte de la información sintetizada por las escalas a partir de las cuales se genera los modelos más exitosos. Esta tarea no es trivial, debido a que en total se generó 78 listas de APVs (ver Anexo N), las que pueden contener entre 1 a 74 a APVs.

Es por esta razón que en la Tabla 11 se define la estructura en la que se sintetizó la información, la que permite dar un primer paso para una visualización de los APVs más influyentes. Esta estructura consiste en sólo 8 listados de APVs.

**Tabla 11: Estructura de la información definida para sintetizar los 78 listados de escalas precursoras.**

Genetic Algorithm		Genetic Algorithm		Análisis Topológico		Análisis Topológico	
ATPS-3D	ATPS-LIN	DRT 3D	DRT LIN	ATPS-3D	ATPS-LIN	DRT 3D	DRT LIN
APV 1	APV 1	APV 1	APV 1	APV 1	APV 1	APV 1	APV 1
APV 2	APV 2	APV 2	APV 2	APV 2	APV 2	APV 2	APV 2
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
APV n <sub>1</sub>	APV n <sub>2</sub>	APV n <sub>3</sub>	APV n <sub>4</sub>	APV n <sub>5</sub>	APV n <sub>6</sub>	APV n <sub>7</sub>	APV n <sub>8</sub>

La síntesis de los APVs obtenidos para los modelos predictivos del DRT en HIC requiere una reducción de las 6 listas originales a 4 listas finales. Cada una de estas listas tiene un bajo número de APVs, razón por la que esta síntesis se hizo de forma directa, mediante la unión de los subconjuntos de APVs.

Por otro lado, la síntesis de los APVs obtenidos para los modelos de ATPS requiere una reducción de las 72 listas de APVs originales a 4 listas finales. En este caso, el alto número de APVs por lista no permite una metodología aditiva (caso anterior) ni visual. Para lograr esta síntesis, se desarrolló la metodología resumida en la Figura 23 y Figura 24, y descrita a continuación:

Primero, se agrupó los listados correspondientes a las dos metodologías de búsquedas de APVs precursoras de las escalas generadas mediante algoritmos de redes neuronales, permitiendo redundancia. De esta forma se redujo las 72 listas a 48 (son 24 listas por metodología de búsqueda de escalas, y se funcionó dos de los tres grupos de 24, quedando finalmente dos grupos de 24 listas).

Luego, se trabajó por separado las listas correspondientes a las cuatro combinaciones dadas por los dos tipos de algoritmos más exitosos, los de redes neuronales y el *Genetic Algorithm*; y los dos tipos de modelos (modelo lineal y modelo 3D). Para cada algoritmo se tiene 24 listas de APVs, las que se dividen en 12 listas para los modelos lineales y 12 listas para los modelos 3D, las que se sintetizó en 1 lista por modelo (las 12 listas por modelo se explica por los 12 sistemas ATPS).

En esta segunda etapa la metodología tiene algunas variaciones entre las listas de APVs obtenidas para las escalas generadas con el *Genetic Algorithm* y los de redes neuronales, esto se debe a que las listas asociadas al *Genetic Algorithm* tienen una frecuencia de influencia para cada APV, indicando el nivel de influencia que tuvo cada uno de éstos en la generación de las mejores escalas.

- La lista final de APVs precursoras asociados al Genetic Algorithm se obtuvo sumando las frecuencias individuales de cada APV. La lista final de APVs precursoras para cada uno de los dos tipos de sistemas fisicoquímicos, HIC y ATPS, es aquella que tiene una frecuencia superior al 30% (ver Figura 23).

Sistema	12 Listas de APVs				Lista APVs	Frecuencia
1	APV <sub>11</sub>	...	APV <sub>1j</sub>	...	k <sub>1</sub> =APV <sub>1</sub>	f <sub>1</sub>
	f <sub>11</sub>	...	f <sub>1j</sub>	...	APV <sub>2</sub>	f <sub>2</sub>
2	APV <sub>21</sub>	...	APV <sub>2j</sub>	...	APV <sub>3</sub>	f <sub>3</sub>
	f <sub>21</sub>	...	f <sub>2j</sub>	...	APV <sub>4</sub>	f <sub>4</sub>
3	APV <sub>31</sub>	...	APV <sub>3j</sub>	...	APV <sub>5</sub>	f <sub>5</sub>
	f <sub>21</sub>	...	f <sub>3j</sub>	...	APV <sub>6</sub>	f <sub>6</sub>
4	APV <sub>41</sub>	...	APV <sub>4j</sub>	...	APV <sub>7</sub>	f <sub>7</sub>
	f <sub>41</sub>	...	f <sub>4j</sub>	...	APV <sub>8</sub>	f <sub>8</sub>
5	APV <sub>51</sub>	...	APV <sub>5j</sub>	...	k <sub>9</sub> =APV <sub>9</sub>	f <sub>9</sub>
	f <sub>51</sub>	...	f <sub>5j</sub>	...	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

$$\text{Si } k = APV_{ij} (\forall i, j; i \neq j) \rightarrow f_k = f_k + f_{ij} (\forall k)$$

Figura 23: Esquema metodología de síntesis de listados de escalas precursoras (APVs). En la lista de los APVs por frecuencia se tiene que:  $f_i > f_{i+1}$ .

- Por otro lado, la lista final de APVs precursoras asociados a los algoritmos de redes neuronales se obtuvo mediante dos metodologías:
  - a. Metodología 1: Primero, se contó la cantidad de veces que aparece cada APV en los 12 sistemas ATPS, de esta manera se obtuvo una lista única la que se ordenó según la frecuencia relativa al APV de mayor frecuencia. La lista final para cada uno de los dos sistemas fisicoquímicos se obtuvo al filtrar considerando una frecuencia igual o mayor al 25% (ver Figura 24).
  - b. Metodología 2: En este caso, primero se filtró las listas de cada uno de los 12 APVs, dejando sólo los 7 APVs más influyentes (que son los primeros 7 APVs en cada lista). Luego se procede igual al punto a.

En ambos casos se obtuvo el mismo resultado final.

Sistema	12 Listas de APVs				Lista APVs	Frecuencia
1	APV <sub>11</sub>	...	APV <sub>1j</sub>	...	k <sub>1</sub> =APV <sub>1</sub>	f <sub>1</sub>
2	APV <sub>21</sub>	...	APV <sub>2j</sub>	...	APV <sub>2</sub>	f <sub>2</sub>
3	APV <sub>31</sub>	...	APV <sub>3j</sub>	...	APV <sub>3</sub>	f <sub>3</sub>
4	APV <sub>41</sub>	...	APV <sub>4j</sub>	...	APV <sub>4</sub>	f <sub>4</sub>
5	APV <sub>51</sub>	...	APV <sub>5j</sub>	...	APV <sub>5</sub>	f <sub>5</sub>
6	APV <sub>61</sub>	...	APV <sub>6j</sub>	...	APV <sub>6</sub>	f <sub>6</sub>
7	APV <sub>71</sub>	...	APV <sub>7j</sub>	...	APV <sub>7</sub>	f <sub>7</sub>
8	APV <sub>81</sub>	...	APV <sub>8j</sub>	...	APV <sub>8</sub>	f <sub>8</sub>
9	APV <sub>91</sub>	...	APV <sub>9j</sub>	...	k <sub>9</sub> =APV <sub>9</sub>	f <sub>9</sub>
10	APV <sub>101</sub>	...	APV <sub>10j</sub>	...	.	.
11	APV <sub>111</sub>	...	APV <sub>11j</sub>	...	.	.
12	APV <sub>121</sub>	...	APV <sub>12j</sub>	...	.	.

$$\text{Si } k = APV_{ij} (\forall i, j; i \neq j) \rightarrow f_k = f_k + 1 (\forall k)$$

Figura 24: Esquema metodología de síntesis de listados de escalas precursoras (APVs). En la lista de los APVs por frecuencia se tiene que:  $f_i > f_{i+1}$ .

#### **6.5.4 Estudio de la Ubicación Espacial de los Mejores Resultados**

En estudios anteriores se ha utilizado el APV más cercano a cada mejor escala generada como el APV de mayor influencia en la topología de ésta [4, 9], debido a que la topología deriva de la ubicación espacial.

En el presente análisis se consideró la ubicación espacial de las escalas, de dos formas:

1. Generando una tabla análoga a la Tabla 11, donde se encuentra los listados de APVs ordenados según las distancias a las que se encuentra de las mejores soluciones.
2. A través de una herramienta de representación visual de la composición de un listado de APVs, donde se puede relacionar la composición de los APVs con la distancia relativa a que se encuentran de la escala evaluada.

##### ***Tabla con Listados de APVs Ordenados por Distancia***

Como una herramienta análoga a la Tabla 11, es posible generar una tabla donde los APVs se ordenan por su grado de cercanía a las mejores escalas. Para cada mejor escala se construye una lista de los 74 APVs ordenada según la cercanía. En total son 26 listas de APVs, sobre las cuales se utilizó la metodología descrita en la sección 6.5.3 para los algoritmos de redes neuronales (se filtró cada lista de APVs para trabajar con los 10 APVs más cercanos), pero en este caso, esta metodología se aplicó para todas las mejores soluciones (independientemente de que hayan sido generadas con algoritmos de redes neuronales o con el *Genetic Algorithm*).

##### ***Evaluación de la Composición de un Listado de APVs***

Para evaluar la ubicación de las mejores soluciones se construyó gráficos de composición de los APVs en torno a las mejores soluciones. Cuando se habla de composición de APVs, se hace referencia a la clasificación de éstos en escalas del tipo conformacional, hidrofóbicas y estadísticas; y de la proporción de APVs de cada uno de estos tres tipos. En la Figura 25 se muestra un esquema sobre cómo se construyen estos gráficos.

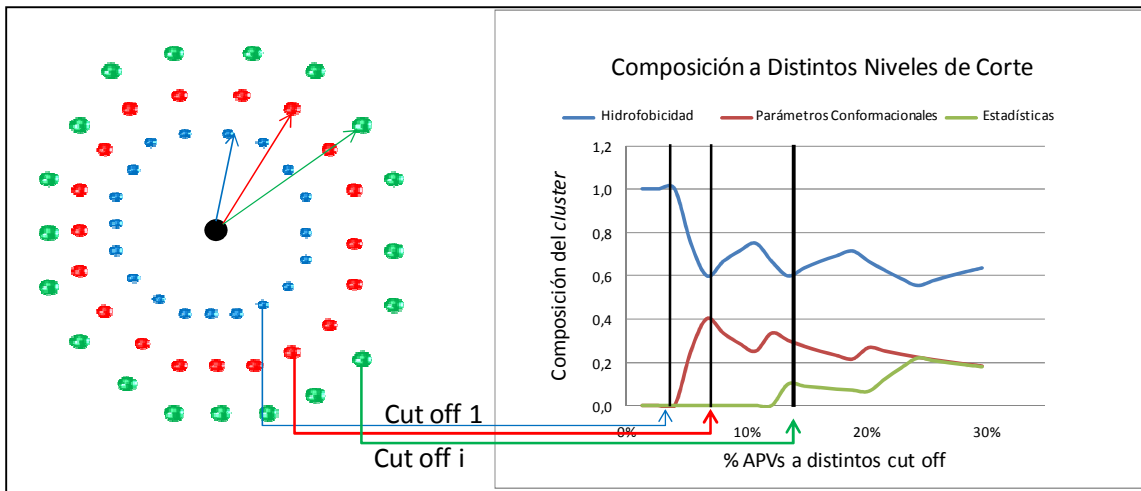


Figura 25: Esquema de la generación de gráficos para evaluar la composición de un

En la Figura 25 se muestra a la izquierda una solución como un punto negro, el que está rodeado de un conjunto de APVs que se ubican formando circunferencias a tres distancias distintas. Por otro lado, a la derecha se encuentra un gráfico cuya ordenada es la composición (fracción) de los distintos APVs, la que depende del porcentaje de APVs que se encuentran dentro de la circunferencia (hiper esfera) a distintos radios o *cut off* (eje de las abscisas).

El primer *cut off* es la distancia a la que está el primer APV, y por lo tanto la composición será 1 para el tipo de APV correspondiente, y cero para los otros dos, sin embargo, lo que se muestra en el eje de las abscisas es el porcentaje de APVs dentro de la circunferencia al radio escogido, en este caso es  $1/74 = 1,34\%$ .

Esta herramienta de representación visual se utilizó para: evaluar el espacio conformado por los APVs en torno a una mejor solución, evaluar la composición de las listas de APVs ordenadas por nivel de influencia a las mejores soluciones, y evaluar las listas de APVs ordenadas por nivel de cercanía a las mejores soluciones.

### 6.5.5 Determinación Final de Escalas Precursoras

Finalmente, para determinar cuáles son los APVs más relevantes en la generación de las mejores escalas, se comparó la Tabla 11, donde se encuentra el conjunto de APVs ordenados según su influencia (entre más influyentes más arriba); con su tabla homóloga de APVs ordenados por cercanía. Los APVs que aparecen con mayor frecuencia en los distintos sistemas, de ambas tablas, se consideraron los más relevantes en la generación de las escalas asociadas a las mejores predicciones, lo cual queda determinado por congruencia entre los dos criterios utilizados: el criterio de cercanía que se sustenta en la topología asociada a cada ubicación espacial, y el criterio de influencia que se sustenta en la metodología utilizada para la generación de las nuevas mejores escalas.

## 7 Resultados y Discusión

Un primer resultado importante son los modelos generados para predecir el comportamiento de proteínas en una cromatografía de interacción hidrofóbica y en 12



sistemas de dos fases acuosas. Estos modelos y sus características (Coeficiente de Pearson, Error de Jacknife y Error Cuadrático Medio) se presentan a continuación en la sección Escalas Asociadas a Mejores Modelos Predictivos para HIC y ATPS (7.1). En ésta se apreciar que se obtiene una mejora significativa en relación a los modelos del mismo tipo existentes a la fecha [4, 6, 9, 13, 23, 54, 62]. Se disminuyó el error asociado a la predicción, medido a través del Error de Jacknife, hasta un 75% para la cromatografía de interacción hidrofóbica y hasta un 99,6% para los sistemas de dos fases acuosas (en la sección 7.5 se analiza la posibilidad de sobre ajuste). La mayor parte de los mejores resultados se obtuvo con los algoritmos *Growing Neuronal Gas* y *Genetic Algorithm*.

## 7.1 Escalas Asociadas a Mejores Modelos Predictivos Para HIC y ATPS

### 7.1.1 Análisis de Disparidad en el Número de Escalas Obtenidas por cada Algoritmo

A partir de los distintos algoritmos utilizados, incluyendo *Kohonen Map* (SOM) y *K-mean*, se generó aproximadamente 344.800 vectores o escalas. En la Tabla 12 se puede ver la cantidad estimada de vectores generados por cada uno de estos algoritmos, y los porcentajes de vectores generados sobre el total.

Tabla 12: **Cantidad de vectores generados por cada uno de los algoritmos utilizados.**

Algoritmo	Nº Vectores	%
Self-Organizing Maps (SOM)	13.820	4,0
K – MEAN	7.100	2,1
Growing Neuronal Gas (GNG)	177.100	51,4
Growing Grid (GG)	65.547	19,0
Genetic Algorithm (GA)	78.000	22,6
Hierarchical Algorithm (HA)	90	0,0
Bisection Algorithm (BA)	3.150	0,9
Restricted Neighborhood Search (RNSC)	1	0,0
Markov Clustering Algorithm (MCL)	5	0,0

La cantidad de escalas obtenidas con cada uno de los algoritmos presentan diferencias significativas, por lo que se necesita verificar si existe un problema para comparar objetivamente los resultados obtenidos por éstos. A continuación se realiza un análisis sobre las razones que limitan el número de soluciones (vectores) obtenidos por los distintos tipos de algoritmos.

#### ***Restricted Neighborhood Search y Markov Clustering Algorithm***

La menor cantidad de escalas por algoritmo se obtuvo para los algoritmos *Restricted Neighborhood Search* (RNS) y *Markov Clustering Algorithm* (MCL). En el caso del MCL sólo se obtuvo cinco resultados distintos, mientras que para el RNS sólo se obtuvo una única solución trivial, que corresponde a un único *cluster* que contiene a todos los APVs estudiados.

El resultado obtenido con el *Restricted Neighbourhood Search* era esperable, debido a que éste es un algoritmo de optimización local basado en una función de costo global que es la suma de la función de costo de cada nodo, la que se calcula de la siguiente forma [134, 135]:

- El número de vecinos (nodo conectado directamente) que no están en el mismo cluster.
- El número de nodos que estando en el mismo cluster no son vecinos.
- Se suma la cantidad total y se divide por dos.

Dado que este algoritmo no incorpora ningún trabajo previo sobre los datos, y que el *input* dado es una matriz de adyacencia de vectores, donde todos los nodos están conectados entre sí [134, 135], es evidente que el *cluster* que minimiza la función de costo es un único *cluster* con todos los APVs en él.

Por otro lado, si bien con el *Markov Clustering Algorithm* se obtuvo un mayor número de *clusters*, este algoritmo se basa en el recorrido de rutas al azar y el número de nodos visitados en un mismo *cluster* [134, 135], lo cual es una medición muy acotada de las propiedades topológicas de un grupo de datos (se puede encontrar otras propiedades en [85, 187]), lo que podría explicar el reducido número de *clusters* encontrado.

### ***Self Organized Maps y K-mean***

Los vectores que se utilizó, asociados a los algoritmos *Self-Organizing Maps* (SOM) y *K-mean*, se obtuvieron en un estudio anterior [9].

Para comparar la cantidad de resultados obtenidos por el algoritmo SOM en relación a los otros algoritmos de redes neuronales, es importante tener en cuenta que al aumentar el número de parámetros del cual depende un algoritmo, también aumentan las combinaciones posibles de valores que pueden tomar estos parámetros, lo que puede producir un aumento en el número de soluciones que el algoritmo es capaz de generar para un set de datos determinados. Por lo tanto, al aumentar el número de combinaciones, puede ser necesario obtener un mayor número de soluciones para garantizar un mismo porcentaje de soluciones exploradas. En este caso el algoritmo SOM, si bien tiene más de 7 variables, sólo se tiene conocimiento de que se hayan hecho variar dos variables [9], las que definen el tamaño de la grilla a utilizar, por lo que desde esta perspectiva es comprensible el menor número de soluciones obtenidas.

Por otro lado el algoritmo *K-mean* tiene un solo parámetro, y el número de soluciones que se obtuvo es del orden del obtenido por el *Bisection Algorithm*.

### ***Bisection Algorithm y Hierarchical Algorithm***

Tanto el *Bisection Algorithm* como el *Hierarchical Algorithm* son algoritmos de *clustering*, al igual que los algoritmos RNS y MCL, sin embargo, son capaces de encontrar distintas soluciones no triviales. Para el caso del *Hierarchical Algorithm* se utilizó todos los *clusters* que puede generar, según la metodología descrita en 6.4.4. En cuanto al *Bisection Algorithm*, se utilizó una cantidad de combinaciones de valores para sus parámetros que incluyó todo el rango permitido, lo que fue posible dado el limitado número de valores admisibles para cada parámetro en la implementación utilizada.

Durante la ejecución del algoritmo en muchos casos se obtuvo los mismos grupos de *clusters*, lo que demuestra que se obtuvo un alto porcentaje de los posibles *clusters* que es capaz de generar.

### **Genetic Algorithm**

El *Genetic Algorithm* obtiene como resultado una escala por cada ejecución, sin embargo, genera una cantidad importante de vectores que evalúa durante su meta-heurística (procedimiento del algoritmo), presentando finalmente sólo una solución.

### **Growing Neuronal Gas y Growing Grid**

La diferencia entre la cantidad de escalas que se generó con los algoritmos *Growing Grid* (GG) y *Growing Neuronal Gas* (GNG) se deben principalmente a dos razones:

- En ambos casos se utilizó la misma cantidad de valores por parámetro, pero el algoritmo GNG posee más parámetros que deben ser ajustados, lo que permite explicar la generación de 1,33 veces más escalas con el algoritmo GNG.
- El algoritmo GNG tiene una mayor multiplicidad de resultados para cada conjunto de valores escogido para sus parámetros, por lo que fue necesario utilizar 5 repeticiones por cada conjunto de valores, a diferencia de las 3 repeticiones utilizadas en el algoritmo GG. Esto explica que con el algoritmo GNG se hayan generado 1,6 veces más escalas.

En conjunto, lo anterior explica la capacidad del algoritmo GNG de generar 2,2 veces más escalas respecto al algoritmo GG, lo que contrasta con los resultados de la Tabla 12 que muestra que el algoritmo GNG generó 2,7 veces más escalas. La diferencia anterior es razonable al considerar que la cantidad de soluciones (escalas) obtenidas en una ejecución difiere entre estos algoritmos. Adicionalmente, el número de soluciones puede diferir entre las distintas ejecuciones hechas con un mismo algoritmo al utilizar distintos parámetros.

Del análisis anterior, se establece que existen varios factores que influyen en la capacidad de generar distintas escalas para cada algoritmo en un mismo intervalo de tiempo. Los factores más relevantes que se identifican son: el número de parámetros, el tiempo de ejecución de cada algoritmo, y la posibilidad de automatizar la ejecución de los algoritmos para distintas combinaciones de parámetros.

Por lo tanto, el número de soluciones a obtener con un algoritmo puede ser considerado como un parámetro del desempeño del éste, y por lo tanto una comparación entre los resultados que se obtuvo con los distintos algoritmos utilizados en el presente trabajo es una comparación objetiva desde el punto de vista discutido.

### **7.1.2 Tiempo Requerido Para Ajuste de Modelos Utilizando Distintos Algoritmos**

Un resultado relevante, desde un punto de vista práctico del uso de algoritmos de *clustering* y optimización, es el tiempo que se utilizó para generar nuevas escalas y ajustar los modelos del DRT y el coeficiente de partición. A continuación, en la Tabla 13 se muestra los tiempos aproximados utilizados con cada uno de los algoritmos, diferenciando para HIC y ATPS.

Los tiempos que se muestran en la Tabla 13 incluyen el tiempo requerido para los siguientes procesos: preparación de datos para ser ingresados como *input* a cada algoritmo, ejecución del algoritmo (programa o *script* – no se considera la escritura del código), ajuste de los modelos, obtención del Error de Jackknife, búsqueda de escalas asociadas a mejores modelos (incluye escritura de algoritmos), y determinación de las características del modelos y sus coeficientes.

**Tabla 13: Algoritmos de clustering y optimización y tiempos que se utilizó para generar nuevas escalas y ajustar los modelos para la el predicción del DRT y el coeficiente de partición K. Los tiempos señalados se obtienen considerando los tiempos de ejecución con un procesador Intel® Core™ 2 Duo CPU T7250 @ 2 GHz - 2 GHz, 2 GB de memoria RAM, bajo un sistema operativo Windows Vista de 32 bits.**

Algoritmo	Sigla	Tiempo Utilizado HIC	Tiempo Utilizado ATPS	Tipo de Implementación
Growing Neuronal Gas	GNG	~1 meses	~5 meses	Propia en base a pseudocódigo [16]
Growing Grid	GG	~3,5 semanas	~3,5 meses	Propia en base a pseudocódigo [126]
Self-Organizing Maps	SOM	-	~2 semanas*	Resultados obtenidos de [9, 122]
K-means	-	-	~1 semana*	Resultados obtenidos de [9]
Bisection Algorithm	-	~3 días	-	Software gCluto [78]
Hierarchical Algorithm	HA	~4 días	-	Biblioteca de Matlab [131]
Markov Clustering Algorithm	MCL	~2 días	-	Aplicación web [107, 132-135]
Restricted Neighbourhood Search	RNS	~2 días	-	Aplicación web [106, 107, 134, 135]
Genetic Algorithm	GA	2 semanas	~1,5 meses	Biblioteca de Matlab [131]

\* Los tiempos señalados no incluyen la ejecución de los algoritmos, debido a que las escalas obtenidas y utilizadas corresponden a los obtenidos por otros autores.

Como se puede ver en la Tabla 13, los algoritmos que requirieron más tiempo para la obtención de los modelos ajustados son los del tipo Redes Neuronales (GG y GNG). Esto coincide con dos de los tres algoritmos a partir de los cuales se obtiene mayores escalas, que son: GNG, GG y *Genetic Algorithm*. Por otra parte, con el *Genetic Algorithm* se obtiene un 22,6% de las escalas generadas (versus un 19% y 51,4% de escalas que se obtiene con GG y GNG respectivamente - ver Tabla 12), pero éste sólo requirió entre un 30-50% del tiempo que se utilizó con los algoritmos de redes neuronales. El tiempo que se requirió para obtener los nuevos modelos del DRT y ATPS ajustados, al utilizar cada uno de los otros algoritmos no supera el 13% del demandado por GNG, y con éstos se obtiene en total un 7% de las escalas generadas.

### 7.1.3 Resultados Para la Cromatografía de Interacción Hidrofóbica (HIC)

Los resultados que se obtuvo para la cromatografía de interacción hidrofóbica se muestran en las cinco secciones siguientes: Línea Base – Mejores Resultados con APVs, Mejores Resultados por Algoritmo, Presentación de los Mejores Modelos, y Escalas con Mejores Resultados.

### **Línea Base de Comparación – Mejores Resultados con APVs**

Para poder determinar si los resultados obtenidos corresponden a una mejora en el poder predictivo de los modelos, se requiere de una referencia de comparación. En este caso, para la comparación se utilizó como referencia los mejores resultados ya obtenidos con APVs. De esta manera, la comparación permite determinar la contribución del uso de técnicas de análisis topológico y de optimización en la búsqueda de mejores escalas de propiedades aminoacídicas, por sobre los estudios experimentales y teóricos.

A continuación se muestra, en la Tabla 14 y la Tabla 15, los mejores resultados en términos de los Errores de Jackknife y las escalas con las que se ha obtenido estos resultados.

**Tabla 14: Atributos de los modelos de HIC obtenidos con APVs con base en la literatura.**

Tipo	Atributo	Valor
3D	MSE <sub>jk</sub> (x10 <sup>2</sup> )	12,98
	Pearson	0,92
	MSE (x10 <sup>2</sup> )	8,75
Modelo Lineal III	MSE <sub>jk</sub> (x10 <sup>2</sup> )	13,50
	Pearson	0,93
	MSE (x10 <sup>2</sup> )	7,38

**Tabla 15: Escalas utilizadas para obtención de mejores modelos con APVs con base en la literatura.**

TIPO	Escala	Descripción
3D	Wertz and Scheraga [180]	Fracción de aminoácidos no expuestos a solventes en 20 proteínas.
Lineal III	Wilson [181]	Constantes de Hidrofobicidad obtenidas a partir del tiempo de retención de péptidos en HPLA

### **Mejores Resultados por Algoritmo**

En la Tabla 16 se muestra los Errores de Jackknife de los mejores modelos encontrados por algoritmo para la predicción del comportamiento de proteínas en una cromatografía de interacción hidrofóbica. Se aprecia que los mejores resultados se han obtenido con los algoritmos *Growing Neuronal Gas* (GNG), *Growing Grid* (GG) y *Genetic Algorithm* (GA).

**Tabla 16: Error de Jacknife de modelos con mejores resultados por algoritmo.**

<b>Algoritmo</b>	<b>Modelo</b>	<b>MSE<sub>jk</sub> (x10<sup>2</sup>)</b>
APVs	3D	12,98
	Lineal	13,5
Self-Organizing Maps (SOM)	3D	14,5
	Lineal	12,3
<i>K-mean</i>	3D	13
	Lineal	12,9
Growing Neuronal Gas (GNG)	3D	7,7
	Lineal	8,6
Growing Grid (GG)	3D	6,1
	Lineal	14,1
Genetic Algorithm (GA)	3D	6,1
	Lineal	3,3
Hierarchical Algorithm (HA)	3D	13,5
	Lineal	13,5
Bisection Algorithm (BA)	3D	12,4
	Lineal	13,5
Markov Clustering Algorithm (MCL)	3D	11,7
	Lineal	44,4
Restricted Neighborhood Search (RNSC)	3D	-
	Lineal	-

Al comparar los resultados con los de la Tabla 16, se aprecia que para los mejores modelos tipo 3D (que utiliza la estructura tridimensional de las proteínas para el cálculo de hidrofobicidad) se logró una disminución de un 53% del valor del Error de Jacknife. Esto significa que se disminuyó en un 53% el error asociado al poder predictivo del modelo. Este resultado es compartido por los algoritmos: *Growing Grid* y *Genetic Algorithm*.

Con los modelos lineales, que no requieren la estructura tridimensional de las proteínas, se obtuvo una disminución del error asociado al poder predictivo de 36,3% con los algoritmos de análisis topológico (GNG), y una disminución de 75,6% con el *Genetic Algorithm*.

### **Presentación de los mejores modelos**

En la Tabla 17 se muestra el Error de Jacknife, el Pearson ( $R^2$ ) y el Error Cuadrático Medio (MSE) para los mejores modelos.

Al examinar la Tabla 17 es se aprecia que el uso de algoritmos de análisis topológico y de optimización, dentro de la metodología de construcción de nuevos modelos, permite disminuir significativamente el error asociado al poder predictivo ( $MSE_{jk}$ ) sin afectar el nivel de ajuste ( $R^2$ ) de los modelos a los datos experimentales; por el contrario, el Coeficiente de Pearson aumenta entre 2,2% y 5,4%, variación que puede ser considerada significativa, ya que el Pearson base es de 0,92 y 0,93 respectivamente.

Estos resultados son muy favorables, ya que el Coeficiente de Pearson y el Error de Jacknife no tienen una correlación significativa al evaluar los resultados obtenidos con un mismo algoritmo (éste es menor a 0,7 en todos los casos estudiados). Es decir, al mejorar el Error de Jacknife el Coeficiente de Pearson puede aumentar, disminuir o mantenerse. Esto sugieren que para los modelos cromatográficos no se hubo un sobre ajuste de los modelos a los datos.

**Tabla 17: Característica de los modelos con mejores niveles de ajuste de los DRT en HIC.**

Tipo	Atributo	A. Análisis Topológico A. Genético		
		APVs	Mejor	Mejor
3D	MSE <sub>jk</sub> (x10 <sup>2</sup> )	12,98	6,13	6,12
	Pearson	0,92	0,97	0,97
	MSE (x10 <sup>2</sup> )	8,75	3,10	2,92
Mod Lineal III	MSE <sub>jk</sub> (x10 <sup>2</sup> )	13,5	8,56	3,39
	Pearson	0,93	0,95	0,98
	MSE (x10 <sup>2</sup> )	7,38	5,19	2,62

Como se indicó en la introducción, los modelos utilizados para la predicción del tiempo de retención adimensional en una cromatografía de interacción hidrofóbica, tienen la forma de la ecuación 13:

$$DRT = b_0 + b_1\Gamma + b_2\Gamma^2 \quad (13)$$

Donde  $b_0$ ,  $b_1$  y  $b_2$  son los coeficientes obtenidos a través del procedimiento de los mínimos cuadrados, y  $\Gamma$  es la hidrofobicidad superficial media de una proteína.

A pesar que las escalas de hidrofobicidad están escaladas en el rango [0,1], en la práctica el dominio de los modelos es muy restringido, y depende de la hidrofobicidad superficial media del set de proteínas utilizadas para ajustar el modelo, los que varían para cada escala-modelo. Por ejemplo, los dominios y recorridos de los modelos de la Tabla 18 son:

- Modelo 3D – A. Topológicos:  $\mathbb{D}\epsilon[0,39; 0,50] \rightarrow \mathbb{R}\epsilon[0,00; 0,87]$
- Modelo Lineales – A. Topológicos:  $\mathbb{D}\epsilon[0,07; 0,14] \rightarrow \mathbb{R}\epsilon[0,00; 0,80]$
- Modelo 3D – A. Genético:  $\mathbb{D}\epsilon[0,27; 0,40] \rightarrow \mathbb{R}\epsilon[0,00; 0,92]$
- Modelo Lineales – A. Genético:  $\mathbb{D}\epsilon[0,35; 0,52] \rightarrow \mathbb{R}\epsilon[0,00; 0,95]$

Salvo el modelo lineal obtenido con algoritmos de análisis topológico, los modelos son muy similares a los de Salgado *et. al.* [9], cuyos dominios son:

- Modelos Lineal I:  $\mathbb{D}\epsilon[0,34; 0,39] \rightarrow \mathbb{R}\epsilon[0,00; 0,82]$
- Modelo Lineal II:  $\mathbb{D}\epsilon[0,35; 0,50] \rightarrow \mathbb{R}\epsilon[0,00; 0,93]$
- Modelo Lineal III:  $\mathbb{D}\epsilon[0,33; 0,42] \rightarrow \mathbb{R}\epsilon[0,00; 0,71]$

A continuación, en la Tabla 18, se muestra los valores de los coeficientes de los modelos correspondientes a los mejores resultados. Para cada coeficiente se muestra el intervalo de confianza, los cuales son de una magnitud similar a los obtenidos por Salgado *et. al.* [9].

**Tabla 18: Coeficientes de los mejores modelos del DRT en HIC.**

Tipo	Coeficiente	Algoritmos de Análisis Topológico			Algoritmo Genético		
		Valor	Min	Max	Valor	Min	Max
3D	$b_2$	-77,81	-144,70	-10,00	-53,51	-98,44	-8,57
	$b_1$	77,69	19,02	136,40	43,27	-0,15	86,68
	$b_0$	-18,52	-31,37	-5,68	-7,83	-18,30	2,64
Mod. Lineal	$b_2$	<b>-214,62</b>	-290,50	-138,80	-31,63	-51,75	-11,52
	$b_1$	58,21	39,39	77,30	32,83	15,96	49,70
	$b_0$	-3,15	-4,25	-2,04	-7,57	-11,09	-4,04

En la Tabla 18 se destaca el modelo lineal obtenido con algoritmos de análisis topológico. Este modelo tiene un coeficiente  $b_2$  inusualmente alto, el que se explica a partir de la Tabla 19 y Tabla 20, que se encuentran en la siguiente sección. En estas tablas se presenta las escalas a partir de las cuales se obtiene los modelos, que son la base del cálculo de la hidrofobicidad superficial media (ASA) del grupo de proteínas utilizadas para el estudio (ver sección 6.1.1).

En la Tabla 19 se puede ver que la escala utilizada para obtener el resultado del modelo lineal contiene cinco coeficientes con valores negativos, en contraste con el resto de las escalas que sólo tienen coeficientes positivos. Al utilizar una escala con coeficientes negativos, la presencia de aminoácidos asociados a éstos en la superficie disminuye la hidrofobicidad superficial media. Esto, a su vez, produce un desplazamiento hacia la izquierda en el plano cartesiano del modelo cuadrático en cuestión (en el eje  $x$  se grafica la hidrofobicidad y en el eje  $y$  el DRT), respecto a los otros modelos, es decir, mantiene el mismo rango de tiempos de retención (entre cero y uno) pero con valores menores de hidrofobicidad superficial media.

### **Escalas Generadas Asociadas a Mejores Modelos**

En esta sección se expone las escalas a partir de las cuales se logra la construcción de los mejores modelos (ver la Tabla 19). Los coeficientes de las distintas escalas se muestran en su forma normalizada y escalada, en el intervalo  $[0,1]$  pero a pesar de esto se ve que existen algunos coeficientes con valor menor a cero, en las escalas obtenidas con los algoritmos que realizan análisis topológico, lo cual se puede deber a que se está representado un valor hidrofóbico de mayor magnitud al encontrado en el conjunto de APVs (ver sección 6.2). Esto no ocurre con el *Genetic Algorithm*, lo cual es lógico, debido a que éste únicamente puede recombinar componentes de los 74 APVs.

Al evaluar los Coeficientes de Correlación entre estas escalas se aprecia que ésta es mayor según el tipo de algoritmo con el que fueron generadas, por sobre el tipo de modelo. Las escalas obtenidas a partir de algoritmos de análisis topológico poseen un coeficiente de 0,76; mientras que las obtenidas con el *Genetic Algorithm* poseen un Coeficiente de Correlación de solo 0,68. La correlación entre escalas obtenidas con algoritmos distintos es mayor a 0,5 en solo un caso (0,64) de cuatro.



**Tabla 19: Mejores Escalas obtenidas con algoritmos que utilizan la topología de los APVs.**

Nº	Aminoácido	3D [MSE <sub>jk</sub> ]	Lineal [MSE <sub>jk</sub> ]
1	ALA	0,52	0,39
2	ARG	0,30	0,05
3	ASN	0,39	-0,05
4	ASP	0,45	-0,11
5	CYS	0,70	0,72
6	GLN	0,38	-0,05
7	GLU	0,42	0,08
8	GLY	0,38	0,64
9	HIS	0,52	-0,01
10	ILE	0,82	0,89
11	LEU	0,82	0,56
12	LYS	0,19	-0,25
13	MET	0,74	0,49
14	PHE	0,92	0,56
15	PRO	0,60	0,13
16	SER	0,38	0,06
17	THR	0,43	0,12
18	TRP	0,94	0,53
19	TYR	0,73	0,35
20	VAL	0,71	0,64

**Tabla 20: Mejores escalas obtenidas con el *Genetic Algorithm*.**

Nº	Aminoácido	3D [MSE <sub>jk</sub> ]	Lineal [MSE <sub>jk</sub> ]
1	ALA	0,08	0,22
2	ARG	0,38	0,35
3	ASN	0,62	0,16
4	ASP	0,43	0,91
5	CYS	0,62	0,52
6	GLN	0,47	0,55
7	GLU	0,56	0,13
8	GLY	1,00	0,19
9	HIS	0,45	0,32
10	ILE	0,00	0,92
11	LEU	0,06	0,9
12	LYS	0,66	0,23
13	MET	0,12	0,51
14	PHE	0,16	0,64
15	PRO	0,73	0,00
16	SER	0,66	0,36
17	THR	0,31	0,54
18	TRP	0,19	0,75
19	TYR	0,34	0,57
20	VAL	0,09	1,00

### 7.1.4 Resultados Para Sistemas de Dos Fases Acuosa (ATPS)

Los resultados obtenidos para los sistemas de dos fases acuosas se muestran en las siguientes cinco secciones, de la misma forma como se expusieron para la cromatografía de interacción hidrofóbica: Línea Base – Mejores Resultados con APVs, Mejores Resultados por Algoritmo, Presentación de los Mejores Modelos, Escalas con Mejores Resultados.

#### Línea Base - Mejores Resultados con APVs

Para poder determinar si los resultados obtenidos corresponden a una mejora en el poder predictivo de los modelos se requiere de una referencia de comparación. En este caso, para la comparación se utilizó como referencia los mejores resultados ya obtenidos con APVs. De esta manera, la comparación permite determinar la contribución del uso de técnicas de análisis topológico y de optimización en la búsqueda de mejores escalas de propiedades aminoacídicas, por sobre los estudios experimentales y teóricos.

A continuación, en la Tabla 21, Tabla 22 y Tabla 23, se muestra los mejores resultados obtenidos con APVs en términos de los Errores de Jackknife, y las escalas con las que se obtuvieron estos resultados.

**Tabla 21: Valores base para la comparación de los resultados obtenidos con los APVs con base en la literatura. Corresponden a los atributos: Error de Jackknife, coeficiente de Pearson y el Error Cuadrático Medio para cada uno de los sistemas de dos fases acuosas.**

Tipo	Atributo	SISTEMAS DE DOS FASES ACUOSAS											
		PEG-F 0,0%	PEG-F 0,6%	PEG-F 8,8%	PEG-S 0,0%	PEG-S 0,6%	PEG-S 8,8%	PEG-C 0,0%	PEG-C 0,6%	PEG-C 8,8%	PEG-D 0,0%	PEG-D 0,6%	PEG-D 8,8%
3D	MSE <sub>jk</sub> (x10 <sup>2</sup> )	3,9	5,8	25,1	19,2	7,3	27,3	2	7,1	47,2	3,1	9,1	7,7
	Pearson	0,71	0,76	0,79	0,11	0,84	0,79	0,86	0,78	0,79	0,54	0,7	0,72
	MSE (x10 <sup>2</sup> )	3	4	18,7	15,8	4,2	20,3	1,3	5,2	36,1	2,3	7,1	5,5
Mod Lineal	MSE <sub>jk</sub> (x10 <sup>2</sup> )	7,7	11,1	34,3	19,2	10,9	31,6	5	12,5	58	0,8	5	3,5
	Pearson	0,17	0,39	0,68	0,3	0,69	0,76	0,49	0,52	0,73	0,9	0,87	0,87
	MSE (x10 <sup>2</sup> )	5,8	8,2	27	14,5	7,7	23,2	3,7	9,7	44,4	0,6	3,3	2,8
	Modelo	III	I	II	II	III	II	I	II	I	III	II	I

**Tabla 22: Escalas con las cuales se construyeron los modelos tipo 3D utilizados como caso base.**

ESCALAS UTILIZADAS PARA OBTENCIÓN DE MEJORES MODELOS (TIPO 3D)				
Nº Sistema	Sistema	NaCl % (p/p)	Escala	Descripción <sup>1</sup>
1	Fosfato	0	Wilson et al. [181]	Escala de hidrofiliidad derivada de la retención de péptidos en una HPLC
2	Fosfato	0.6	Wilson et al. [181]	Escala de hidrofiliidad derivada de la retención de péptidos en una HPLC
3	Fosfato	8.8	Meek [130]	Coefficiente de Retención en HPLC a pH 7,4
4	Sulfato	0	Wilson et al. [181]	Escala de hidrofiliidad derivada de la retención de péptidos en una HPLC
5	Sulfato	0.6	Welling et al. [179]	Valores de antigenicidad ( <i>Antigenicity value</i> )
6	Sulfato	8.8	Sweet and Eisenberg [178]	Hidrofobicidad obtenida por optimización ( <i>optimized matching</i> )
7	Citrato	0	Wilson et al. [181]	Escala de hidrofiliidad derivada de la retención de péptidos en una HPLC
8	Citrato	0.6	Browne et al. [156]	Coefficientes de retención obtenidos de HFBA
9	Citrato	8.8	Meek [130]	Coefficiente de Retención en HPLC a pH 7,4
10	Dextrano	0	Meek [130]	Coefficiente de Retención en HPLC a pH 7,4
11	Dextrano	0.6	Meek [130]	Coefficiente de Retención en HPLC a pH 7,4
12	Dextrano	8.8	Meek [130]	Coefficiente de Retención en HPLC a pH 7,4

<sup>1</sup> Se incluye una breve descripción del tipo de escala al que corresponden.

**Tabla 23: Escalas utilizadas en la elaboración de los modelos del tipo lineal, utilizados como caso base. Se incluye una breve descripción del tipo de escala al que corresponden.**

ESCALAS UTILIZADAS PARA OBTENCIÓN DE MEJORES MODELOS (LINEAL)				
Nº Sistema	Sistema	NaCl % (p/p)	Escala	Descripción
1	Fosfato	0	Parker <i>et. al.</i> [144]	Escala de hidrofiliidad derivada de la retención de péptidos en una HPLC
2	Fosfato	0.6	Welling et al. [179]	Valores de antigenicidad ( <i>Antigenicity value</i> )
3	Fosfato	8.8	Meek [130]	Coefficiente de Retención en HPLC, pH 7,4
4	Sulfato	0	Erikkson [166]	Escala de hidrofobicidad basada en energía libre de transferencia en un sistema etanol-agua
5	Sulfato	0.6	Welling et al. [179]	Valores de antigenicidad
6	Sulfato	8.8	Cowan and Whittaker [158]	Índices de hidrofobicidad determina con una HPLC a pH 7,5
7	Citrato	0	Welling et al. [179]	Valores de antigenicidad
8	Citrato	0.6	Meek [130]	Coefficiente de Retención en HPLC a pH 7,3
9	Citrato	8.8	Meek [130]	Coefficiente de Retención en HPLC a pH 7,4
10	Dextrano	0	Jesior [188]	Composición espacial de vecinos
11	Dextrano	0.6	Jesior [188]	Composición espacial de vecinos
12	Dextrano	8.8	Welling et al. [179]	Valores de antigenicidad

### **Mejores Resultados por Algoritmo**

En la Tabla 24 se muestra los Errores de Jacknife de los mejores modelos encontrados para cada algoritmo y sistema de dos fases acuosas. En general, los modelos con mayor poder predictivo se obtuvieron con los algoritmos *Growing Neuronal Gas* (GNG) y *Genetic Algorithm* (GA), salvo para el sistema 10, en el cual el algoritmo *Growing Grid* supera al GNG, pero no al algoritmo GA.

Para analizar los resultados, en la Tabla 25 se muestra la variación porcentual de los Errores de Jacknife de la Tabla 24, respecto a la línea base de comparación (Tabla 21), que correspon de a los resultados obtenidos directamente con los 74 APVs.

A través del tipo de modelo que utiliza la estructura tridimensional de las proteínas (3D) para el cálculo de hidrofobicidad, se logró una disminución máxima de un 93,2% del valor del Error de Jacknife, para el sistema 6, lo que indica que se disminuyó en un 93,2% el error asociado al poder predictivo del modelo. Este resultado se obtuvo con el algoritmo GNG, lo que contrasta con el algoritmo GA que obtuvo una disminución máxima de 82%.

Para el caso de los modelos lineales, que no requieren la estructura tridimensional de las proteínas, se obtuvo una disminución máxima en el error asociado al poder predictivo a través de algoritmos de análisis topológico de 97,6% (sistema 9), y con el *Genetic Algorithm* la disminución máxima del Error de Jacknife fue de 87,8% (sistema 2).

**Tabla 24: Mejores resultados por algoritmo para cada uno de los 12 sistemas de dos fases acuosas, según los valores del Error de Jacknife ( $MSE_{jk}$ ).**

[ $MSE_{jk}$ ]		SISTEMAS DE DOS FASES ACUOSAS											
Algoritmo	Modelo	PEG-F			PEG-S			PEG-C			PEG-D		
		0,0%	0,6%	8,8%	0,0%	0,6%	8,8%	0,0%	0,6%	8,8%	0,0%	0,6%	8,8%
APVs	3D	3,9	5,8	25,1	19,2	7,3	27,3	2,0	7,1	47,2	3,1	9,1	7,7
	Lineal	7,7	11,1	34,3	19,2	10,9	31,6	5,0	12,5	58,0	0,8	5,0	3,5
K-mean	3D	3,9	5,8	20,9	19,2	5,8	23,7	2,0	7,1	35,3	3,1	8,4	7,7
	Lineal	6,2	8,7	30,0	15,4	7,6	31,5	4,1	11,0	57,4	0,8	4,4	3,5
SOM	3D	4,0	5,2	9,2	18,5	6,6	15,0	1,7	4,7	10,1	2,9	7,9	8,0
	Lineal	6,5	8,2	27,9	15,6	6,7	26,5	4,1	8,4	47,4	0,3	2,2	3,6
GNG	3D	3,5	4,3	8,4	7,6	4,5	1,9	1,2	3,9	9,3	0,6	4,7	5,9
	Lineal	0,2	1,3	1,7	0,8	1,0	0,1	0,3	2,2	1,6	0,2	0,6	0,7
GG	3D	3,8	4,4	9,4	14,2	5,2	11,8	1,5	3,8	11,7	2,7	7,5	6,4
	Lineal	7,3	9,6	32,2	20,4	13,4	38,0	4,3	11,6	62,4	2,1	8,4	6,4
GA	3D	4,3	2,4	6,5	6,9	4,0	9,5	4,2	2,9	16,5	1,8	5,2	1,4
	Lineal	3,2	1,3	12,6	13,6	1,2	19,9	3,1	5,0	26,0	0,4	1,3	1,9

**Tabla 25: Mejores resultados por algoritmo para cada uno de los 12 sistemas de dos fases acuosas, como porcentajes de reducción del Error de Jacknife respecto a la línea base ( $\%MSE_{jk}$ ). Los colores son una combinación de rojo, verde y azul, los que representan los porcentajes más bajos, intermedios y altos respectivamente.**

[% $MSE_{jk}$ ]		SISTEMAS DE DOS FASES ACUOSAS											
Algoritmo	Modelo	PEG-F			PEG-S			PEG-C			PEG-D		
		0,0%	0,6%	8,8%	0,0%	0,6%	8,8%	0,0%	0,6%	8,8%	0,0%	0,6%	8,8%
k-Mean	3D	-	-	17	-	21	13	-	-	25	-	8	-
	Lineal	19	22	13	20	30	0	18	12	1	-	12	-
SOM	3D	- 3	10	63	4	10	45	15	34	79	6	13	- 4
	Lineal	16	26	19	19	39	16	18	33	18	63	56	- 3
GNG	3D	10	26	67	60	38	93	40	45	80	81	48	23
	Lineal	97	88	95	96	91	100	94	82	97	75	88	80
GG	3D	3	24	63	26	29	57	25	46	75	13	18	17
	Lineal	5	14	6	- 6	- 23	- 20	14	7	- 8	-163	- 68	- 83
GA	3D	- 10	59	74	64	45	65	-110	59	65	42	43	82
	Lineal	58	88	63	29	89	37	38	60	55	50	74	46

### Presentación de los mejores modelo

En la Tabla 26 y en la Tabla 28, se muestra el Error de Jacknife, el Pearson ( $R^2$ ) y el Error Cuadrático Medio (MSE) para los mejores resultados obtenidos para cada uno de los sistemas de dos fases acuosas.

A partir de la Tabla 26 y la Tabla 28 se observa que el uso de algoritmos de análisis topológico y de optimización para la construcción de nuevos modelos permitió disminuir significativamente el error asociado al poder predictivo ( $MSE_{jk}$ ).

Por otro lado, a diferencia de los resultados obtenidos para la cromatografía, el Coeficiente de Pearson de los mejores modelos de los sistemas de dos fases acuosas experimentó aumentos y disminuciones con respecto al caso base, lo cual concuerda

con los resultados obtenidos por Salgado *et. al.* [4], quien utiliza 36 APVs para construir los modelos y compara las diferencias entre el modelos 3D y los modelos lineales (propuestos por él).

En los resultados que se obtuvo a partir de los algoritmos de análisis topológico, los modelos tipo 3D exhiben disminuciones en el Coeficiente de Pearson en los sistemas 6, 10 y 11; siendo la media una disminución del 8%. Por su parte, en los modelos del tipo lineal se observa disminuciones para todos los sistemas con excepción del sistema 1. La media es una disminución del Coeficiente de Pearson de un 40,3% con respecto al caso base.

Para los mejores modelos obtenidos a partir del *Genetic Algorithm* se observa un aumento significativo del Coeficiente de Pearson en casi todos los sistemas (salvo para los resultados omitidos y los modelos lineales 10, 11 y 12). En promedio, los modelos del tipo 3D presentan un aumento en el Coeficiente de Pearson de 40,8%, y los modelos tipo lineal del 56,1%.

Como se comentó en la sección análoga para HIC, no hay una correlación clara entre el Error de Jacknife y el Coeficiente de Pearson, razón por la cual no sorprenden los resultados de este último coeficiente. En el caso de los resultados obtenidos a partir de algoritmos de análisis topológico y modelos lineales, el Coeficiente de Pearson de los modelos disminuyó de forma significativa, sin embargo, esto no implica un fracaso de la metodología, ya que el criterio de selección de las escalas se basa únicamente en la máxima disminución del Error de Jacknife.

Para facilitar el análisis, se generó la Tabla 27 y la Tabla 29, donde se muestra la variación de los Errores de Jacknife, Pearson y el Error Cuadrático Medio; respecto a la línea base de comparación. En todos los casos, un aumento porcentual positivo significa una mejora en el coeficiente, es decir, un aumento porcentual en el Coeficiente de Pearson significa un mejor ajuste, mientras que para el Error de Jacknife significa una disminución del error de predicción.

Adicionalmente, en la Tabla 27 y la Tabla 29 se aprecia una gran diferencia entre los resultados obtenidos con los algoritmos de análisis topológico y el *Genetic Algorithm*. Mientras con el primero se obtuvo mejores resultados con el modelo tipo 3D, respecto de los lineales, con el *Genetic Algorithm* se obtuvo peores resultados con el modelo tipo 3D (salvo en 5 casos), respecto de los modelos tipo lineal (salvo en 3 casos). Lo anterior se puede deber a que el *Genetic Algorithm* está restringido a la recombinación de escalas para la generación de las nuevas, sin que se permita mutación (se restringe su capacidad de realizar pequeñas modificaciones).

Por otro lado, se observa que, en general, los modelos de los sistemas 10, 11 y 12, que corresponden al sistema PEG/Dextrano, presentan mayor dificultad para mejorar su poder predictivo y su ajuste, lo que se puede deber a que éstos tienen una línea base muy buena, especialmente para el caso de los modelos lineales.

Por otra parte, se ha publicado sobre las dificultades existentes para obtener buenos modelos a bajas concentraciones de sal [4, 71]. En los resultados del presente estudio, se observa mejoras cuyas magnitudes podrían ser asociadas a este fenómeno en los sistemas 1, 2, 3, 7, 8 y 9 (ver Tabla 27); que corresponden a los sistemas PEG-Fosfato

y PEG-Citrato. Sin embargo, al revisar los coeficientes finales en la Tabla 26 no se observa que el poder predictivo sea mejor a bajas concentraciones de sal, por el contrario, el ajuste tiende a mejorar en los modelos construidos para los sistemas con mayores concentraciones de sal. Por lo tanto, en el presente estudio no se aprecia una tendencia clara de la capacidad predictiva de los modelos en función de la concentración de sal.

**Tabla 26: Parámetro estadísticos asociados a los mejores modelos obtenidos con algoritmos de análisis topológico.**

		SISTEMAS DE DOS FASES ACUOSAS											
Tipo	Atributo	PEG-F 0,0%	PEG-F 0,6%	PEG-F 8,8%	PEG-S 0,0%	PEG-S 0,6%	PEG-S 8,8%	PEG-C 0,0%	PEG-C 0,6%	PEG-C 8,8%	PEG-D 0,0%	PEG-D 0,6%	PEG-D 8,8%
3D	MSE <sub>jk</sub>	3,46	4,32	8,41	7,57	4,48	1,85	1,22	3,86	9,30	0,52	4,71	5,88
	Pearson	0,74	0,83	0,93	0,22	0,91	0,03	0,91	0,89	0,96	0,19	0,57	0,80
	MSE	2,70	2,99	6,30	-0,98	2,43	52,15	0,87	2,70	6,62	-0,28	2,58	4,00
Mod Lineal	MSE <sub>jk</sub>	-0,01	2,43	1,79	7,08	0,19	4,79	0,34	2,15	1,40	0,00	0,65	0,38
	Pearson	0,33	0,34	0,27	0,05	0,25	0,15	0,38	0,11	0,42	0,34	0,48	0,64
	MSE	-1,30	0,29	0,35	12,49	-3,07	11,60	-0,14	5,93	-14,70	-0,72	-3,05	-0,26
	Modelo	II	I	I	I	I	III	II	II	II	III	II	II

**Tabla 27: Indicadores del resultado para los mejores modelos obtenidos con algoritmos de análisis topológico. Resultados expresados en porcentajes de variación respecto a línea base de manera que las variaciones positivas sean favorables para los distintos coeficientes.**

		SISTEMA DE DOS FASES ACUOSAS [%]											
Tipo	Atributo	PEG-F 0,6%	PEG-F 8,8%	PEG-S 0,0%	PEG-S 0,6%	PEG-S 8,8%	PEG-C 0,0%	PEG-C 0,6%	PEG-C 8,8%	PEG-D 0,0%	PEG-D 0,6%	PEG-D 8,8%	PEG-F 0,0%
3D	MSE <sub>jk</sub>	11	26	66	61	39	93	39	46	80	83	48	24
	Pearson	4	9	18	100	8	-96	6	14	22	-65	-19	11
	MSE	10	25	66	106	42	-157	33	48	82	112	64	27
Mod Lineal	MSE <sub>jk</sub>	100	78	95	63	98	85	93	83	98	100	87	89
	Pearson	94	-13	-60	-83	-64	-80	-22	-79	-42	-62	-45	-26
	MSE	122	96	99	14	140	50	104	39	133	220	192	109
	Modelo	II	I	I	I	I	III	II	II	II	III	II	II

Tabla 28: Indicadores del resultado para los mejores modelos obtenidos con el *Genetic Algorithm*. Las celdas en blanco corresponden a sistemas para los cuales los mejores resultados de la optimización genética no logran superar los valores del caso base en ningún parámetro, por el contrario, aumentan de forma muy significativa.

		SISTEMA DE DOS FASES ACUOSAS											
Tipo	Atributo	PEG-F 0,0%	PEG-F 0,6%	PEG-F 8,8%	PEG-S 0,0%	PEG-S 0,6%	PEG-S 8,8%	PEG-C 0,0%	PEG-C 0,6%	PEG-C 8,8%	PEG-D 0,0%	PEG-D 0,6%	PEG-D 8,8%
3D	MSE <sub>jk</sub>	-	2,45	6,53	21,71	3,99	-	-	2,92	-	-	-	1,39
	Pearson	-	0,92	0,95	0,27	0,92	-	-	0,91	-	-	-	0,95
	MSE	-	1,52	4,62	16,00	2,17	-	-	2,26	-	-	-	1,04
Mod. Lineal III	MSE <sub>jk</sub>	4,70	1,35	13,56	16,11	1,42	23,83	3,10	4,95	38,89	2,31	8,47	3,86
	Pearson	0,68	0,95	0,89	0,48	0,97	0,83	0,73	0,86	0,82	0,69	0,76	0,85
	MSE	3,18	0,92	10,62	12,27	0,92	17,10	2,27	3,57	30,62	1,73	5,76	3,07

Tabla 29: Indicadores del resultado para los mejores modelos obtenidos con el *Genetic Algorithm*. Resultados expresados en porcentajes de variación respecto a línea base de manera que las variaciones positivas sean favorables para los distintos coeficientes.

		SISTEMA DE DOS FASES ACUOSAS											
Tipo	Atributo	PEG-F 0,0%	PEG-F 0,6%	PEG-F 8,8%	PEG-S 0,0%	PEG-S 0,6%	PEG-S 8,8%	PEG-C 0,0%	PEG-C 0,6%	PEG-C 8,8%	PEG-D 0,0%	PEG-D 0,6%	PEG-D 8,8%
3D	MSE <sub>jk</sub>	-	58	74	-13	45	-	-	59	-	-	-	82
	Pearson	-	21	20	145	10	-	-	17	-	-	-	32
	MSE	-	62	75	-1	48	-	-	57	-	-	-	81
Mod Lineal III	MSE <sub>jk</sub>	39	88	60	16	87	25	38	60	33	-189	-69	-10
	Pearson	300	144	31	60	41	9	49	65	12	-23	-13	-2
	MSE	45	89	61	15	88	26	39	63	31	-188	-75	-10

En la Tabla 30 y Tabla 31 semuestra los valores de los coeficientes de los modelos correspondientes a los mejores resultados. Como se indicó en la introducción, estos modelos tienen la forma descrita en la ecuación 22, a continuación:

$$\log K = R \log \Gamma - R \log \Gamma_0 \quad (22)$$

Donde  $\Gamma$  y  $R$  son los coeficientes del modelo que se desea determinar. Para cada caso se muestra el intervalo de confianza en que se encuentran los coeficientes.

De la metodología descrita en la sección 6.2 se sabe que las escalas de hidrofobicidad están escaladas y normalizadas en el rango  $[0,1]$ , lo cual es coherente con los modelos cuyo dominio está restringido al intervalo  $\mathbb{D} \in [0,1]$  debido a la función logarítmica. Esto, porque el recorrido de  $\log K$  se encuentra en el intervalo  $\mathbb{R} \in [0, -1]$ , de manera que se cumpla con que  $K \in [0,1]$ .

En la práctica cada modelo tiene un dominio más restringido, al igual que en el caso de HIC, sin embargo el conjunto de los distintos modelos calculados utilizan el intervalo completo.



En la Tabla 30 y Tabla 31 se muestra el logaritmo de  $\Gamma$ , utilizado para linealizar el modelo.

**Tabla 30: Coeficientes de mejores modelos del coeficiente de partición obtenidos con los algoritmos de análisis topológico**

		SISTEMAS DE DOS FASES ACUOSAS											
Tipo	Atributo	PEG-F 0,0%	PEG-F 0,6%	PEG-F 8,8%	PEG-S 0,0%	PEG-S 0,6%	PEG-S 8,8%	PEG-C 0,0%	PEG-C 0,6%	PEG-C 8,8%	PEG-D 0,0%	PEG-D 0,6%	PEG-D 8,8%
3D	log $\Gamma_0$	-0,30	-0,71	-0,44	-2,89	-0,62	-4,48	-0,72	-0,43	-0,47	-1,23	-1,09	-0,44
	Max log $\Gamma$	0,00	-0,18	-0,26	2,84	-0,30	119,11	-0,35	-0,19	-0,34	2,90	0,43	-0,09
	Min log $\Gamma$	-0,59	-1,25	-0,62	-8,61	-0,95	-128,07	-1,09	-0,66	-0,61	-5,37	-2,61	-0,80
	R	16,13	4,27	-42,26	-0,35	5,86	0,12	3,33	24,47	33,48	-0,14	0,36	-17,10
	Max R	5,08	6,43	3,32	0,27	7,83	3,49	4,50	33,78	40,38	0,22	0,09	-7,45
	Min R	27,17	2,11	-42,66	-0,97	3,90	-3,24	2,17	15,17	26,57	-0,50	0,63	-26,75
Mod Lineal	log $\Gamma_0$	-2,40	-0,78	-1,75	3,32	0,97	-0,10	-0,84	0,59	-1,36	-0,99	-1,00	-1,95
	Max log $\Gamma$	8,81	0,84	1,11	-42,66	-2,39	5,43	0,66	-7,64	0,18	0,50	0,17	-0,45
	Min log $\Gamma$	-13,62	-2,41	-4,61	49,30	4,34	-5,63	-2,34	8,81	-2,91	-2,47	-2,17	-3,44
	R	-0,22	0,61	1,05	0,14	0,31	-0,48	0,44	0,20	2,01	0,41	0,76	0,36
	Max R	-0,44	1,26	2,40	2,00	0,94	1,95	0,86	1,55	3,72	0,81	1,31	0,57
	Min R	-2,47	-0,05	-0,31	-1,73	-0,33	-2,90	0,01	-1,16	0,29	0,00	0,20	0,15

**Tabla 31: Coeficientes de mejores modelos del coeficiente de partición obtenidos con optimización genética. Las celdas en blanco corresponden a sistemas para los cuales los mejores resultados de la optimización genética no logran superar las metodologías de análisis topológico.**

		SISTEMAS DE DOS FASES ACUOSAS											
Tipo	Atributo	PEG-F 0,0%	PEG-F 0,6%	PEG-F 8,8%	PEG-S 0,0%	PEG-S 0,6%	PEG-S 8,8%	PEG-C 0,0%	PEG-C 0,6%	PEG-C 8,8%	PEG-D 0,0%	PEG-D 0,6%	PEG-D 8,8%
3D	log $\Gamma_0$	-	0,57	-0,45	-0,45	0,27	-	-	-0,43	-	-	-	-0,39
	Max log $\Gamma$	-	0,34	0,03	-0,10	-0,31	-	-	-0,07	-	-	-	0,21
	Min log $\Gamma$	-	0,80	-0,93	-0,81	0,84	-	-	-0,80	-	-	-	-0,98
	R	-	11,12	6,44	6,44	-6,22	-	-	10,08	-	-	-	5,23
	Max R	-	14,75	11,03	1,03	4,12	-	-	15,91	-	-	-	10,51
	Min R	-	7,49	1,85	1,85	-16,56	-	-	4,25	-	-	-	-0,05
Mod Lineal	log $\Gamma_0$	-0,21	-0,33	-0,41	-0,15	-0,18	-0,34	-0,29	-0,26	-0,36	-0,30	-0,29	-0,21
	Max log $\Gamma$	-0,03	-0,22	0,20	1,08	0,12	0,21	-0,03	-0,05	0,17	2,03	1,94	0,50
	Min log $\Gamma$	-0,40	-0,43	-1,02	-1,38	-0,47	-0,88	-0,55	-0,46	-0,89	-2,63	-2,52	-0,93
	R	11,52	-19,54	6,42	1,42	7,96	5,08	-11,06	-8,04	-6,07	1,54	1,75	2,52
	Max R	18,13	-14,83	12,82	6,02	16,47	10,38	-3,79	-3,14	0,57	7,84	9,01	7,17
	Min R	4,91	-24,26	0,02	-3,18	-0,56	-0,23	-18,33	-12,94	-12,72	-4,76	-5,51	-2,13

Los coeficientes de los modelos predictivos del coeficiente de partición en los sistemas ATPS toman valores positivos y negativos, dependiendo del dominio en el cual se encuentra la hidrofobicidad superficial media (ASA) de las proteínas, la que depende de cada escala.

## Escalas Generadas Asociadas a Mejores Modelos

En la Tabla 41 a 43 (ver Anexo O) se muestra las escalas a partir de las cuales se construyen los mejores modelos obtenidos. Al igual que para el caso de HIC algunos coeficientes de los resultados obtenidos con los algoritmos de análisis topológico son negativos, lo que denota un nivel de hidrofobicidad que sale del rango encontrado en los coeficientes de los 74 APVs.

A diferencia de HIC, para los sistemas de dos fases acuosas no se aprecia una relación significativa entre las correlaciones asociadas a los tipos de modelos y algoritmos (probabilidad de Fisher de 0,54 en un análisis de varianza de un factor entre los Coeficientes de Correlación de los distintos grupos de escalas). A pesar de esto, se distinguen pequeños *clusters* de escalas que tienen un Coeficiente de Correlación superior a 0,6 y 0,7. La interpretación de estos resultados es que existen escalas con distintas ubicaciones espaciales, y pocas escalas muy cercanas entre ellas. Sin embargo, el número de escalas cercanas entre ellas (medido a través del Coeficiente de Pearson) es mayor entre los resultados obtenidos en un mismo sistema-algoritmo-modelo.

### 7.1.5 Análisis Comparativo Para los Modelos HIC y ATPS

En total se generó 52 mejores modelos para los distintos tipos de algoritmos y modelos, los que se pueden desglosar por: sistema, tipo de algoritmo utilizado y tipo de modelo; según se muestra en la Tabla 32:

Tabla 32: Desglose de los modelos generados por tipo.

Tipo Sistema-Tipo Algoritmo-Tipo Modelo	N° Modelos	Variación Media MSE <sub>jk</sub> [-]	Variación Media Pearson [-]	% n° Casos Mejoras MSE <sub>jk</sub> [-]	% n° Casos Mejoras Pearson [-]
HIC-A. Topológico-3D	1	45%	4%	100%	100%
HIC-A. Topológico-Lineal	1				
HIC-A. Genético-3D	1	64%	5%	100%	100%
HIC-A. Genético-Lineal	1				
ATPS-A. Topológico-3D	12	70%	58%	100%	42%
ATPS-A. Topológico-Lineal	12				
ATPS-A. Genético-3D	12	57%	60%	58%	100%*
ATPS-A. Genético-Lineal	12				

\* Porcentaje referido a los casos que exhiben mejora del MSE<sub>jk</sub>.

Un primer resultado interesante es la disminución del Error de Jackknife en 42 de los 52 modelos (80,8%). En los sistemas cromatográficos el rango de mejora es de 37%-75% con un promedio de 54%. Por otro lado en los sistemas de dos fases acuosas el rango de mejora es de 11%-99,6% con un promedio de 64%.

Debido a que se utilizó el Error de Jackknife para buscar los mejores resultados, resulta interesante examinar con más detalle la variación de los Coeficientes de Pearson (ajuste) de los modelos, debido a que éstos no presentan una alta correlación.

Para los modelos cromatográficos los Coeficientes de Pearson del caso base de comparación son considerablemente altos (0,92 – 0,93), los que se logran aumentar en un porcentaje pequeño pero significativo (2,2% - 5,4%), ya que el máximo valor admisible para este coeficiente es 1, por lo tanto el máximo aumento posible es un 8,7%.

Para el caso de los sistemas de dos fases acuosas se obtuvo aumentos de mayor magnitud en el Coeficiente de Pearson, aunque también se obtuvo una alta variabilidad en los resultados (-96% a 300%). En este caso la línea base de comparación contiene valores del Coeficiente de Pearson entre 0,11 – 0,87 con un promedio de 0,66; lo cual explica la posibilidad de grandes aumentos. Por otro lado, la disminución se puede explicar debido a la falta de correlación entre el Coeficiente de Pearson y el Error de Jackknife. En el 82% de los casos en que disminuye el Coeficiente de Pearson se debe a algoritmos de análisis topológico. Al respecto, hay que considerar que si bien estos algoritmos generan nuevas escalas conservando propiedades topológicas, éstas pueden conservar buenas y malas propiedades presentes en las escalas originales, lo que puede generar ruido en los resultados. Esto se puede evitar en futuros trabajos, utilizando un Coeficiente de Pearson mínimo, dentro del criterio de selección de los mejores modelos-escalas.

A pesar de esta variabilidad, es interesante que para las mejores soluciones obtenidas con el *Genetic Algorithm*, para la cromatografía de interacción hidrofóbica, se obtiene mejoras en el Coeficiente de Pearson y el Error de Jackknife en 2 modelos de un total de 2; y en 14 modelos de un total de 24, en los sistemas de dos fases acuosas.

En el caso de las soluciones obtenidas a través de algoritmos de análisis topológico, se obtuvo mejoras en 2 modelos cromatográficos de un total de 2, considerando ambos coeficientes, y en 10 modelos de ATPS de un total de 24.

Los resultados que se presentó corresponden a los obtenidos para el mejor modelo lineal en cada caso. No se incluyó un apartado de análisis de diferencia de resultados obtenidos entre los distintos tipos de modelos lineales, ya que no se aprecian tendencias que los diferencien, aunque sí se puede resaltar que el modelo lineal III es el que obtiene los mejores resultados en HIC y en todos los sistemas ATPS a partir de las escalas obtenidas por el *Genetic Algorithm*. En el caso de los resultados obtenidos para ATPS a partir de escalas generadas con Algoritmos de Análisis Topológico, los mejores resultados se obtuvieron en un 50% con el modelo lineal II, en un 33,3% con el modelos lineal I y en un 16,7% con el modelos lineal III.

## ***7.2 Análisis de las Escalas Generadas que Permiten Obtener los Mejores Modelos***

A través de la metodología que se utiliza en este estudio, los mejores resultados fueron presentados para 13 sistemas físico-químicos, para los cuales se obtuvieron modelos del tipo 3D y lineales, y con algoritmos de 2 tipos (52 sistemas-tipo de modelo-tipo de algoritmo); lográndose un aumento de entre un 11% y un 99,6% del poder predictivo. De forma simultánea, dentro de los modelos con aumento del poder predictivo (42 de 52 casos), se obtuvo mejoras en el nivel de ajuste de entre un 4% y un 300%, en 28 casos de 42; por otra parte, el nivel de ajuste disminuyó en 14 de los 42 casos.

El éxito de la metodología empleada es significativo y mayor al logrado a través de otros estudios que modifican el tipo de modelo incorporando más información. Un ejemplo es el caso del trabajo de Riveros [71], que utiliza la hidrofobicidad y el Coeficiente de Solvatación para mejorar las predicciones, y que consigue una disminución del error asociado a la predicción de entre 4% y 34,1%, en un 38,4% de los modelos utilizados.

Los resultados mostrados en esta tesis, en la sección 7.1, comprueban la utilidad práctica de estos algoritmos, sin embargo, es complejo entender cómo fueron formadas las escalas a partir de las cuales se obtuvo los mejores modelos, lo que genera una resistencia natural a validar los resultados. Es por esto que en la presente sección se trabaja sobre los siguientes puntos:

- a. Determinar cuáles son las principales escalas con origen en la literatura que se utilizaron como precursoras en la generación de los modelos más exitosos. Este es el objetivo de la sección 7.2.1.
- b. Establecer la validez de los APVs precursores. Este es el objetivo de la sección 7.2.2.

### **7.2.1 Análisis del Origen de las Mejores Escalas**

El análisis del origen de las mejores escalas, corresponde al estudio de los APVs (*aminoacidic protein vector* más relevantes a través de los cuales se obtiene los mejores modelos desde el punto de vista de su capacidad de predicción. Los APVs más relevantes se entienden como aquellos seleccionados por los criterios descritos en la sección 6.5 como los más cercanos, determinado por la distancia euclidiana, o más influyentes. A su vez, los más influyentes corresponden a los APVs que contienen la mayor parte de la información sintetizada por las escalas a partir de las cuales se generó los modelos más exitosos. Este conjunto de APVs se pueden entender como claves para la generación de estas nuevas escalas, y sus propiedades.

La relevancia de este análisis radica en que permite sentar las bases para la discusión sobre la relación entre los APVs que generan las escalas más exitosas, lo que se encuentra en las secciones 7.3, y 7.4.

Esta sección se basa en las listas de APVs más influyentes y en las listas de APVs más cercanos que se obtuvo utilizando la metodología explicada en la sección 6.5. A partir de estas listas, es posible obtener los APVs más relevantes para el posterior análisis fenomenológico (sección 7.3). Sin embargo, las metodologías utilizadas presentan un nivel de abstracción semejante a los algoritmos de *clustering*, por lo que los resultados se analizan a través de técnicas desarrolladas en el presente trabajo, como son los gráficos de composición introducidos en la sección 6.5.4, y además, para facilitar la comprensión se incluye un ejemplo que correlaciona éstos gráficos con la ubicación espacial de un listas de APVs en el plano.

Los resultados de esta sección permiten realizar un análisis de la diferencia topológica entre los resultados obtenidos con el *Genetic Algorithm* y los algoritmos de *clustering*, específicamente los de redes neuronales.

### ***APVs Más Influyentes y Más Cercanos a las Mejores Soluciones***

A través de las metodologías descritas en la sección 6.5, es posible analizar cada uno de los resultados obtenidos con los algoritmos de *clustering* y el *Genetic Algorithm* donde se encuentran las mejores escalas, en conjunto con los APVs, para determinar cuáles son los más influyentes. La información obtenida requiere ser sintetizada para que sea posible un análisis visual. En la sección 6.5.3 se presentó la estructura escogida para mostrar los APVs.

En la Figura 26, a continuación, se muestra las listas de APVs más influyentes y de APVs más cercanos, los que están ordenados de manera que los más influyentes y cercanos se encuentran en la parte superior de la lista, desglosados por sistema, modelo y tipo de algoritmo empleado. Las listas se generaron a partir de las mejores escalas, cuyos Coeficientes de Pearson no disminuyeron de forma significativa.

LISTAS DE APVs POR NIVEL DE INFLUENCIA							
Genetic Algorithm		Genetic Algorithm		Análisis topológico		Análisis topológico	
ATPS - 3D	ATPS - LIN	HIC - 3D	HIC - LIN	ATPS - 3D	ATPS - LIN	HIC - 3D	HIC - LIN
10	16	15	12	51	7	59	22
7	9	3	10	58	10	60	
4	4	13	11	7	11	61	
5	5	18	17	8	12		
11	7		55	9	55		
13	6			10			
20	11			11			
6	12			16			
9	13			21			
12	15			25			
19	14			32			
15	10			37			
16	19			42			
55	2			44			
				45			
				46			
				53			
				57			

LISTAS DE APVs POR CERCANÍA							
Genetic Algorithm		Genetic Algorithm		Análisis topológico		Análisis topológico	
ATPS - 3D	ATPS - LIN	HIC - 3D	HIC - LIN	ATPS - 3D	ATPS - LIN	HIC - 3D	HIC - LIN
25	8	15	12	25	29	62	25
32	25	14	25	32	25	28	29
8	32	25	32	8	32	36	22
34	34	13	55	12	60	25	23
20	23	69	10	23	61	34	61
41	20	72	8	53	23	24	60
70	22	8	34	67	22	27	32
42	26	32	30	26	30	32	30
49	36	34	62	42	67	33	67
55	30	42	23	51	42	49	34
60	42			69			
	61			70			
	70			72			
	12			34			
	24			41			
	41			49			
	49			58			

Figura 26: Síntesis de los APVs más influyentes y cercanos a las mejores soluciones. Los números en fondo azul corresponden a escalas hidrofóbicas, los que están en fondo rojo a escalas de origen conformacional, y los en fondo verde a escalas de origen estadístico.

En la Figura 26 cada APV se representa mediante un número que lo identifica (ver Anexo P) en un fondo de color que corresponde a su clase.

La información contenida en la Figura 26 permite realizar varios análisis, dentro de los cuales son de gran utilidad los gráficos de la composición de APVs, cuya metodología se detalla en la sección 6.5.4. A continuación, se presenta en la Figura 27 un gráfico de composición de APVs para cada uno de los tipos de sistemas físico-químicos, según su influencia o cercanía a las mejores escalas. Esta figura se construyó generando una lista única a partir de la frecuencia y grado de influencia o cercanía de cada APV.

## Composición Global de APVs Por Influencia y Cercanía Espacial

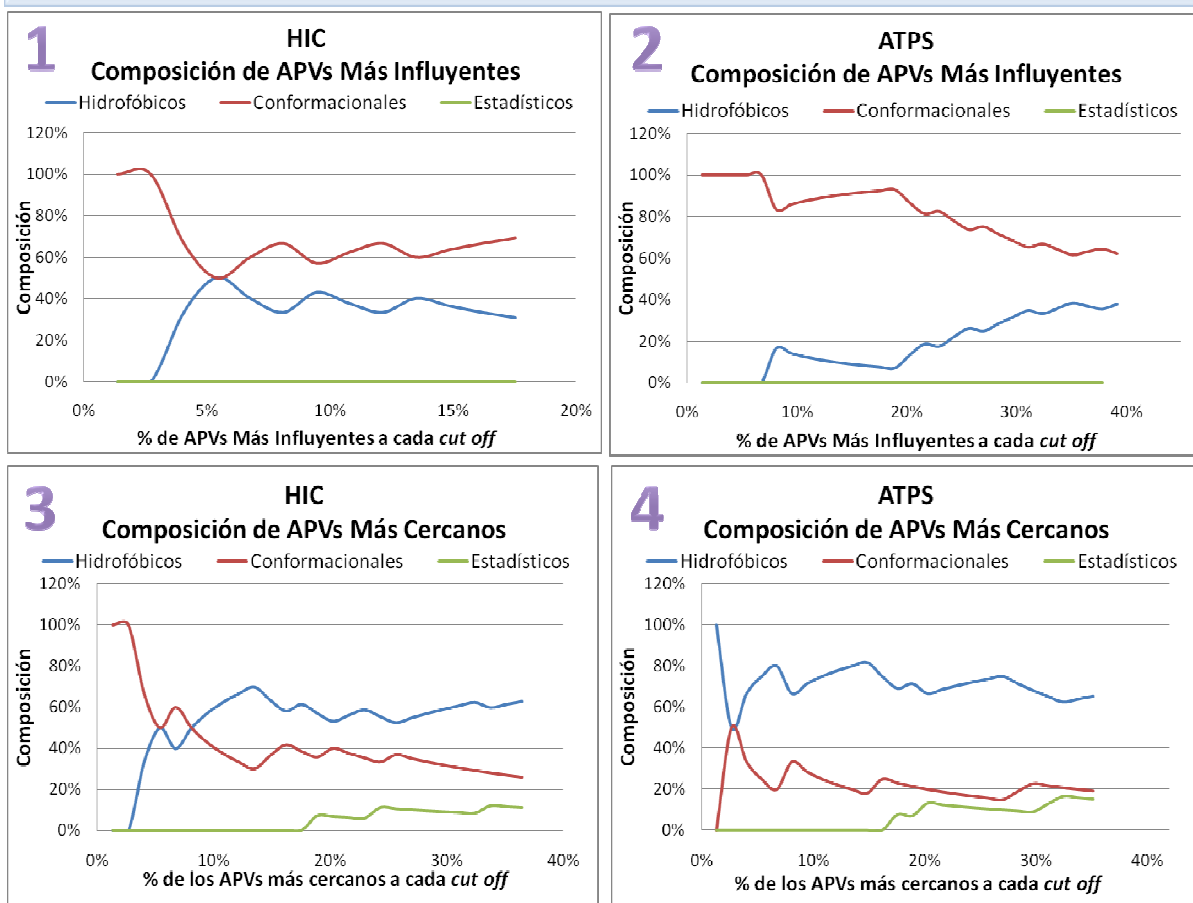


Figura 27: Composición global de APVs para ATPS y HIC. En azul se representa el porcentaje de APVs hidrofóbicos, en rojo los de tipo conformacional y en verde los del tipo estadístico.

En la Figura 27 se muestran cuatro gráficos. En los primeros dos se relaciona la composición de los APVs con el porcentaje de los APVs más influyentes a un *cut off* (que en este caso indica el número de APVs más influyentes incluido en el conjunto). En los gráficos 3 y 4 se muestra la composición de APVs versus el porcentaje de APVs más cercanos a un *cut off* (que en este caso representa un radio dentro de cuya hiper esfera se encuentran el conjunto de APVs).<sup>1</sup>

En las listas de la Figura 27 se observa que los mejores resultados obtenidos tanto para la cromatografía de interacción hidrofóbica como los sistemas de dos fases acuosas están influenciadas principalmente por APVs de origen conformacional, sin embargo, las zonas donde están ubicados contienen un mayor porcentaje de APVs hidrofóbicos respecto a los APVs conformacionales, porcentaje que se mantiene relativamente constante con la distancia. Tanto al comparar los gráficos 1 y 3 de la figura 27, como también los gráficos 2 y 4 de la misma figura, se aprecia una inconsistencia entre la ubicación espacial de las mejores escalas generadas y los APVs más influyentes a éstas.

<sup>1</sup> Para una descripción detallada de la metodología usada para generar éstos gráficos, ver la sección 6.5.4

Esta inconsistencia se puede apreciar también en la Figura 26, donde están las listas de APVs más influyentes que corresponden a los mejores resultados generados con el *Genetic Algorithm*. Para el caso de las listas de APVs más influyentes generadas con los algoritmos de redes neuronales, se aprecia una influencia de una variedad de tipos de APVs, por lo que no es tan claro si existe consistencia entre la ubicación espacial de las mejores escalas generadas y los APVs más influyentes a éstas.

En la siguiente sección se profundiza el análisis de las diferencias entre las listas de APVs obtenidas con el *Genetic Algorithm* y los obtenidos con algoritmos de redes neuronales.

### ***Genetic Algorithm* v/s Algoritmo Topológicos**

En la sección anterior se muestran los gráficos de composición global de APVs en relación a su influencia y cercanía a las mejores escalas obtenidas, los que presentan una inconsistencia que se aprecia en la Figura 26.

Para realizar una comparación entre las listas de APVs más influyentes y más cercanos obtenidas con los algoritmos de redes neuronales y el *Genetic Algorithm*, es importante considerar qué tan relevante es la consistencia entre los gráficos de composición de APVs, obtenido a través de la búsqueda de los APVs más influyentes, y los más cercanos a las mejores soluciones encontradas.

Los APVs más influyentes corresponden a los utilizados como base por los algoritmos generadores de nuevas escalas. Los algoritmos de redes neuronales utilizados, los que en adelante también se mencionan bajo el nombre de “Algoritmos de Análisis Topológico”, realizan una transferencia de las propiedades topológicas entre los APVs y las nuevas escalas [16, 126]; entre más influencia tiene un APV sobre una escala, mayor será la transferencia de las propiedades topológicas. Esta transferencia se realiza en los algoritmos de redes neuronales acercando la nueva escala al APV en cuestión, debido a que la topología es un conjunto de propiedades que derivan de la ubicación espacial relativa. Por otro lado, al sintetizar las listas de APVs (originalmente 78) se utiliza la misma metodología para la búsqueda de los APVs más influyentes y cercanos, la que varía ligeramente entre los tipos de algoritmos.

Por lo tanto, hay una relación importante entre la influencia que pueda tener un APV y su cercanía con una escala.

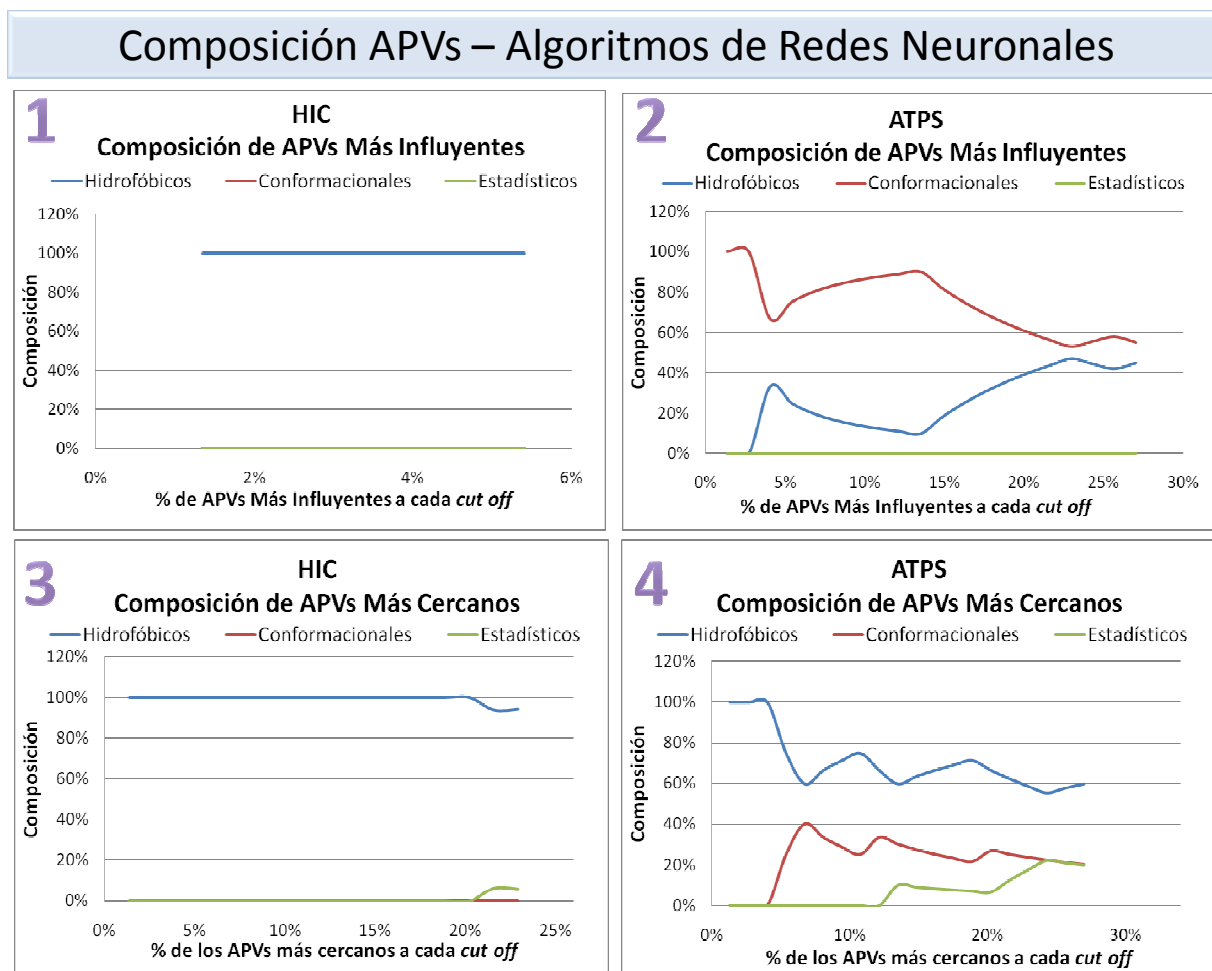
Respecto al carácter de la relación entre influencia y cercanía de los distintos tipos de APVs, lo esperable es que:

- Si la influencia de una mejor escala o un conjunto de mejores escalas se debe a APVs de sólo un tipo, entonces los APVs más cercanos debieran ser casi exclusivamente del mismo tipo. En adelante a esto se le llamará Consistencia Tipo I.
- Si la composición de APVs, cuya influencia en una mejor escala o un conjunto de mejores escalas, es muy baja en APVs de un tipo, entonces los APVs más cercanos no debieran tener una composición significativa para este tipo de APV. En adelante a esto se le llamará Consistencia Tipo II.
- Si la composición de APVs, cuya influencia en una mejor escala o un conjunto de mejores escalas, incluye de forma significativa más de un tipo de APV,



entonces los APVs más cercanos a una mejor solución debieran incluir de forma significativa estos mismos tipos de APVs. En adelante a esto se le llamará Consistencia Tipo III, la cual es estudiada en más detalle en la sección 7.2.1.

En base a lo discutido hasta ahora, en la presente sección se evaluará los gráficos de composición por influencia y cercanía para las escalas más exitosas, separando las encontradas por los algoritmos de redes neuronales y el *Genetic Algorithm*, en la Figura 28 y en la Figura 29 respectivamente.



**Figura 28: Composición de APVs más influyentes y cercanos de las mejores soluciones encontradas con los algoritmos de redes neuronales. En azul se representa el porcentaje de APVs hidrofóbicos, en rojo los de tipo conformacional y en verde los del tipo estadístico.**

En la Figura 28 se observa cómo en los resultados obtenidos para los algoritmos de redes neuronales existe una perfecta consistencia entre la composición de los APVs más influyentes y cercanos asociados a los mejores resultados en una cromatografía de interacción hidrofóbica, esta es una Consistencia Tipo I. Para los APVs asociados a los sistemas de dos fases acuosas también se observa consistencia, pero Consistencia Tipo III. En este caso la consistencia es menos intuitiva, razón por la cual en la sección 7.2.1 se explica cómo es posible, en base a una misma solución, obtener un gráfico de composición por influencia y cercanía como los de la Figura 28 (gráficos 2 y 4).

## Composición APVs – Algoritmo Genético

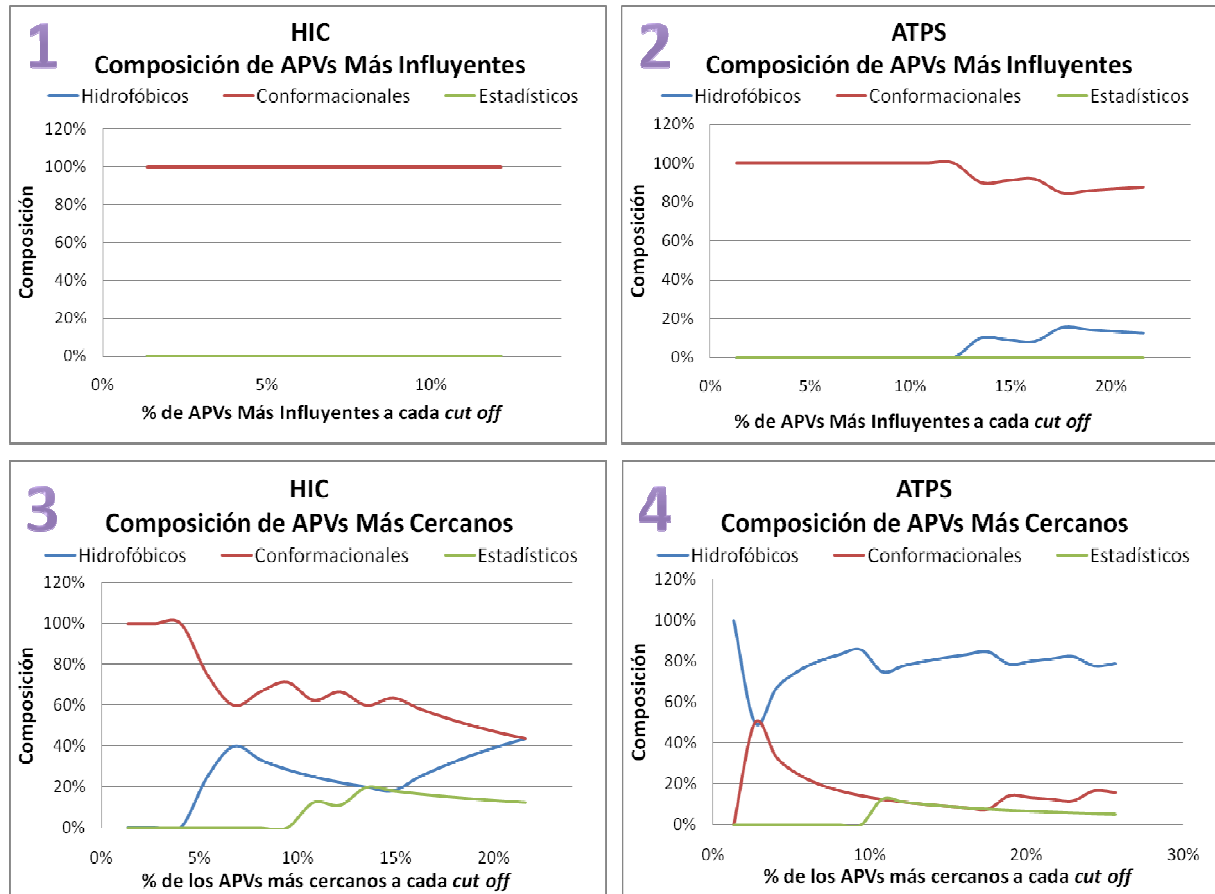


Figura 29: Composición de APVs más influyentes y cercanos de las mejores soluciones encontradas con el *Genetic Algorithm*. En azul se representa el porcentaje de APVs hidrofóbicos, en rojo los de tipo conformacional y en verde los del tipo estadístico.

A diferencia de las composiciones encontradas para los algoritmos de redes neuronales, en la Figura 29 se aprecia una clara inconsistencia para ambos sistemas. En el caso de la composición de los APVs más influyentes en HIC se aprecia una influencia exclusiva por escalas del tipo conformacional, sin embargo, la ubicación espacial de las mejores escalas está entre una zona con APVs conformacionales (los más cercanos), dos zonas con APVs hidrofóbicos (con un *cluster* cercano y otro en la zona más lejana considerada), y una con APVs del tipo estadístico (entre las dos zonas de *clusters* hidrofóbicos). En el caso de los gráficos de composición de APVs para los sistemas de dos fases acuosas la inconsistencia es aún más clara, dado que la relación entre los APVs más influyentes y los más cercanos es opuesta a la esperada. La influencia es prácticamente conformacional, y los APVs más cercanos son principalmente de carácter hidrofóbico.

Por otro lado, en la Figura 28 se observa una diferencia importante entre los gráficos de composición de APVs asociados a HIC (1 y 3) y los asociados a ATPS (2 y 4). Mientras los gráficos 1 y 2 se aprecia una composición casi exclusiva de APVs del tipo hidrofóbicos, en los gráficos 3 y 4 se aprecia una composición no despreciable en APVs del tipo conformacional. Esto se puede deber a que el sistema ATPS es más complejo que HIC, en el cual ocurren interacciones entre una mayor variedad de moléculas, por

lo que las diferencias conformacionales entre cada proteína en ATPS puede tener mayor relevancia. En las secciones 7.3, se realiza una discusión más extensa sobre este punto.

La principal conclusión que se obtiene en esta sección es que las mejores escalas producidas por el *Genetic Algorithm*, y los APVs más influyentes o cercanos a éstas, no deben ser utilizadas en un análisis sobre la hidrofobicidad y la relación de ésta con los modelos utilizados, debido a que no existe consistencia entre la ubicación espacial de las mejores escalas generadas y los APVs más influyentes sobre éstas.

Para una discusión más extensa sobre la validez de las mejores escalas obtenidas con los algoritmos de redes neuronales, y su uso en las discusiones posteriores, ver la sección 7.2.2.

### ***Interpretación de Gráficos de Composición y Consistencia Tipo III***

En la sección anterior, en la Figura 28 se aprecia la variación de la composición de APVs más influyentes y cercanos, al considerar distintos conjuntos clasificados por los criterios correspondientes. Dentro de la discusión se determinó que existe Consistencia Tipo III entre la composición de APVs más influyentes y cercanos a las mejores soluciones de las ATPS. Si bien los gráficos 3 y 4 de la Figura 28 están compuestos por APVs del tipo hidrofóbico y conformacional, la composición varía de forma distinta con respecto a la influencia y la cercanía, y alcanza órdenes de magnitud diferentes. De hecho, a simple vista no es intuitivo comprender que ambos gráficos provengan de un mismo conjunto de mejores escalas.

A modo de facilitar la comprensión y, al mismo tiempo, validar las conclusiones de la sección 7.2.1, en la presente sección se analiza cómo se puede interpretar estos gráficos de composición de APVs, en función del ejemplo mostrado en la Figura 30.

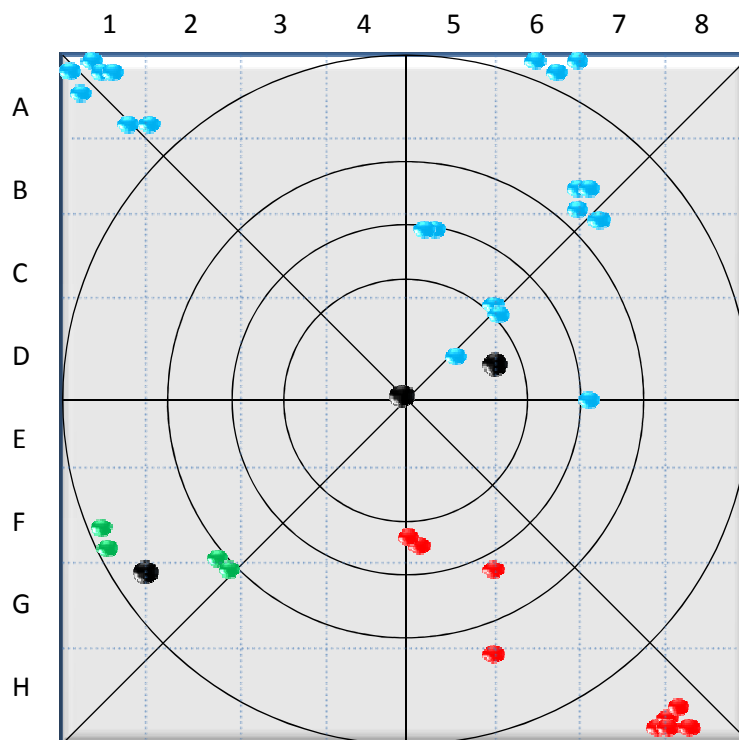


Figura 30: Ejemplo de distribución en el plano de nuevas escalas (puntos negros), y APVs hidrofóbicos (celeste), conformacionales (rojo), y estadísticos (verdes).

En la Figura 30 se muestra un ejemplo de una posible distribución de escalas y APVs de los tres tipos, que permite explicar simultáneamente los gráficos de composición 3 y 4 de la Figura 28. Este plano está definido dentro de un cuadrado, que a su vez está particionado por tres círculos y tres rectas con el objetivo de permitir un rápido cálculo intuitivo (no exacto) de las distancias entre la escala en evaluación, que corresponde al punto negro en el centro del plano, y el resto de los APVs. Los APVs están representados mediante puntos con la misma nomenclatura de colores que los gráficos de composición. Los azules corresponden a escalas de hidrofobicidad, los rojos a conformacionales y los verdes a escalas con origen estadístico.

Del análisis acucioso del ejemplo de la Figura 30, que se realiza en la Figura 31 y en la Figura 32, se puede obtener información clarificadora para la interpretación de la Consistencia Tipo III.

Inicialmente no se considerarán los dos *clusters* de APVs más distantes del centro (uno hidrofóbicos en la esquina superior izquierda y uno conformacional en la esquina superior derecha), ni las escalas adicionales a la del centro del plano (puntos negros). De esta manera se obtiene la Figura 31.

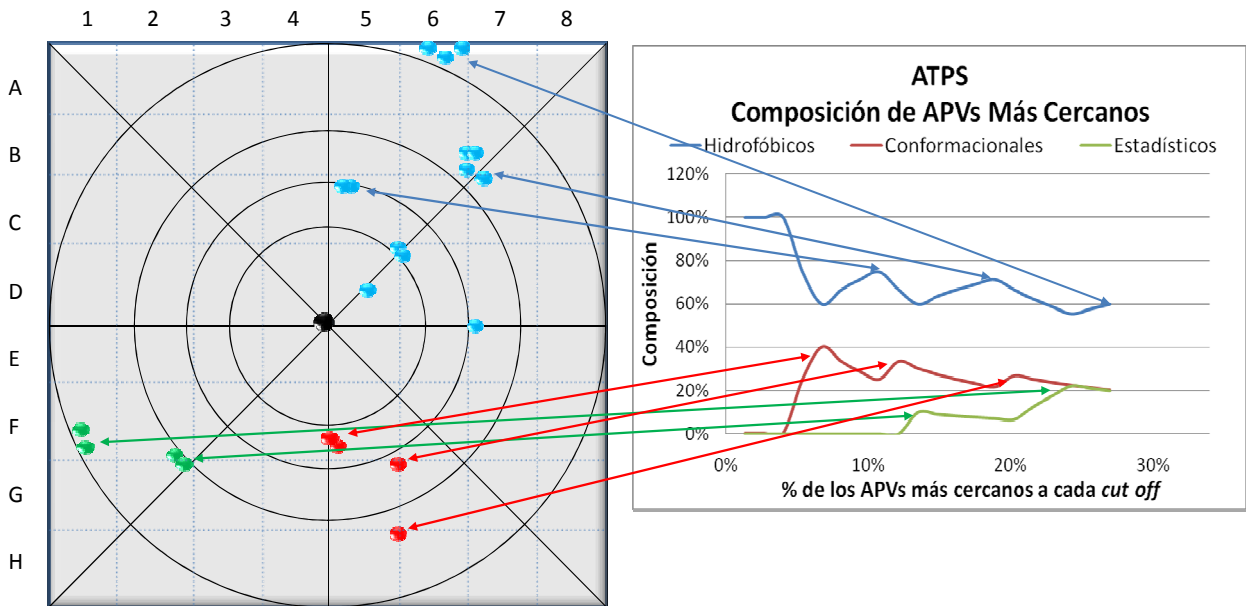


Figura 31: Esquema de concordancia entre el gráfico n°4 de la Figura 28 y el ejemplo de la Figura 30, de escalas y APVs en el espacio.

En la Figura 31 se muestra el espacio conformado por una escala y un grupo de APVs, y se relaciona con el gráfico número 4 de la Figura 28, que muestra la composición de APVs por cercanía para los mejores resultados obtenidos en ATPS a través de los algoritmos de redes neuronales. En esta figura se muestra claramente una correlación entre distintos *peaks* y valles, con la ubicación de los APVs en el espacio.

El total de APVs con los que se trabajó son 74, de los cuales el 29% corresponde a 21. Los APVs más cercanos a la escala en el centro son del tipo hidrofóbicos, y se encuentran en el primer cuadrante. (D5) Luego hay dos APVs conformacionales que corresponden al primer *peak* de estos APVs en la gráfica (F-4/5 y F-6). Luego siguen 3 APVs más del tipo hidrofóbicos que se encuentran en los límites de primer cuadrante (2 en C-5 y uno en E/D-7), los que coinciden con el primer *peak* de este tipo de APV en la gráfica.

En este punto hay un 11% de APVs sobre 74, lo que equivale a 8 APVs, lo que coincide con la cantidad explicada hasta ahora. De estos 8 solo 2 son conformacionales (25%) y 6 son hidrofóbicos (75%), lo que concuerda con el gráfico.

Los siguientes 2 *peak* están explicados por el APV del tipo conformacional (rojo) en el límite de la posición G-5, y los 2 APV del tipo estadístico en la posición G/F-2. En el extremo opuesto se encuentra cuatro APVs hidrofóbicos que explican el segundo *peak* hidrofóbico (B-6/7). Los últimos tres *peak* se deben a los APVs o *clusters* de APVs restantes del tipo conformacional en H-5 (que coincide con el 20% de APVs), estadísticos en F-1 e hidrofóbicos en A-6.

El 20% de APVs sobre un total de 74 corresponde a 15, de los cuales se observan 10 hidrofóbicos (67%), 3 conformacionales (20%) y 2 estadísticos (13%); lo que nuevamente tiene perfecta concordancia con el gráfico de composición.

Finalmente, del 29% (21 APVs) hay 13 APVs hidrofóbicos (61%), 4 conformacionales (19%) y 4 estadísticos (19%); lo cual nuevamente concuerda con el gráfico de composición.

En conclusión, a través del ejemplo propuesto en la Figura 30 se puede explicar perfectamente el resultado del gráfico de composición de APVs por cercanía en torno a las mejores soluciones. Por otro lado, a través de la Figura 31 se ilustra cómo esta misma escala y los mismos APVs, y también algunos más distantes, pueden explicar el gráfico de composición por influencia de la Figura 28.

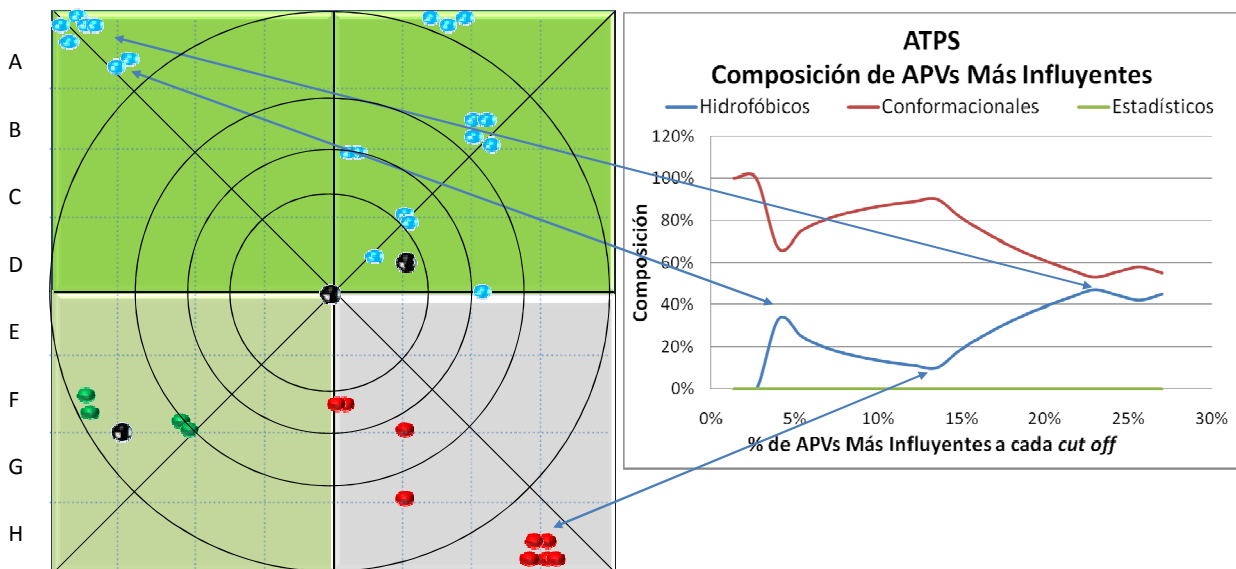


Figura 32: Esquema de concordancia entre el gráfico n°2 de la Figura 28 y el ejemplo de la Figura 30, de escalas y APVs en el espacio.

Para entender cómo la escala central de la Figura 32 puede estar asociado a un gráfico de composición de APVs más influyentes, como el mostrado en la misma figura, es necesario considerar los conceptos básicos a través de los cuales los algoritmos de redes neuronales utilizan los APVs para influir sobre la topología de las nuevas escalas (ver sección 3.3.5), y cómo se obtiene las listas de APVs más influyentes (sección 6.5). En resumen, y de forma simplificada, la transferencia topológica se produce acercando la escala al APV que influye en él, para lo cual se identifica la escala más cercana a cada APV. En la Figura 32 se observa que para todos los APVs del primer cuadrante existe otra escala más cercana (punto negro en D-6), y para los APVs de tipo estadístico (del tercer cuadrante) también existe otra escala más cercana en F/G-2. De esta manera, la escala central sólo se ve influida por los APVs del segundo y cuarto cuadrante.

Dentro de los APVs que influyen en la escala, los más cercanos son 4 del tipo conformacional, y luego 2 del tipo hidrofóbicos (A-2), los que explican el primer *peak*. Luego, siguen 5 APVs conformacionales y 5 hidrofóbicos, los que explican los siguientes 2 *peaks*. Al considerar el 23% de APVs más influyentes (17 APVs) de la Figura 32, se aprecia un total de 16 APVs influenciando a la escala central, de los cuales 9 son conformacionales (56%) y 8 son hidrofóbicos (50%).

Algunas de las aproximaciones y simplificaciones realizadas en este ejemplo son:

- Se trabajó en  $\mathbb{R}^2$  en vez de  $\mathbb{R}^{20}$ .
- La distribución de los APVs es un ejemplo que no está basado en la verdadera distribución de los 74 APVs utilizados, aunque sí en varias de sus características tales como: zona características para los APVs hidrofóbicos separada de las zonas donde se encuentran los APVs conformacionales y estadísticos (zona superior del plano); una zona donde se encuentran APVs de distintas clases en ubicaciones espaciales muy cercanas entre sí, en relación a las zonas más características para cada tipo de APV (razón por la que no se ve el gran cluster de escalas hidrofóbicas). Lo anterior se basa en un análisis de componentes principales de los 74 APVs realizado por Salgado et. al. [9].
- La búsqueda de escalas más influyentes es un proceso iterativo, el cual fue reducido a un paso.

Si bien en este ejemplo se ha hecho varias aproximaciones y simplificaciones, a través de él es posible verificar de forma intuitivo que el resultado obtenido en los gráficos 4 y 5 de la Figura 28, es viable a partir de una misma escala (o un grupo de escalas), y por lo tanto poseen una Consistencia Tipo III, según lo definido en la sección 7.2.1.

### ***Identificación de los APVs Más Relevantes para Generar las Mejores Soluciones***

En la sección 7.2.1 se expone las listas de APVs más relevantes (influyentes y cercanos) para la obtención de las mejores escalas. Luego, se determina que los gráficos de composición de APVs no son consistentes para las mejores escalas obtenidas con el *Genetic Algorithm*, pero sí son consistentes las generadas con los algoritmos de redes neuronales.

Para identificar los trabajos más relevantes se utiliza la Figura 26, y coherentemente con los resultados expuestos, se descarta las listas de APVs más influyentes en las mejores escalas obtenidas con el Genetic Algorithm.

En la Figura 26 se puede distinguir algunos APVs de mayor coincidencia entre las listas de APVs generadas por influencia y cercanía, los que se identifican en la Figura 33, a continuación.

LISTAS DE APVs POR NIVEL DE INFLUENCIA A MEJORES SOLUCIONES				LISTAS DE APVs POR CERCANÍA A MEJORES SOLUCIONES							
Análisis topológico		Análisis topológico		Genetic Algorithm		Genetic Algorithm		Análisis topológico		Análisis topológico	
ATPS- 3D	ATPS - LIN	DRT 3D	DRT LIN	ATPS- 3D	ATPS - LIN	DRT 3D	DRT LIN	ATPS- 3D	ATPS - LIN	DRT 3D	DRT LIN
51	7	59	22	25	8	15	12	25	29	62	25
58	10	60		32	25	14	25	32	25	28	29
7	11	61		8	32	25	32	8	32	36	22
8	12			34	34	13	55	12	60	25	23
9	55			20	23	69	10	23	61	34	61
10				41	20	72	8	53	23	24	60
11				70	22	8	34	67	22	27	32
16				42	26	32	30	26	30	32	30
21				49	36	34	62	42	67	33	67
25				55	30	42	23	51	42	49	34
32				50	42			69			
37					61			70			
42					70			72			
44					12			34			
45					24			41			
46					41			49			
53					49			58			
57											

Figura 33: Listas de APVs más relevantes (influyentes y cercanos) para mejores escalas generadas con algoritmos de redes neuronales.

En base a la Figura 33 es posible realizar un análisis de frecuencia de APVs coincidentes entre las listas de APVs de mayor influencia y mayor cercanía, cuyo resultado final se puede ver en la Tabla 33.

Tabla 33: APVs seleccionados para análisis teórico (ver referencias de cada APV en Anexo Q).

N° APV	N° Repeticiones	
	Individuales	En Cluster
25	9	6
32	9	6
60	4	3
61	4	3
10	3	2
11	2	2
8	6	-
42	6	-
22	4	-
12	4	-

En este análisis se consideró escalas con un mínimo de 4 coincidencias. Un caso muy especial son los APVs que aparecen juntos en más de una oportunidad, como los APVs 25 y 32 [143, 144] que aparecen juntos 6 veces, los 60 y 61 [145] que aparecen juntos 3 veces, y finalmente la 10 y la 11 [150, 151] que aparecen juntos en 2 oportunidades, aunque éstos últimos sólo aparecen en las listas por influencia.

### 7.2.2 Análisis Matemáticos y Bibliográficos de los APVs Más Cercanos

De acuerdo a lo presentado en la sección 7.2.1, existe evidencia que indica que los APVs más relevantes obtenidos para las mejores escalas, que se generó a partir de algoritmos de redes neuronales, se pueden utilizar para realizar un análisis detallado de los APVs y su relación con la hidrofobicidad y los tipos de modelos utilizados. Sin embargo, en esta sección se recopila los argumentos mostrados, más otros adicionales, para validar los APVs.



La necesidad de una sección de validación radica en que las metodologías utilizadas tienen un alto nivel de abstracción gracias al uso de herramientas de *clustering* y optimización. Además, se diseñó e implementó una metodología específica para la obtención de los APVs más relevantes, lo que se explica en la sección 6.5.

En las sub secciones siguientes se expone las principales evidencias.

### Validación del Ordenamiento y Selección de APVs en Base a Frecuencia

Como se comentó con anterioridad, la Figura 26 se obtuvo como resumen de los resultados de un estudio de frecuencia de los APVs a partir de los 78 listados originales ordenados por influencia y de los 78 ordenados por cercanía (ver sección 6.5). Si consideramos sólo los APVs obtenidos con algoritmos que realizan análisis topológico, entonces, la Figura 26 debe poseer un orden lógico que permita obtener los APVs más relevantes, según se ilustra en la Tabla 35.

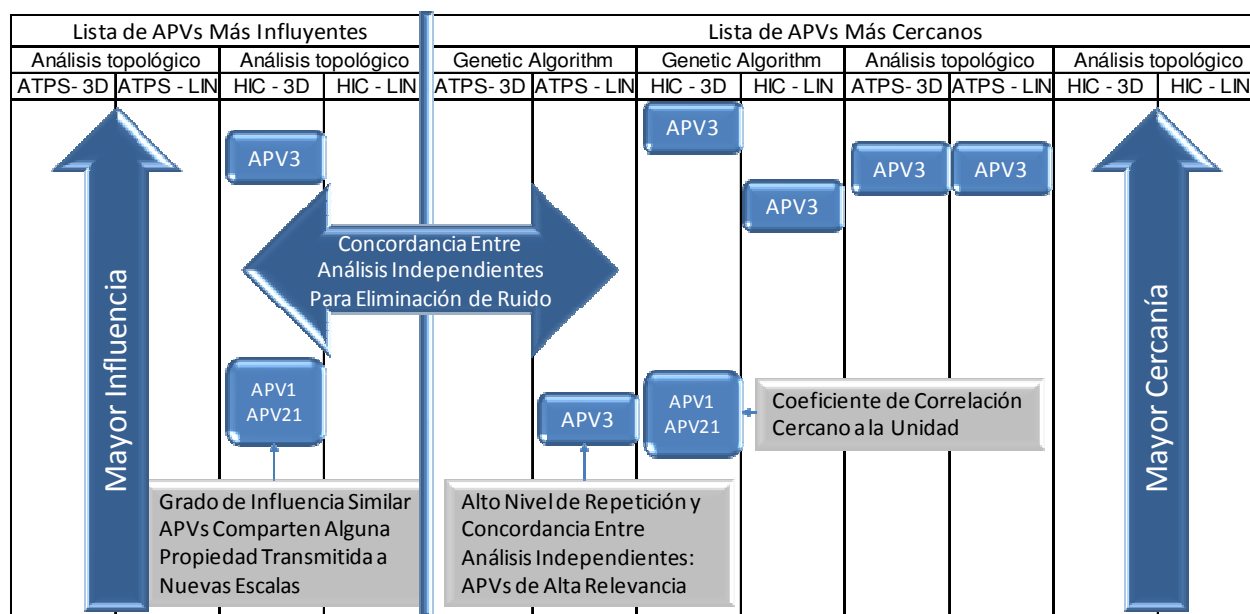


Figura 34: Esquema de contenido y orden esperado de las listas de APVs de la Figura 33.

En la Figura 34 se muestra un diagrama de las listas de APVs, y el orden o lógica intrínseca esperable, lo que responde a la metodología utilizada para generar estas listas:

- Se utilizó la frecuencia sobre los APVs más influyentes para ordenar las listas (flecha vertical a la izquierda en la Figura 34)
- En los APVs más cercanos el orden por cercanía es directo (flecha vertical a la derecha en la Figura 34).
- En base a los dos puntos anteriores, se favorece la aparición de APVs muy cercanos formando clusters en las listas ordenadas por influencia. Además, estos APVs deberían exhibir un alto Coeficiente de Correlación. Para el caso de las listas de APVs ordenadas por influencia esto es menos intuitivo, sin embargo, se puede ver ejemplos de estos tipos de cluster en la sección 7.2.2.

De acuerdo a la Figura 34 es posible definir grupos de APVs concordantes, es decir, APVs que resaltan por aparecer repetidamente en las distintas listas, ya sean las

obtenidas a partir de los conceptos de influencia y/o cercanía. Se distinguen los siguientes tipos de concordancia:

- Concordancia Tipo I: se obtiene de la coincidencia entre los APVs de los análisis independientes desarrollados a partir de los conceptos de influencia y cercanía. Se espera que no sea producto del azar y, por lo tanto, que la concordancia exista sólo para un número reducido de APVs.
- Concordancia Tipo II: se obtiene de la coincidencia entre los APVs de las listas obtenidas para los distintos modelos (3D o lineal) y un mismo sistema experimental (HIC o ATPS); o para los APVs de las listas obtenidas para modelos iguales (3D o Lineal) y sistemas experimentales distintos (HIC o ATPS).

Según esta definición, la Concordancia Tipo II se puede encontrar al analizar internamente las listas de APVs obtenidas a partir de los conceptos de cercanía e influencia, o de la coincidencia entre los APVs de los análisis independientes desarrollados a partir de los conceptos de influencia y cercanía. Debido a lo anterior, se puede distinguir los siguientes casos:

- Concordancia tipo II, no Tipo I
- Concordancia simultanea del tipo I y II

Es posible estudiar ambas concordancias a la vez. Para ello se utiliza la Concordancia Tipo II al evaluar la frecuencia de APVs en las listas ordenadas solo por un análisis de cercanía o influencia (ver APV<sub>3</sub> en Figura 34). Luego, sobre este análisis, se observa la frecuencia de los APVs con Concordancia Tipo I.

A partir de los conceptos previamente definidos se analiza la frecuencia de los APVs de cada tipo en la Figura 33, cuyo resultado se muestra a continuación en la Figura 35.

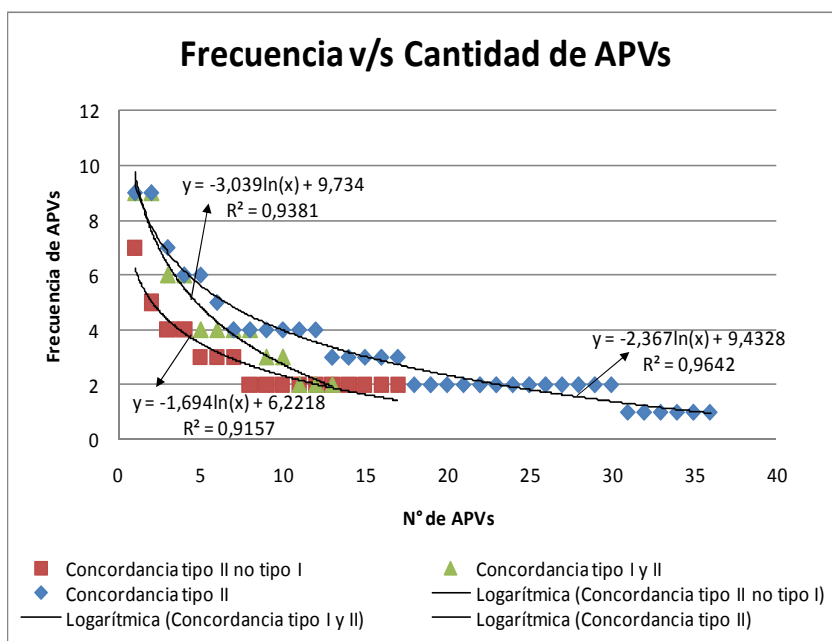


Figura 35: Evaluación de Concordancia Tipo I y Concordancia Tipo II.

En la Figura 35 se muestra que la distribución de frecuencia de Concordancia Tipo II no es producto del azar, ya que sigue una función logarítmica de pendiente negativa. La función logarítmica indica que solo un pequeño grupo de APVs tiene alta frecuencia, y que la mayoría tienen una frecuencia baja o nula, lo que indica la existencia de un grupo reducido de APVs más relevantes (influyentes y cercanos), y un gran número de APVs de relevancia menor.

Se puede contrastar el resultado obtenido para la concordancia del tipo II con la del tipo I y II o la del tipo II que no tienen Concordancia Tipo I (ver Figura 35). Nuevamente, los resultados respaldan que los APVs con Concordancia Tipo I y II, son el grupo de mayor frecuencia (y en este caso, los más relevantes). Por el contrario, en los APVs que tienen Concordancia Tipo II pero no Tipo I, los niveles de frecuencia bajan. La misma conclusión se puede hacer respecto al nivel de selectividad, lo cual se ve reflejado en la magnitud del coeficiente que acompaña al logaritmo, el cual aumenta al considerar Concordancia Tipo I y II, pero disminuye al considerar Concordancia Tipo I y no II.

Se puede estudiar con mayor detalle la Concordancia Tipo I, al comparar el coeficiente del logaritmo de la lista de APVs obtenidas con los criterios I y II, pero considerando solo modelos lineales o modelos tipo 3D. Los resultados se muestran en la Tabla 34.

**Tabla 34: Variación del coeficiente del logaritmo entre modelos tipo 3D y lineales. A partir de éstos se puede ver que se mantiene el nivel de selectividad de los APVs más cercanos e influyentes.**

Modelos	Coeficiente del logaritmo [Frec. APVs/log(n° APVs)]
Solo tipo 3D	-3,265
Solo tipo Lineal	-2,983

Al comparar los resultados de la Tabla 34 con los coeficientes de la Figura 35, se aprecia que aún después de reducir el número de listas de APVs incluidas en el análisis, utilizando sólo las que corresponden a modelos que utilizan información tridimensional de proteínas o sólo modelos lineales que utilizan sólo la composición de las proteínas; el nivel de selectividad se mantiene por sobre el encontrado al considerar únicamente Concordancia Tipo II, o Concordancia Tipo I, no Tipo II. Esto se produce a pesar que se redujo los niveles de frecuencia por un menor número de listas. En atención a lo anterior es posible concluir que el nivel de selectividad alto es un parámetro robusto frente a cambios en el número de modelos y sistemas utilizados para evaluar la frecuencia derivada de la Concordancia Tipo II.

La conclusión final de la presente sección es que las tablas de las Figura 26 y la Figura 33 poseen información que no es producto del azar, y cuyo ordenamiento permite un análisis para obtener los APVs más relevantes (influyentes y cercanos) para el análisis que se presenta en las secciones siguientes.

### ***Análisis Bibliográfico - Teórico de APVs en Clusters en Figura 33***

En la Figura 33 se observa un pequeño grupo de APVs que aparecen agrupados en forma reiterada, lo que se espera no sea producto del azar. En efecto, a continuación se muestra evidencia de la relación que éstos poseen:

- Los APVs n° 25 y 32 [143, 144], que corresponden a los de mayor frecuencia, son ambas escalas de hidrofobicidad obtenida por medios experimentales. Pese a que estas escalas se obtuvieron a través de sistemas distintos (HIC y ATPS), éstos poseen un alto Coeficiente de Correlación (0,91). La escala n° 25 fue obtenida por Fouchere y Pliska a través del cálculo de la constante de Hansch ( $\pi$ ) en octanol/agua [143, 189], utilizando la estructura N $\alpha$ -acetil-amino-acido amida para los 20 aminoácidos. Por su parte, la escala n° 32 es fue obtenida por Parker et. al. [144] a través de una HPLC, al correlacionar el tiempo de retención con actividad antigénica utilizando el péptido Ac-Gly-X-X-(Leu)3-Lys2-Amine, donde X es el aminoácido a ser estudiado.
- Los APVs n° 60 y 61 son escalas hidrofóbicas obtenidas por Jesior [145] en un mismo estudio sobre la vecindad de aminoácidos, utilizando distintos diámetros para definir la vecindad. El cálculo de hidrofobicidad se desarrolla a partir de un proceso iterativo que utiliza la siguiente expresión: n° a. a. hidrofóbicos/n° a. a. hidrofílicos en la vecindad de cada aminoácido, estadístico que converge independientemente del punto de partida. Las distintas escalas se obtienen al considerar distintos rangos del radio en el cálculo del estadístico. Ésta relación entre las escalas también se ve reflejado en su Coeficiente de Correlación, que es de 0,99.
- Finalmente, los APVs n° 10 y 11 [150, 151] son escalas conformacionales basadas en trabajos de predicción de estructura secundaria, en los cuales las escalas generadas reflejan el potencial de cada aminoácido de pertenecer a distintas estructuras secundarias. En el trabajo de Chou y Fasman (escala n° 10) el potencial se calcula para cuatro tipos de estructuras secundarias:  $\alpha$ -hélice,  $\beta$ -sheet, turn, y coil. En el estudio de Levitt (APV n° 11) el potencial se calcula sólo para tres tipos de estructuras secundarias:  $\alpha$ -hélice,  $\beta$ -sheet y turn, pero a diferencia del estudio de Chou y Fasman incluye los 20 aminoácidos. Esta diferencia puede explicar que el Coeficiente de Correlación de estas escalas sea de 0,83.

### ***Análisis de la Robustez del Tipo de Consistencia***

Durante el desarrollo de la metodología que se discutió en la sección 7.2.1 se descartó las listas de APVs provenientes de las mejores soluciones cuyos modelos tienen un bajo Coeficiente de Pearson. Para evaluar el nivel de robustez del análisis realizado se construyó los gráficos de composición de APVs (ver Figura 36) de las listas empleadas (ver Figura 28) con los gráficos de composición obtenidos sin filtrar por Pearson, donde se aprecia que se mantiene la Consistencia Tipo I en los resultados obtenidos para la cromatografía de interacción hidrofóbica (gráficos 1 y 3) y la Consistencia Tipo III para los sistemas de dos fases acuosas (gráficos 2 y 4).

## Composición APVs – Algoritmos de Redes Neuronales

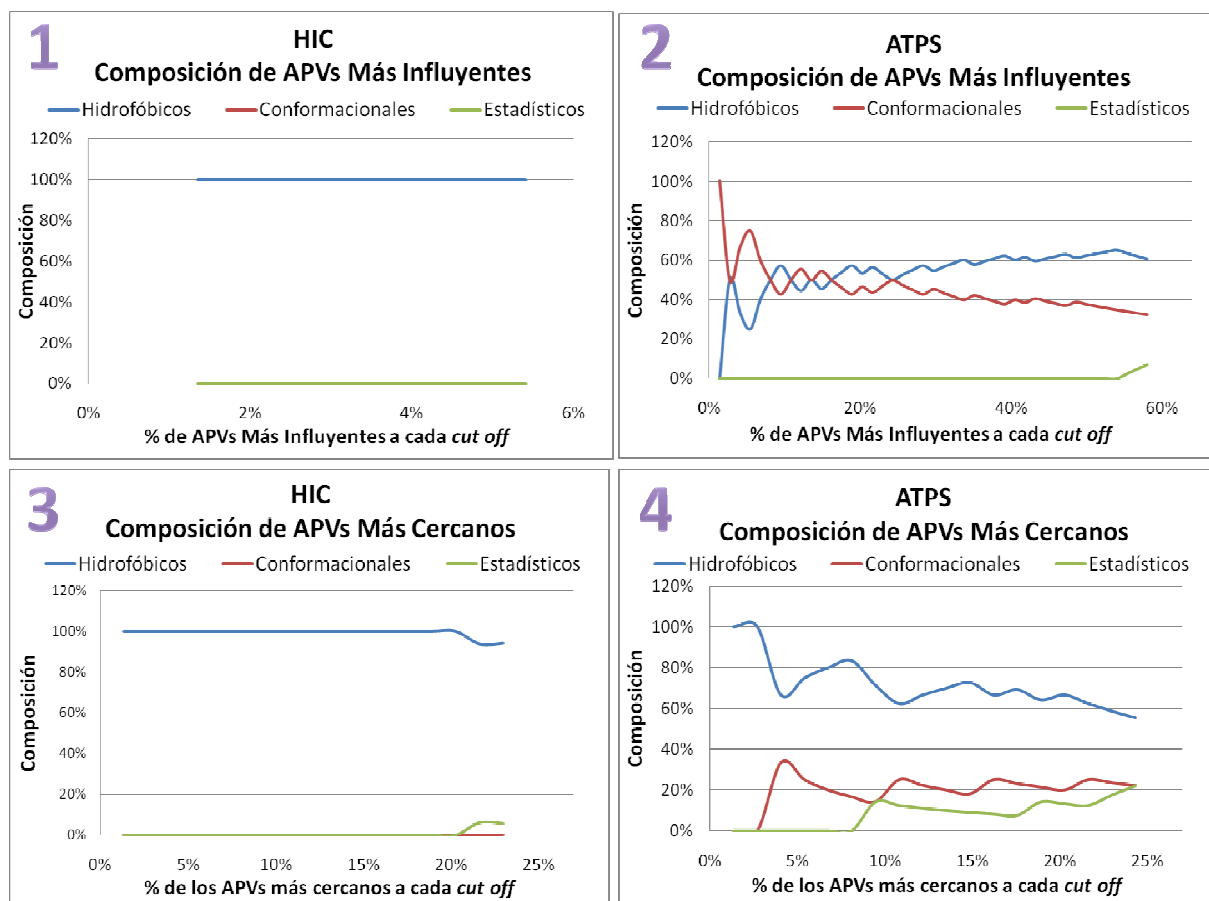


Figura 36: Composición de APVs más influyentes y cercanos de las mejores soluciones encontradas con algoritmos de análisis topológico (redes neuronales). En azul se representa el porcentaje de APVs hidrofóbicos, en rojo los de tipo conformacional y en verde los del tipo estadístico.

### Relación Topológica Entre Escalas Generadas y Originales

En la sección 7.2.1 se comprobó que los resultados obtenidos con el *Genetic Algorithm* no poseen consistencia entre los gráficos de composición por influencia y cercanía. Esto es evidencia de que el *Genetic Algorithm* no utiliza la información topológica de los APVs para generar nuevas escalas, sin embargo, la ubicación espacial de las mejores escalas generadas mediante este algoritmo es muy similar al de las escalas generadas por algoritmos de redes neuronales.

En la sección de Escalas Asociadas a Mejores Modelos Predictivos Para HIC y ATPS (7.1), se observa que a través del *Genetic Algorithm* se creó escalas que permiten obtener modelos con alto poder predictivo. Durante el desarrollo de la metodología se utilizó el *Genetic Algorithm* para evaluar una cantidad de escalas comparable a los algoritmos de redes neuronales, utilizando una cantidad de tiempo significativamente menor.

Por lo tanto, el algoritmo de optimización *Genetic Algorithm* (al menos bajo las restricciones empleadas en este estudio) es una herramienta válida para generar nuevas escalas de hidrofobicidad, el cual es de gran utilidad en la generación de

nuevos modelos cuando no se cuentan con APVs obtenidos con las mismas condiciones del sistema en estudio.

En caso que el estudio que se realiza requiera seleccionar los APVs con información topológica más adecuada para la generación de una escala óptima, por ejemplo para un análisis teórico posterior como el que se realiza en la sección 7.3, se requiere utilizar el enfoque de los algoritmos de redes neuronales, ya que a través de los gráficos de composición por influencia y cercanía se logró demostrar que este tipo de algoritmos presentan consistencia entre la topología de los APVs utilizados y los más cercanos a la escala generada, lo que evidencia el uso de información topológica de los APVs para la generación de nuevas escalas.

La lógica inherente de los APVs en la Figura 34, y la coherencia obtenida entre los gráficos de composición por influencia y cercanía (con los algoritmos de análisis topológico), evidencian una fuerte correlación entre los APVs seleccionados y las mejores escalas obtenidas. En atención a lo anterior, se anticipa que los estudios en los cuales se generó los APVs seleccionados aportan información relevante para una discusión del fenómeno de hidrofobicidad en los sistemas estudiados.

### ***7.3 Análisis de la Propiedad de Hidrofobicidad a la Luz de los Resultados Obtenidos***

En la presente sección se realiza un análisis fenomenológico a partir de los trabajos asociados a los APVs más influyentes y cercanos a las mejores escalas, obtenidos en la sección anterior (7.2.1 y 7.2.2).

En una primera etapa se revisa los trabajos, cuyos resúmenes se pueden encontrar en el Anexo 79, y se introduce el concepto de hidrofobicidad. Luego, se analiza la información contenida en estos trabajos, desde la perspectiva de las estructuras de las proteínas, y se analiza la relación con el concepto de hidrofobicidad. Finalmente, se analizan las similitudes y diferencias entre los sistemas experimentales montados para la obtención de los APVs y los datos utilizados para generar los modelos en el presente trabajo.

#### **7.3.1 El concepto de hidrofobicidad**

La hidrofobicidad se puede describir como la tendencia del agua a excluir moléculas no polares [190, 191]. Este fenómeno se produce por la red formada por puentes de hidrógeno (de carácter altamente dinámico) entre moléculas de agua líquida. Grupos polares como el -OH en la molécula de metanol pueden formar puentes de hidrógeno, y (por lo tanto no interrumpen la red de puentes de hidrógenos) por lo tanto no contribuyen a aumentar la hidrofobicidad de la molécula. Sin embargo, una molécula que consta sólo de una estructura de hidrocarburo, por ejemplo hexano, no puede participar en puentes de hidrógeno con moléculas de agua, lo cual perturba la red. Por ende, al introducir hexano en agua aumenta la entropía, y por lo tanto, el sistema adquiere una configuración en que las moléculas reducen su área de interfaz con el agua. Esta es la configuración más probable y estable, ya que se minimiza la energía libre del sistema. Según la definición termodinámica de hidrofobicidad, cuyo concepto

se acaba de explicar, ésta no es propiedad de una única molécula, sino el resultado de la interacción de muchas moléculas dentro de un sistema acuoso.

Aplicando la definición de hidrofobicidad descrita a los aminoácidos, se deduce que éstos no poseen un valor universal de hidrofobicidad, razón por la cual existe más de una escala asociada a los mismos. Estas escalas dependen tanto de la técnica de medición como del sistema en que se encuentran. Además, la variación del comportamiento de cada aminoácido es distinta al variar el sistema. Por ejemplo, Nozaki y Tandford [192] obtuvieron como resultado que el triptófano y la tirosina tienen un comportamiento altamente hidrofóbicos, mientras que Wolfenden *et. al.* [152] encontraron que dicho comportamiento es muy hidrofílico, respecto a los otros aminoácidos. En ambos casos se trata de aminoácidos con un grupo aromático en su estructura, lo que contribuye al comportamiento hidrofóbico, pero el triptófano posee un átomo de nitrógeno formando tres enlaces (uno a un átomo de hidrógeno), y por su parte, la tirosina contiene un átomo de oxígeno formando dos enlaces (unido con un átomo de hidrógeno), por lo que ambos se pueden comportar como aminoácidos polares de cadena larga, dependiendo por ejemplo del pH o de la concentración de sal del sistema, pudiendo formar puentes de hidrógeno con otras moléculas como el agua.

Dado que los sistemas HIC y ATPS se han descrito como fenómenos físico-químicos fuertemente relacionados con la hidrofobicidad, y que los modelos fueron diseñados y utilizados inicialmente considerando el uso de escalas hidrofóbicas, una de las hipótesis iniciales en esta tesis es que las escalas asociadas a la medición de la hidrofobicidad deberían ser las más influyentes en el diseño de nuevas escalas con los algoritmos de análisis topológico; sin embargo, en la Tabla 48 del Anexo Q se aprecia que un 35% de los APVs provienen del estudio conformacional de las proteínas.<sup>2</sup>

Los trabajos asociados a los APVs escogidos pueden ser clasificados en dos categorías: los que utilizan sistemas físico-químicos para diseñar experimentos con el objetivo de aislar y cuantificar el fenómeno de hidrofobicidad; y los que estudian la hidrofobicidad a través del rol, comportamiento, o características de los aminoácidos dentro de proteínas, o alguna otra propiedad medible que se relacione con la estructura proteica. Estos dos tipos de estudios pueden diferir mucho en las metodologías que utilizan, sin embargo, ambos aportan aspectos importantes sobre qué es y cómo se mide la hidrofobicidad [130, 143-151].

A continuación se realiza una discusión sobre el fenómeno de hidrofobicidad a partir de los trabajos asociados a los APVs escogidos.

### **7.3.2 Aproximación a la Hidrofobicidad a Partir de la Estructura de las Proteínas**

Los trabajos asociados a los APVs que se escogió y que utilizan la estructura de las proteínas son los n°: 8 [147], 10 [150], 11 [151], 12 [148], 22 [146], 60 y 61 [145] (ver trabajos asociados a los índices señalados en la Tabla 48, en Anexo Q y Anexo R).

---

<sup>2</sup> Para entender porqué esta hipótesis no se cumple, es necesario comprender los estudios asociados a los APVs escogidos (ver anexo J).

A partir de estos trabajos, una primera conclusión es que existe un principio que gobierna el plegamiento de las proteínas desde su estructura nativa hasta la terciaria [145-152, 193], sin embargo, éste no se puede explicar únicamente a través de la interacción entre residuos hidrofóbicos, sino que también depende de interacciones entre residuos polares e interacciones hidrofóbicas entre puentes de hidrógeno de dos aminoácidos [145-147]. Las estructuras secundarias, que se produce durante la primera etapa del plegamiento, contienen un mayor número de aminoácidos polares [145, 146] (teoría del glóbulo fundido, una de diversas teorías), encontrándose además que la proporción de grupos polares que forman enlaces intramolecular mediante puentes de hidrógeno dentro de una proteína es independiente del peso molecular de éstas (después de un valor límite  $PM^*$ ), y que es prácticamente constante e igual a ~50% [146] (estudio de 1976). Lo anterior sugiere que existe una estructura básica de proteínas polares, aparentemente de cadena larga [150]. Por otra parte, se ha encontrado evidencia que indica que en la vecindad de los aminoácidos, en una esfera de radio 3 Å al interior de proteínas, la mayor parte de ellos son hidrofílicos [145]. Por lo tanto, durante la primera etapa del plegamiento existe una estructura que se origina principalmente en base a aminoácidos identificados como polares.

A pesar de lo anterior, es relevante considerar que a partir de cada puente de hidrógeno también se produce una estabilización hidrofóbica producto del desplazamiento de moléculas de agua, la que se ha cuantificado en 25 [cal/mol] [194, 195], lo que contrasta con los 0,44 a 1,0 [Kcal/mol] del puente de hidrógeno (valor medido en una configuración del tipo Asp-Asp [196]). Esto evidencia un comportamiento dual (hidrofóbicos o hidrofílicos) de algunos aminoácidos [141].

Por otro lado, existe evidencia que indica que el mecanismo a partir del cual se forma la estructura terciaria considera principalmente aminoácidos del tipo hidrofóbicos, por ejemplo, el área superficial expuesta de los residuos hidrofóbicos disminuye mayormente durante el plegamiento que se produce entre la estructura secundaria y terciaria, en relación a los residuos polares [146].

Anteriormente se expuso evidencia que indica que los aminoácidos cumplen distintas funciones durante el plegamiento de las proteínas. Los aminoácidos hidrofílicos son más relevantes en interacciones tempranas [145, 146], están asociados a una alta energía de estabilización [196], y por lo tanto son los principales responsables del desplazamiento de las primeras moléculas de agua. Los aminoácidos hidrofóbicos participan en reordenamientos que requieren una menor energía de estabilización [194, 195]. Debido a que los aminoácidos mantienen la capacidad de formar el mismo tipo de enlaces, tanto al interior de una proteína como en su superficie, es razonable considerar que el estudio de las estructuras que forman los aminoácidos en su interior puede aportar información sobre las estructuras que forman en su superficie.

En conclusión, los APVs más relevantes para la generación de los modelos más exitosos contienen información importante sobre la estructura de las proteínas y la función de los aminoácidos, que como se ve a continuación se puede correlacionar con el fenómeno de hidrofobicidad.

Los APVs n° 7, 10 y 12, que corresponden a estudios de predicción de la estructura de las proteínas [148-151], se basan en el cálculo de la frecuencia (considerada como un



potencial) con que cada aminoácido se encuentra en una estructura secundaria específica. Como se explicó en su oportunidad, los aminoácidos polares poseen mayor frecuencia en estructuras secundarias respecto a los aminoácidos hidrofóbicos [145-147] (sin ser esta última despreciable). Por otro lado, propiedades moleculares de los aminoácidos como si son voluminosos, si poseen cadenas cortas polares, o si poseen grupos hidroxilos, entre otros, influyen en la formación de estructuras secundarias [151]. Por lo tanto, no es extraño que haya una relación entre el potencial de asociación a alguna de estas estructuras secundarias y su comportamiento hidrofóbico al interior de proteínas. De hecho Chothia indica que los aminoácidos polares que forman puentes de hidrógenos intermoleculares deben ser considerados como hidrofóbicos [146].

En el trabajo asociado a los APVs n° 60 y 61, se estudia la vecindad de los aminoácidos en proteínas y se obtiene varias escalas de hidrofobicidad a través del siguiente estadístico:  $n^\circ \text{ a. a. hidrofóbicos} / n^\circ \text{ a. a. hidrofílicos}$ , el que considera los aminoácidos dentro de un radio  $r_i$  para cada tipo de aminoácido. La metodología empleada inicialmente se basa en el supuesto de que los aminoácidos hidrofóbicos y polares están mayormente rodeados por aminoácidos del mismo tipo, sin embargo, ésta es iterativa y converge sin importar la elección de los dos aminoácidos iniciales (uno es el más hidrofílico y el otro el más hidrofóbico), lo que elimina el sesgo del supuesto inicial. Por lo tanto, existe una relación directa entre la relación de la hidrofobicidad y estos APVs.

En el trabajo asociado al APV n° 8, se estudia la pérdida del área superficial (global y por aminoácidos) durante el plegamiento de las proteínas. A partir del gráfico del área del estado estándar v/s la pérdida de área superficial durante el plegamiento se aprecia tres tendencias claras, a partir de lo cual es posible clasificar los aminoácidos de la siguiente manera:

- Aminoácidos hidrofóbicos: (Gly), Ala, Cys, Val, Ile, Leu, Met, Phe y Trp.
- Aminoácidos medianamente polares: (Gly), Ser, Thr, His y Tyr.
- Aminoácidos polares: (Gly), Pro, Asp, Asn, Glu, Gln, Lys, y Arg.

Esta clasificación es coherente con la estructura molecular de los distintos aminoácidos, ya que casi todos los aminoácidos que son catalogados como hidrofóbicos poseen una cadena de carbono sin grupos funcionales (Ala, Val, Ile, Leu, Met, Phe). Las excepciones son Cys, que forma puentes disulfuro, por lo que no interactúa como una molécula polar normal, y Try, el que tiene una cadena de tres carbonos acoplada a un anillo tipo fenol y una molécula de nitrógeno, por lo que también es un caso atípico. Los medianamente polares son todos aminoácidos con grupos alcohol (-OH) o amoníaco en un anillo (2NH-), lo cual también constituye un caso atípico. Los aminoácidos polares corresponden a ácidos (-COOH) o contienen moléculas de nitrógeno sin impedimento estérico (-NH<sub>2</sub> o -NH<sub>3</sub>). Por lo tanto, esta correcta clasificación de los aminoácidos evidencia la relación entre la pérdida del área superficial y la hidrofobicidad.

Como resumen de esta sección se puede decir que:

- a. El conjunto de los estudios asociados a los APVs más relevantes para la generación de los modelos más exitosos contienen información importante sobre la estructura de las proteínas y la función de los aminoácidos. Además, este conocimiento se origina como resultado del uso de distintas metodologías de estudio.

- b. Los estudios asociados a los APVs más relevantes para la generación de los modelos (encontrados por los algoritmos topológicos) están relacionados con el fenómeno de hidrofobicidad.
- c. Los aminoácidos y moléculas en general no se deben clasificar únicamente como hidrofóbicos o hidrofílicos. Los fenómenos asociados a la hidrofobicidad podrían verse mejor representados mediante una escala de hidrofobicidad y adicionalmente una que denote el carácter polar o de hidrofiliidad.

### **7.3.3 Aproximación a la Hidrofobicidad a Partir de Procesos de Purificación de Proteínas**

En los trabajos asociados a los APVs n° 25, 32 y 42 [130, 143, 144] (ver Anexo R), al igual que los estudios base de los tipos de modelos utilizados [4, 9] no se considera cómo el sistema puede interactuar con zonas específicas de las proteínas, ya que utilizan la hidrofobicidad media u otros estadísticos globales de la hidrofobicidad de una molécula. Sin embargo, esto puede ser útil para aislar la contribución de un sistema específico a las interacciones hidrofóbicas.

Acorde a lo señalado anteriormente, existe evidencia que muestra que los modelos obtenidos a partir de APVs asociados a procesos físico-químicos bajo condiciones específicas, pueden perder poder predictivo al cambiar las condiciones, tal como le ocurrió a Meek [130] al utilizar dos pH diferentes (2,1 y 7,4) o a Browne *et. al.* [156] al utilizar los reactivos ácido trifluoroacético y ácido heptafluorobutírico.

Dado que la contribución de un sistema específico a la hidrofobicidad involucra mecanismos complejos, es fundamental el estudio global de la hidrofobicidad en estos sistemas bajo las condiciones específicas en que se requiere obtener buenas predicciones.

A continuación se analiza los trabajos asociados a las escalas n° 25, 32 y 42 [22, 24, 54]; con el objetivo de evaluar las similitudes y diferencias de los montajes experimentales respecto a los utilizados para generar la información base a partir de la cual se ajustan los modelos utilizados en esta tesis. La discusión se realiza en forma separada para cada uno de los sistemas estudiados.

#### ***Comparación Sistemas de Cromatografía de Interacción Hidrofóbica***

Una diferencia importante entre las dos escalas hidrofóbicas seleccionadas por los algoritmos de análisis topológico es que utilizaron estructuras proteicas muy diferentes. Mientras Parker *et. al.* utilizó estructuras específicas para aislar la contribución de cada aminoácido [144], Meek utilizó estructuras más complejas que se encuentran en los sistemas vivos, como: triglicina, dimetionina, oxitocina, neurotensina,  $\alpha$ -Melanotropina, entre otros [130]. Según el análisis de la sección 7.2.1, el trabajo de Parker *et. al.* es más influyente para los mejores APVs, aún cuando se utilizó el tiempo de retención de proteínas globulares. En conclusión, el trabajo experimental de Parker *et. al.* permite aislar la contribución de cada aminoácido a la hidrofobicidad de las proteínas de forma más precisa, en relación a la metodología empleada por Meek, quien utilizó una técnica indirecta.

Es posible encontrar más diferencias entre los estudios de Parker *et. al.* y Meek en los sistemas cromatográficos utilizados, evaluando las similitudes con el sistema que se utilizó para obtener los datos base del ajuste de los modelos empleados en esta tesis (Lienqueo *et. al.* [54]). A continuación se presenta en la Tabla 35 las distintas características de los protocolos utilizados por Parker *et. al.* [144], Meek [130] y Lienqueo *et. al.* [54].

**Tabla 35: Características de las cromatografías realizadas por Lienqueo *et. al.* [54], Parker *et. al.* [144] y Meek [130].**

Características	Lienqueo	Parker	Meek
Sistema Cromatográfico	Sistema cromatográfico de flujo rápido y alta resolución (FPLC, Pharmacia, Uppsala, Sweden) equipado con loop de inyección de 500- $\mu$ l	Sistema cromatográfico de alta resolución (Spectra-Physics SP8700 + SP8750 módulo combinado con un Hewlett- Packard HP 1040A) equipado con loop de inyección de 500- $\mu$ l	Sistema cromatográfico de alta resolución (HPLC) con gradiente de acetonitrilo obtenido de Fisher
Tipo de Cromatografía	HIC	Fase Reversa	Fase Reversa
Columna	1 ml fenil-sefarosa de flujo rápido. Tamaño de partícula de 30 $\mu$ m, y concentración de ligando de 25 $\mu$ mol/ml de medio	SynChropak RP-18 (C-18) (250 x 4.1 m i.d.). Tamaño de partícula de 6.5 $\mu$ m y tamaño de poro de 300 Å de carbón al 7,5%	-
Tasa de Flujo	0.75 ml/min	1 ml/min	1 ml
Volumen	10 columnas	-	-
Eluyente	Gradiente decreciente de sulfato de amonio A: 20 mM Bis-Tris pH 7.0 más 2 M sulfato de amonio B: 20 mM Bis-Tris pH 7.0	A: 10 mM (NH <sub>4</sub> ) <sub>2</sub> HPO <sub>4</sub> /0.1 M NaClO <sub>4</sub> acuoso B: 0.1 M NaClO <sub>4</sub> in 40% H <sub>2</sub> O y 60% acetonitrilo	A: 0,1 M NaClO <sub>4</sub> /0% acetonitrilo + buffer 5 mM de fosfato a pH 7.4 B: 0,1 M NaClO <sub>4</sub> /60% acetonitrilo (v/v)
Gradiente	El gradiente utilizado fue 7.5% B/min	-	0.5 ml B/min
Tipo Proteínas	Monomérica	Péptidos	Péptidos Varios
pH	7.0 [H <sub>3</sub> O <sup>+</sup> ]	7.0 [H <sub>3</sub> O <sup>+</sup> ]	7.4 [H <sub>3</sub> O <sup>+</sup> ]

En la Tabla 35 se puede apreciar que la cromatografía utilizada por Lienqueo *et. al.* y las utilizadas por Parker *et. al.* y Meek difieren en el tipo, mientras la primera es HIC, las otras dos son de fase reversa. Esto implica que el medio de los sistemas de Parker *et. al.* y Meek genera interacciones hidrofóbica de una magnitud considerablemente mayor a la del medio utilizado por Lienqueo *et. al.*

A continuación se comparan los sistemas de Lienqueo *et. al.* y Parker *et. al.*, debido a que este último corresponde al estudio más influyente.

A pesar de la cantidad de propiedades de un sistema HIC, como se puede ver en el resumen de la Tabla 35, algunos estudios [197] indican que las propiedades más relevantes para la retención de proteínas son la concentración y tipo de sal [35]; y la densidad y tipo de ligando hidrofóbico [198].

El ligando hidrofóbico utilizado en el estudio de Lienqueo *et. al.* [54] es del tipo fenólico (ver Figura 37 parte A), y la estructura del ligando utilizado por Parker *et. al.* es una cadena carbonada de 18 carbonos (ver Figura 37 parte B). Lienqueo *et. al.* utilizan ligandos fijos a una matriz de *spheroidal silica*, con concentración de ligando de 25

$\mu\text{mol/ml}$  de medio. Por otro lado, para el caso de Parker *et. al.* se desconoce el soporte de los ligandos, aunque estos son compatibles con la misma matriz [193].

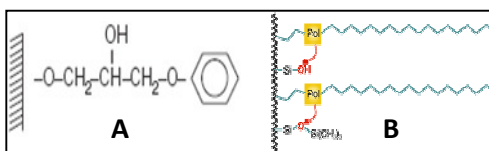


Figura 37: Ejemplos de distintos tipos de ligandos utilizados en una cromatografía de interacción hidrofóbica [29].

Otro factor relevante es la concentración y tipo de sal. En el caso de la cromatografía utilizada por Lienqueo *et. al.*, se trata de una cromatografía de interacción hidrofóbica, en la cual la elución se produce por el tipo y concentración de sal. En el caso de la cromatografía utilizada por Parker *et. al.* (y la de Meek) es de fase reversa, por lo que el principal compuesto que genera la elución no es la sal, es el acetonitrilo. A diferencia de la sal, que permite partir con un medio altamente hidrofóbico y luego se disminuye la hidrofobicidad para la elución, el acetonitrilo es una molécula altamente hidrofóbica que compite con los péptidos desplazándolos de la matriz, y haciendo que éstos eluyan.

Según los antecedentes mostrados, no se posee evidencia de similitud en ninguno de los principales factores que afectan la retención en HIC, y la magnitud de las interacciones hidrofóbicas son diferentes. Por lo tanto, la elección del APV de Parker *et. al.* no se puede atribuir a la similitud de las condiciones cromatográficas.

Por otro lado, la elección del APV desarrollado por Parker *et. al.* como la escala hidrofóbica más influyente asociada a HIC, puede deberse a los méritos relativos de las metodologías para aislar la hidrofobicidad de cada aminoácido, minimizando posibles interferencias. Las evidencias encontradas, y que diferencian al estudio de Parker *et. al.* sobre el de Meek, u otros, son las siguientes:

- El estudio de Parker *et. al.* considera el efecto de los bordes mediante el uso de una estructura básica similar de estructura lineal para el estudio de cada aminoácido (en vez de utilizar aminoácidos libres, péptidos de cadena larga o proteínas).
- El estudio de Parker *et. al.* utiliza pH 7 con el objetivo de evitar la formación de grupos funcionales cargados.

Por otro lado, bajo condiciones hidrofóbicas tan fuertes es posible que en los péptidos se produzcan interacciones entre sus distintas zonas, o entre distintos péptidos. En este sentido Parker *et. al.* utilizan un péptido la siguiente forma: Ac-Gly-X-X-(Leu)<sub>3</sub>-Lys<sub>2</sub>-Amine. A continuación se resumen los componentes de este péptido en la siguiente tabla:

Tabla 36: Detalle componentes de la estructura del péptido de Parker [144] (Ac-Gly-X-X-(Leu)<sub>3</sub>-Lys<sub>2</sub>-Amine).

Abreviatura	Nombre	Residuo
Ac	Grupo Carboxilo	-COOH
Gly	Glicina	-H
Leu	Leucina	-CH <sub>2</sub> CH <sub>2</sub> (CH <sub>3</sub> ) <sub>2</sub>
Lys	Lisina	-(CH <sub>2</sub> ) <sub>3</sub> NH <sub>3</sub> <sup>+</sup>
Amine	Amino	-NH <sub>3</sub> <sup>+</sup>

Como se puede apreciar en la Tabla 36, el péptido de Parker *et. al.* sólo contiene un aminoácido hidrofóbico, Leucina, el cual está posicionado junto a los dos aminoácidos cuya hidrofobicidad se desea caracterizar. Los aminoácidos hidrofóbicos no pueden interactuar plenamente al estar posicionados uno junto al otro (siempre quedará la mayor parte de la superficie expuesta), por lo que se descarta interacción significativa entre los aminoácidos de un mismo péptido. Las interacciones entre péptidos también están controladas en cierta medida, al haber un esqueleto de aminoácidos con grupos polares; como el ácido carboxílico, el grupo amina al final y el aminoácido lisina. Dado que la matriz y los ligandos no tienen grupos polares, las interacciones hidrofóbicas predominante se produce entre el péptido y la matriz. Finalmente, el aminoácido hidrofóbico asegura un mínimo de interacción hidrofóbica, lo que permite caracterizar también a los aminoácidos polares.

Es claro que la construcción peptídica de Parker *et. al.* posee muchas ventajas eliminando fuentes de variabilidad desde la experimentación, lo que no se puede asegurar para el caso de Meek. Este último, además utiliza un método indirecto para determinar la hidrofobicidad de cada aminoácido, que consiste en una optimización del Coeficiente de Correlación entre los valores empíricos y teóricos de hidrofobicidad obtenidos para los 25 péptidos, variando los coeficientes de hidrofobicidad de cada aminoácido utilizando un análisis de sensibilidad como estrategia para examinar el dominio.

### ***Comparación de los Sistemas de Dos Fases Acuosa***

El estudio de Fauchère *et. al.* [143] corresponde al único APV construido en base a un sistema de dos fases acuosas (octanol/agua) dentro de los APVs que se seleccionó en la sección 7.2.1. Al igual que en el caso del APV de Parker *et. al.* [144] para HIC, Fauchère *et. al.* utilizan péptidos de estructura corta y estándar para todos los aminoácidos, con el objetivo de simular el enlace peptídico y evitar cargas en la cadena principal. Además, se utiliza la estructura N $\alpha$ -acetil-amino-acido amida para estudiar el efecto de los extremos carboxilos y aminos.

Es posible obtener más antecedentes de los méritos del APV de Fauchère *et. al.* a partir de un análisis comparativo del protocolo utilizado por éste, respecto al que se utilizó para obtener los datos base del ajuste de los modelos empleados en esta tesis (Andrews *et. al.* [22]). A continuación, en la Tabla 37, se muestran los principales antecedentes de los sistemas utilizados por Fauchère *et. al.* y Andrews *et. al.*

**Tabla 37: Características de los ATPS realizadas por Andrews *et. al.* [22] y Fauchère *et. al.* [143]. La sigla PEG corresponde a Polietilenglicol.**

Característica o Propiedad	Estudio de Andrews	Estudio de Fauchère
Parámetro	Coeficiente de Partición	Coeficiente $\pi$ (en base a coeficiente de distribución, y Coeficiente de Partición para Prolina y Cisteína)
Sistema utilizado	PEG 4000/ phosphate PEG 4000/ sulfate PEG 4000/ citrate PEG 4000/ dextrano	-
Temperatura	Ambiente	Ambiente
Buffer	-	0,1N HCl o 0,1 N NaOH
Sal utilizada	NaCl	-
Concentración de Sal	(0.0%, w/w), (0.6%, w/w) y (8.8%, w/w)	-

Como se puede ver en la Tabla 37, una de las principales diferencias es el parámetro utilizado para medir la calidad de la separación. El estudio de Fauchère *et. al.* usa el coeficiente de Hansch ( $\pi$ ) [46], parámetro que permite aislar la contribución hidrofóbica de las cadenas laterales (-R) de los aminoácidos, lo que puede generar diferencias para los aminoácidos con átomos ionizados. Según se puede ver en la misma tabla, la información disponible de los sistemas utilizados por Andrews *et. al.* [22] y Fauchère *et. al.* [143], no permite una comparación del resto de los parámetros.

#### **7.4 Representación de la Hidrofobicidad Mediante Los Modelos Desarrollados**

Los modelos predictivos que se utilizó en esta tesis se basan en la hidrofobicidad superficial media de las proteínas (ver sección 6.1), lo cual produce un sesgo que puede llegar a ser significativo, ya que estos modelos predicen un mismo tiempo de retención y un mismo coeficiente de partición para dos proteínas, sin considerar si las zonas hidrofóbicas se encuentran localizadas en, por ejemplo, uno o cuatro *clusters*.

Los modelos utilizados para predecir el comportamiento de las proteínas en los sistemas físico-químicos, HIC y ATPs, poseen entre dos y tres grados de libertad. Por otro lado, a partir de la Tabla 35 y la Tabla 37 se puede ver que el número de parámetros más relevantes que influyen en la interacción hidrofóbica entre el sistema utilizado y una proteína oscila entre 9 y 14. Durante el ajuste de los modelos las propiedades de las proteínas varían y las del sistema permanecen constantes, por lo tanto, los coeficientes de los modelos utilizados como: la hidrofobicidad intrínseca y la resolución de ATPs, y los coeficientes del modelo cuadrático de HIC (sección 6.1); solo poseen la capacidad de representar en términos globales e inespecíficos las propiedades del sistema. Es importante aclarar que los modelos lineales tienen 21 grados de libertad, pero éstos se utilizó para aproximar la hidrofobicidad superficial media (o en términos generales la ASP).

Finalmente, al utilizar algoritmos de *clustering* o de optimización para la generación de nuevas escalas, éstas son consideradas como variables (aunque con restricciones que

derivan de la topología de los APVs, en el caso de los algoritmos de *clustering*). Cuando las escalas de propiedades aminoacídicas se utilizan como variables, aumentan los grados de libertad de la metodología global utilizada, lo que posibilita incorporar información derivada de la estructura de las proteínas, según se ve reflejado en los resultados discutidos en la sección 7.3.2. Esto probablemente contribuye a un ajuste de carácter específico a cada proteína (cómo ésta interactúa con el sistema puntual).

## 7.5 *Análisis de las Características de Sobre Ajuste de los Modelos*

En la sección de resultados se obtuvo nuevos modelos para predecir el comportamiento de proteínas en HIC y en 12 ATPs. Estos modelos presentan mejoras significativas desde un punto de vista de su poder predictivo, estimado a través del Error de Jacknife, en relación a los modelos del mismo tipo reportados anteriormente en la literatura [4, 6, 9, 13, 23, 54, 62]. El error asociado a la predicción se logró disminuir hasta un 75% para la cromatografía de interacción hidrofóbica y hasta un 99,6% para los sistemas de dos fases acuosas.

Con la metodología descrita en la sección 6.1, se construyó nuevos modelos en base a nuevas escalas utilizando herramientas de *clustering*, incluyendo Algoritmos de Redes Neuronales, además de un algoritmo de optimización (ver sección 6.4). El uso de este tipo de herramientas hace complejo determinar el número de parámetros ajustables agregados (o grados de libertad equivalentes agregados), respecto a la metodología que utiliza APVs con base en la literatura.

En la sección anterior se vio que el aumento del número de parámetros ajustables posibilita incorporar información derivada de la estructura de las proteínas, según se ve reflejado en los resultados discutidos en la sección 7.3.2. Por otro lado, en el análisis de la sección 7.2 se aprecia evidencia de transferencia de información topológica desde los APVs a las escalas generadas con los Algoritmos de Análisis Topológico. A pesar de esto, no se ha descartado la existencia de sobre ajuste, ya que persiste la incertidumbre sobre el número de parámetros ajustables agregados.

En esta sección se presenta los antecedentes disponibles, y se realiza un análisis con el objetivo de avanzar hacia la determinación de la existencia de un sobre ajuste, enfocado en el caso de HIC.

### 7.5.1 Modelos utilizados

La ecuación básica que relaciona el DRT con la hidrofobicidad de las proteínas es la ecuación 14, donde  $\Gamma$  es la hidrofobicidad de una proteína, y  $b_i$  son los coeficientes del modelo cuadrático:

$$DRT = b_0 + b_1\Gamma + b_2\Gamma^2 \quad (14)$$

La hidrofobicidad de cada proteína se calculó con el modelo tipo 3D a partir de la ecuación 15:

$$\Gamma = \sum_{i \in A} \hat{r}_i \varphi_i \quad (15)$$

Donde  $A$  es el conjunto de 20 aminoácidos posibles,  $\varphi_i$  es el componente número  $i$  de un vector de propiedad aminoacídica (APV),  $\hat{r}_i$  representa la fracción de la superficie proteica ocupado por el aminoácido de clase  $i$ .

La diferencia entre el modelo 3D y los modelos lineales está en que estos últimos no utilizan la ecuación 15 directamente. En este caso la ecuación 15 se utiliza para el cálculo de la hidrofobicidad de una base de datos de 1982 proteínas. Luego, se ajusta los modelos lineales para predecir la hidrofobicidad a partir de la composición aminoacídica, por lo que con el uso de modelos lineales no se agregó parámetros ajustables adicionales para la predicción del DRT.

### 7.5.2 Consecuencia del uso de la escala de propiedades aminoacídicas como variable

En la presente sección se estudia cómo el cambio metodológico, del uso de APVs con base en la literatura a la generación de nuevos APVs, puede influir en el ajuste de los nuevos modelos a los DRT experimentales de un grupo de 12 proteínas (ver proteínas en sección 6.1.1).

Los DRT son valores experimentales (fijos). Al analizar la ecuación 15 se puede apreciar que la hidrofobicidad relativa y absoluta de cada proteína varía al modificarse la escala de hidrofobicidad (APV). Por lo tanto, al modificar la escala hidrofóbica utilizada cambian los valores de hidrofobicidad utilizados en la ecuación 14, lo que se puede apreciar en el esquema de la Figura 38 a continuación.

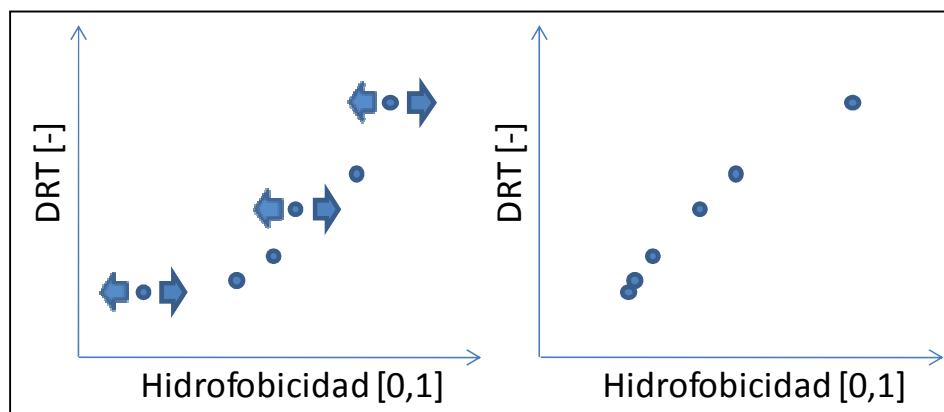


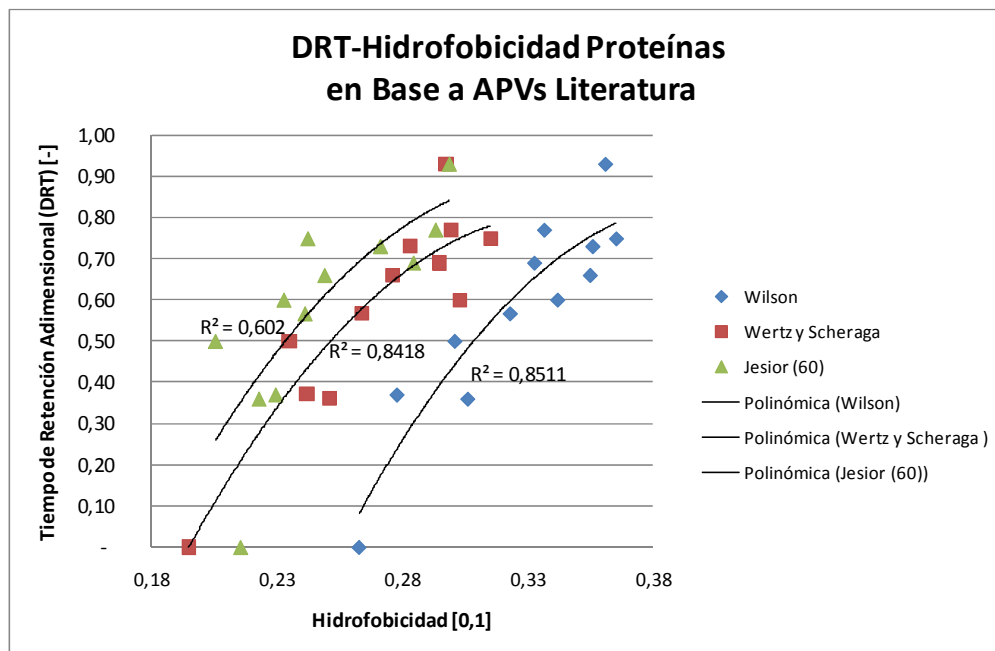
Figura 38: Esquema del efecto de un cambio de escala de hidrofobicidad sobre la ubicación espacial de una proteína en la gráfica del DRT en función de la hidrofobicidad. A la derecha se representa el efecto global si se pudiera modificar la ubicación de cada proteína de forma independiente, es decir, que el cambio de ubicación de una proteína no influya sobre la ubicación del resto de las proteínas en el gráfico.

En la Figura 38 cada punto representa a una proteína. Además, en la figura se ilustra que al modificar la escala de hidrofobicidad se produce un desplazamiento de los puntos sobre los cuales se ajustan los modelos (ver cambios entre los gráficos a la izquierda y a la derecha en la Figura 38).

A continuación, en la Figura 39, se muestra un gráfico que relaciona el DRT y la hidrofobicidad que se obtuvo para las 12 proteínas utilizadas para ajustar los modelos creados (ver sección 6.1.1). La hidrofobicidad se calculó con los APVs de Wertz y Scheraga [180], de Wilson [181] y de Jesior [145]. Los dos primeros APVs se utilizó



como referencia en la sección 7.1.3 para analizar los resultados de los modelos generados para HIC, ya que dentro de los 74 APVs estudiados son los que permiten obtener los mejores modelos [4, 9, 13, 54]. El APV de Jesior [145] es uno de los más influyentes para la generación de las escalas a partir de las cuales se creó los mejores modelos para el DRT en HIC (ver sección 7.2.1).



**Figura 39:** DRT en función de la hidrofobicidad de las 12 proteínas utilizadas para la generación de los nuevos modelos. Se incluye el ajuste de un modelo cuadrático en cada caso. La hidrofobicidad se calcula con las escalas de Wilson [179], Wertz y Scheraga [178], y Jesior [143].

Al analizar la Figura 39, y considerando que una misma proteína mantiene la posición en el eje de la ordenada (tiene un valor fijo de DRT), se puede ver las variaciones relativas de la hidrofobicidad de cada proteína al variar la escala. Además, a partir de los coeficientes del ajuste de funciones del tipo cuadrático se puede ver cómo estos cambios de hidrofobicidad repercuten en el ajuste.

En términos generales, se aprecia un desplazamiento horizontal de todas las proteínas en la misma dirección, lo que implica que en cierto grado todas las proteínas varían su posición de forma similar en la gráfica. Sin embargo, al observar en detalle las repercusiones del cambio de escala se aprecia que algunas de estas proteínas no varían igual, en términos relativos entre ellas. Como en este caso se trata de APVs con base en la literatura, es esperable que al utilizar las nuevas escalas generadas la variación relativa de las proteínas no se mantenga.

A continuación se muestra cómo varía la posición de las proteínas, un gráfico del tipo DRT-Hidrofobicidad, con las escalas generadas a partir de los modelos 3D obtenidos con algoritmos de análisis topológico y el *Genetic Algorithm*, y con el modelo lineal y el *Genetic Algorithm*. La ubicación de las proteínas obtenidas con modelos del tipo lineal y con escalas generadas a partir de algoritmos de análisis topológico se encuentra muy a la izquierda en el gráfico, dificultando la visualización por lo que no se muestra, sin embargo, el gráfico DRT-Hidrofobicidad de dicha escala no aporta información crítica para el análisis.

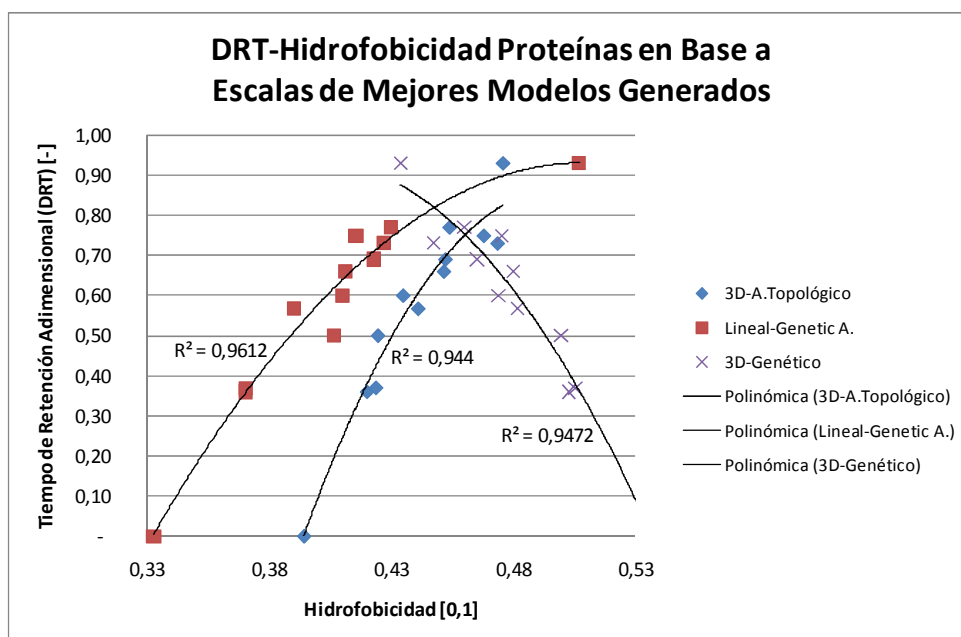


Figura 40: DRT en función de la hidrofobicidad de las 12 proteínas utilizadas para la generación de los nuevos modelos. Se incluye el ajuste de un modelo cuadrático en cada caso. La hidrofobicidad se calcula con las escalas generadas para los modelos 3D con algoritmos de análisis topológico y el *Genetic Algorithm*, y con la escala generada para el modelo lineal con el *Genetic Algorithm*.

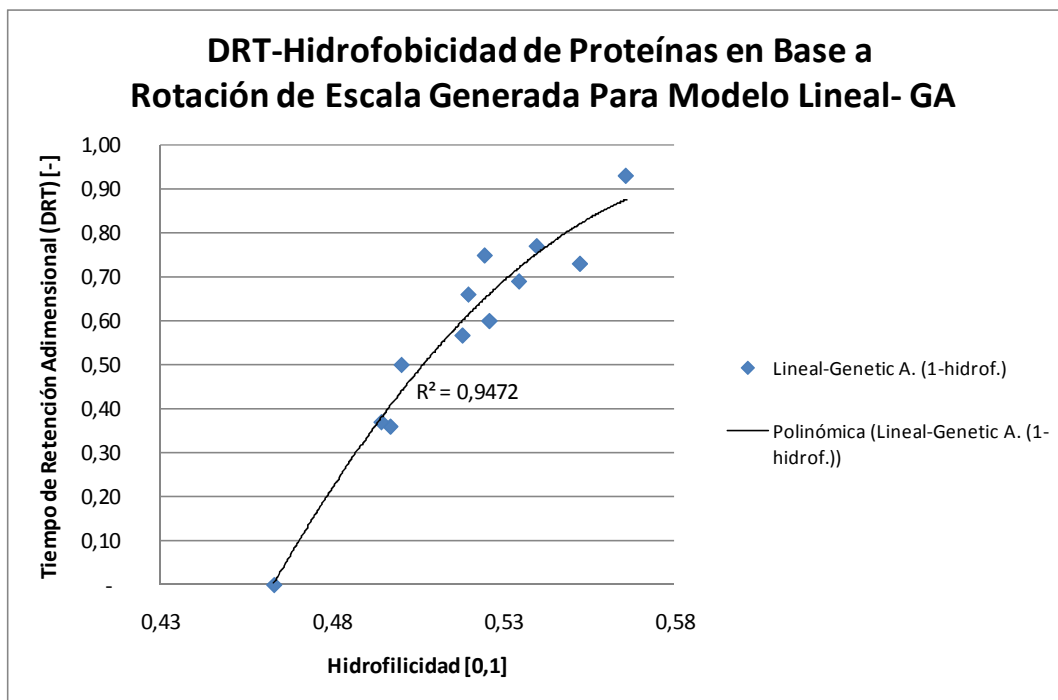
Como se puede ver en la Figura 40, y comparando con el gráfico de la Figura 39, sólo en el caso de la escala obtenida por el *Genetic Algorithm*, y utilizado para generar un modelo del tipo lineal, se puede apreciar claramente una variación de la posición de las proteínas de una forma singular que podría sugerir un sobre ajuste. Como muestra el Coeficiente de Determinación ( $R^2$ ) existe un buen ajuste, sin embargo, la tendencia de los datos se ha modificado completamente respecto al de los APVs de la Figura 39. Desde un punto físico-químico pareciera que se está representando un comportamiento opuesto a la hidrofobicidad (también denominado hidrofiliidad).

Para que esto no sea un sobre ajuste deberían existir escalas que tengan una correlación negativa, por ejemplo, con la de Wertz y Scheraga [180]. Analizando la correlación entre la escala de Wertz y Scheraga [180] con cada uno de los 74 APVs, se encuentran 12 APVs con correlación negativa. Destacan las correlaciones obtenidas entre el APV de Wertz y Scheraga [180] y los APVs de Bhaskaran y Ponnuswamy [162], que corresponde a una índice de flexibilidad en la posición de residuos de aminoácidos en proteínas globulares, y cuyo coeficiente de correlación es de -0,802; de Janin [171], construida en base a la fracción molar de 3220 residuos accesibles, y cuyo coeficiente de correlación de -0,796; y la de Levitt [151], que corresponde a la frecuencia normalizada de participación de aminoácidos en las configuraciones *Beta* y *Turn*, y cuyo coeficiente de correlación es de -0,720.

Por lo tanto, se concluye que existen APVs con base en la literatura con correlación negativa, las que se pueden tratar desde un punto de vista físico-químico como escalas de hidrofiliidad. De esta manera, las escalas de hidrofiliidad generadas se pueden transformar en escalas de hidrofobicidad con la siguiente ecuación:

$$Hidfobicidad = 1 - Hidrofiliidad \quad (40)$$

Para el caso concreto de la escala de hidrofilidad generada con el algoritmo genético y correspondiente a modelos lineales de la Figura 40, luego de transformar los datos se obtuvo el siguiente gráfico DRT-Hidrofobicidad, según se muestra en la Figura 41.



**Figura 41: DRT en función de la hidrofobicidad de las 12 proteínas utilizadas para la generación de los nuevos modelos. La hidrofobicidad de las proteínas se calcula con la escala asociada al modelo lineal que se genera con el Genetic Algorithm (ver Tabla 20), luego de transformar la escala con la ecuación 40.**

Al comparar la Figura 41 con la Figura 40, se aprecia que los modelos cuadráticos ajustados tienen un Coeficiente de Determinación ( $R^2$ ) idéntico, lo que se debe a que la ecuación 40 realiza una rotación de los puntos en  $90^\circ$  respecto al eje x (la rotación se realiza en el plano x-z), de manera que las pendientes de las regresiones lineales son iguales en magnitud y de signo opuesto (cuyo valor es 8,367 para el caso del modelo de la Figura 41), y por lo tanto se mantiene la calidad del ajuste y del nivel predictivo en el nuevo modelo.

En general, a través de los gráficos de DRT-Hidrofobicidad no se logró determinar si existe o podría existir sobre ajuste, debido a que las proteínas varían de ubicación de forma irregular incluso al utilizar distintos APVs con base en la literatura. Por lo tanto, aunque los gráficos DRT-Hidrofobicidad permiten comprender bien cómo actúan los distintos APVs y nuevas escalas sobre la construcción de los nuevos modelos, se requiere un mayor análisis para determinar si ha habido un sobre ajuste.

### 7.5.3 Estudio del Comportamiento Hidrofóbico de las Proteínas

Si fuera posible afectar la posición de cada proteína con DRT conocido (ver sección 6.1.1) de forma independiente, entonces el número de parámetros ajustables (o grados de libertad equivalentes) añadidos en la metodología para la generación de nuevos modelos del DRT sería igual al número de proteínas. Es por esto que en la presente sección se estudian las escalas de propiedades aminoacídicas y sus componentes, la hidrofobicidad de las proteínas, y la relación entre éstas.

Para el estudio de las variables mencionadas se utilizó el área superficial accesible de los aminoácidos en las proteínas estudiadas (ver Anexo S y Z), la ecuación 15 y la escala de Wertz y Scheraga [180] (ver Anexo AA).

Se hizo variar el valor de hidrofobicidad de un aminoácido en la escala de Wertz y Scheraga [180], en un valor fijo e igual a la media de la escala original, con lo que se obtuvo un nuevo valor de hidrofobicidad para cada proteína incluida en el estudio. Este conjunto de valores de hidrofobicidad para cada proteína se puede entender como un perfil de hidrofobicidad. Por lo tanto, al realizar la modificación señalada sobre la escala de Wertz y Scheraga [180] se obtuvo un conjunto de perfiles de hidrofobicidad (ver Anexo T). Sobre estos resultados se realizó un análisis de varianza de un factor, donde se determinó la probabilidad de que los perfiles de hidrofobicidad sean iguales. La probabilidad obtenida es de  $3,23 \cdot 10^{-23}$ . Como la probabilidad es menor a 0,05 se rechaza, con un nivel de confianza del 95%, la hipótesis de que los perfiles de hidrofobicidad sean iguales (ver resultado de la prueba de varianza – Anova en Anexo U).

Lo anterior indica que la forma en que cada aminoácido afecta la hidrofobicidad de cada proteína es diferente. Esto es esperable, debido a que el área superficial accesible (ASA) de cada uno de los 20 aminoácidos es distinta para una proteína, y entre proteínas (ver Anexo S).

Por otro lado, al variar la hidrofobicidad de una proteína a partir de una modificación en el valor de los componentes de un APV, el valor de la hidrofobicidad asociado al resto de las proteínas también varía (ver Anexo V).

Por lo tanto, es relevante estudiar el nivel de dependencia-independencia entre las hidrofobicidades calculadas para el conjunto de 12 proteínas. Para esto se utilizó el mismo análisis sobre el APV de Wertz y Scheraga, en el cual se hace variar el valor de la hidrofobicidad de cada aminoácido, uno a la vez, en una magnitud igual a la media de hidrofobicidad de la escala.

Un primer análisis realizado es un test de varianza de un factor sobre, considerando los perfiles de hidrofobicidad por proteína y no por aminoácido, cuyo valor de hidrofobicidad se ha modificado (ver Anexo V). Como el valor de probabilidad obtenido es de 0,953 no se rechaza la hipótesis de que la hidrofobicidad de las proteínas varía de igual manera, con una confianza del 95% (el valor de rechazo es inferior a 0,05). Por lo tanto, no existe evidencia estadística que indique que las proteínas estudiadas se comportan de manera diferente frente a los distintos cambios realizados en la escala de hidrofobicidad de Wertz y Scheraga.

También se estudia la probabilidad de que las proteínas se comporten de igual forma al ser agrupadas según una clasificación determinada a partir del análisis de sus ASAs (ver Anexo Z), obteniéndose una probabilidad de 0,650. Al igual que en el caso anterior, no existe evidencia estadística que indique que las proteínas estudiadas se comportan de manera diferente frente a los distintos cambios realizados en la escala de hidrofobicidad de Wertz y Scheraga. Es decir, la agrupación propuesta (ver Anexo Z) no es estadísticamente significativa para describir el comportamiento de la varianza entre la hidrofobicidad de las proteínas incluidas en el estudio.

Otra forma de expresar los resultados obtenidos es la siguiente: como la hidrofobicidad de cada proteína  $H_{pi}$  depende de la escala de hidrofobicidad utilizada  $\overline{APV} = (h_1, \dots, h_i, \dots, h_{20})$  (se utiliza sólo una a la vez), entonces existe una función  $g_{ij}$  que relaciona la hidrofobicidad de las proteínas (ver Figura 42). Además, no se logró encontrar evidencia estadística que indique que el comportamiento de la hidrofobicidad de las proteínas, frente a las variaciones hechas a partir de la escala de Wertz y Scheraga [180], en conjunto equivale a más de un parámetro ajustable (equivalente a agregar un grado de libertad en el modelo).

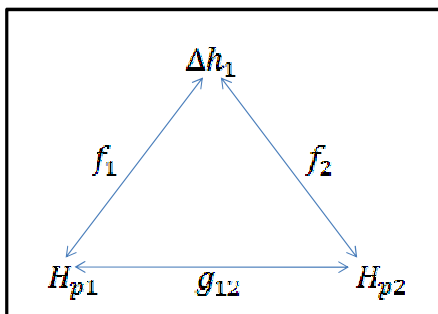


Figura 42: Esquema de relación entre la variación de un componente de una escala de hidrofobicidad, y la hidrofobicidad de dos proteínas.

En resumen, del análisis realizado en la presente sección (7.5) se puede decir que no existe evidencia de un sobre ajuste a los datos realizados por los modelos y la metodología utilizada. Por otra parte, el estudio del comportamiento de la hidrofobicidad sobre las modificaciones hechas a la escalas de Wertz y Scheraga [180], se deduce que no existe evidencia que indique que la metodología utilizada tiene el potencial de agregar más de un parámetro ajustable adicional en comparación con el uso de APVs provenientes de la literatura.<sup>3</sup>

## 7.6 Información Aportada a los Modelos por los APVs Más Relevantes

De los análisis realizados en las secciones anteriores, se desprende que al utilizar algoritmos de redes neuronales es posible crear nuevas escalas que poseen información topológica transmitida desde el conjunto inicial de APVs utilizado. Por otra parte, la evidencia indica que el conjunto de APVs que dan origen a los mejores modelos no es un conjunto azaroso, sino que producto de la etapa de selección de los mejores resultados, el conjunto es bien acotado y contiene básicamente dos tipos de información:

- Por una parte, los APVs seleccionados incluyen estudios experimentales similares a los sistemas en estudio, y coherentemente con lo esperado, sus niveles de influencia y cercanía son los más altos, según se puede ver en el análisis de la sección 7.2.1. Esto es muy relevante, debido a que estos APVs contienen en su topología la información sobre el comportamiento de los 20

<sup>3</sup> Para ver gráficamente la variación de la hidrofobicidad de cada proteína, y para ver los gráficos DRT-Hidrofobicidad de las escalas generadas a partir de las modificaciones realizadas sobre la escala de Wertz y Scheraga 180. Wertz, D.H. y H.A. Scheraga, *Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule.* Macromolecules, 1978. 11(1): p. 9-15. , ir a Anexo Y.

aminoácidos en los sistemas en estudio, como resultado de una interacción inespecífica producto de las propiedades del sistema.

- Por otra parte, los APVs seleccionados incluyen estudios conformacional e hidrofóbicos que contienen información importante sobre la estructura de las proteínas y la función de los aminoácidos, y que está correlacionada con el fenómeno de hidrofobicidad. Esta información es de carácter específico, y por ejemplo permite considerar cambios de comportamiento relativo de algunos aminoácidos al variar las condiciones del sistema en estudio.

Como se discutió en la sección 7.4, el modelo permite utilizar la información topológica contenida en las escalas creadas, información que se complementa con la contenida en las constantes de los modelos, y que permiten a su vez calibrar considerando los cambios existentes entre el sistema experimental utilizado y el sistema experimental del mismo tipo del o los APVs más influyentes, lo cual explica el gran éxito de la metodología empleada para mejorar el poder predictivo de los modelos creados evaluado a través del Error de Jackknife.

## 8 Conclusiones

A partir de la búsqueda de la mayor cantidad de soluciones posibles, considerando las variables tiempo y rigurosidad metodológica, se generó aproximadamente 380.000 nuevas escalas. A partir del análisis de las diferencias en el número de escalas generadas por cada algoritmo, se concluye que éstas se explican principalmente por las características de estos algoritmos, como pueden ser: tipo de algoritmo, número de parámetros ajustables (o número de parámetros que se haga variar), tipo de objetos con que trabaja (vectores o grafos), tipo de propiedad topológica utilizada, existencia de multiplicidad de resultados, interfaz y posibilidad de automatizar cálculos (esta es una característica de la implementación utilizada).

A través de la metodología empleada se obtuvo 52 mejores modelos (considerando los dos tipos de algoritmos, dos tipos de modelos y 13 sistemas), para los cuales se logró un aumento del poder predictivo medido a través del Error de Jackknife, de entre un 11% y un 99,6% en un 80,8% de los casos. De forma simultánea, dentro de los modelos con aumento del poder predictivo se obtuvo mejoras en el nivel de ajuste medido a través del Coeficiente de Pearson, de un 4% a un 300%, en 28 de 42 casos (66,7%); por otra parte, el nivel de ajuste disminuyó en 14 de los 42 casos.

El éxito de la metodología empleada es significativo y mayor al logrado a través de otros estudios que modifican el tipo de modelo incorporando directamente más información. Esto tiene el potencial de ahorrar costos en I+D a la industria biotecnológica.

Dentro de los 42 casos donde se aprecian mejoras del Error de Jackknife, en los sistemas cromatográficos el poder predictivo mejora entre un 37% y 75% con un promedio de 54%. Por otro lado, en los sistemas de dos fases acuosas el rango de mejora es de 11% a 99,6%, con un promedio de 64%.

A través del comportamiento del Error de Jackknife y el Coeficiente de Pearson, se obtuvo evidencia que indica que en los modelos cromatográficos no ha habido un sobre

ajuste de los datos, como tampoco en 34 de los 48 modelos obtenidos para los sistemas de dos fases acuosas.

A partir de los resultados se concluye que los algoritmos de *clustering*, que son del tipo redes neuronales, son los que tienen mayor capacidad para la obtención de mejores nuevas escalas. Una de las principales limitaciones que tienen el resto de los algoritmos de *clustering* es que sólo realizan una agrupación de los APVs, pero no resuelven el problema de generar nuevas escalas, que en esta tesis se realizó mediante el cálculo del centroide de los APVs en cada *cluster*.

A partir del estudio de consistencia entre la ubicación espacial de las mejores escalas generadas y los APVs más influyentes en la generación de éstas, se verificó que no existe transferencia de topología entre los APVs utilizados y las mejores escalas generadas con el *Genetic Algorithm*, pero sí existe transferencia de topología entre los APVs más influyentes y las nuevas escalas generadas con los algoritmos de redes neuronales (casi todos los resultados se obtuvieron con el Algoritmo *Growing Neuronal Gas*), por lo que éstos APVs son de interés y permiten obtener conclusiones adicionales sobre el fenómeno de hidrofobicidad y el uso de información de los modelos utilizados.

A pesar de lo anterior, con el *Genetic Algorithm* se crean escalas que permiten obtener modelos con mayor poder predictivo respecto a otros algoritmos, y la topología de las mejores escalas generadas no es diferenciable a la de las mejores escalas generadas con los algoritmos de redes neuronales (o los APVs de los cuales provienen). Por otro lado, a través del *Genetic Algorithm* se evaluó una cantidad de escalas comparable a los algoritmos de redes neuronales, utilizando una cantidad de tiempo significativamente menor. Por lo tanto, se concluye que el uso del algoritmo de optimización *Genetic Algorithm* (al menos bajo las restricciones empleadas en este estudio) es un enfoque válido para generar nuevas escalas de hidrofobicidad, el cual es de gran utilidad en la generación de nuevos modelos cuando no se cuentan con APVs obtenidos bajo las mismas condiciones del sistema en estudio.

Lo anterior queda respaldado por el análisis de la sección 7.5, del cual se concluye que no existe evidencia de un sobre ajuste a los datos realizados por los modelos, y que la metodología utilizada tenga el potencial de agregar más de un parámetro ajustable adicional, en comparación con el uso de APVs provenientes de la literatura.

A partir de un análisis inverso sobre las mejores escalas obtenidas; basado en las metodologías de los algoritmos de redes neuronales, y en la composición de APVs en el hiper espacio en torno a las mejores escalas; se concluyó que del grupo de los APVs más relevantes (influyentes y cercanos) para la generación de las mejores soluciones: 3 han sido clasificados previamente como escalas hidrofóbicas, obtenidas a través del estudio de sistemas físico químico (HIC y ATPS); 3 como escalas hidrofóbicas, obtenidas a partir de análisis sobre los aminoácidos en proteínas; y 5 como escalas conformacionales, para los cuales se encontró evidencia de su relación con la hidrofobicidad.

A partir de los trabajos en que se obtuvo los APVs más relevante (influyentes y cercanos) para la generación de los modelos más exitosos, se concluye que contienen información que se puede dividir en dos clases:

- Experimentales similares a los sistemas en estudio, cuyos APVs son los de mayor nivel de influencia y cercanía.
- Estudios conformacional e hidrofóbicos que contienen información importante sobre la estructura de las proteínas y la función de los aminoácidos, y que está correlacionada con el fenómeno de hidrofobicidad. Esta información es de carácter específico, y por ejemplo permite considerar cambios de comportamiento relativo de algunos aminoácidos al variar las condiciones del sistema en estudio.

Por otro lado, se concluye que existe un sesgo en los distintos tipos de modelo utilizado debido a que se basan en la hidrofobicidad superficial media. Sin embargo, este sesgo se pudo reducir mediante el uso de algoritmos de *clustering* y optimización, gracias a que el modelo permite utilizar la información topológica contenida en las escalas creadas, información que es complementada con la contenida en las constantes de los modelos, y que permiten a su vez calibrar considerando los cambios existentes entre el sistema experimental utilizado y el sistema experimental del mismo tipo del o los APVs más influyentes. Esto explica el gran éxito de la metodología empleada para mejorar el poder predictivo de los modelos creados evaluado a través del Error de Jackknife.

Otras conclusiones colaterales del estudio, indican que:

- El conjunto de los estudios de la estructura de las proteínas asociados a APVs más relevantes sugieren que los fenómenos asociados a la hidrofobicidad podrían verse mejor representados mediante una escala de hidrofobicidad y adicionalmente una que denote el carácter polar o de hidrofiliidad.
- El hecho de incluir información del tipo conformacional produzca una mejora significativa del poder predictivo de los modelos estudiados, sugiere que la estructura de aminoácidos en la superficie de las proteínas influye en el comportamiento hidrofóbico de las proteínas, por lo que se podrían crear otros métodos indirectos que resulten útiles para mejorar la capacidad predictiva, como por ejemplo, definir la probabilidad de ocurrencia de cada aminoácido en distintos tipos de estructuras que existan en la superficie de las proteínas. También puede ser relevante estudiar la relación entre estas distintas estructuras superficiales y su aporte relativo a la hidrofobicidad, corregido por área superficial hidrofóbica.
- Lo anterior también abre el debate sobre qué relevancia tiene el sesgo mencionado, el cual se puede estudiar al utilizar distintos conjuntos de proteínas con estructura tridimensional, DRT y/o K conocido, dentro de los cuales hayan grupos de proteínas con un mismo tipo de zonas hidrofóbicas en la superficie, y conjunto de proteínas con distintos tipos de zonas hidrofóbicas en la superficie.

Las metodologías que se creó y se utilizó en esta tesis para determinar la información original más relevante utilizada por algoritmos de redes neuronales, en la obtención de



escalas específicas, no está limitada al campo de aplicación en esta tesis, razón por la cual puede ser de utilidad en una amplia gama de aplicaciones. Esta metodología cuenta con la ventaja de haber sido validada en un extenso análisis y bajo distintas perspectivas. Además, el uso del enfoque de los modelos utilizados en esta tesis ha demostrado un gran potencial predictivo por su capacidad de integrar distinto tipo de información, incluyendo información generada a partir de sistemas complejos donde no se logra entender por completo el comportamiento conjunto de las partes existentes, y en que existen distintas metodologías de medición de una propiedad o distintas propiedades involucradas en el comportamiento global que se desea predecir. Por lo tanto, es probable que este enfoque tenga múltiples aplicaciones para el estudio de sistemas complejos.

Finalmente, se concluye que se ha cumplido el objetivo de esta tesis, al generar nuevos APVs que permiten mejorar modelos del tiempo de retención adimensional (DRT) para HIC y del coeficiente de partición (K) para doce sistemas de dos fases acuosas, utilizando herramientas de *clustering*. Además, a partir de la evidencia de transferencia de información topológica detectada, y de cómo se utilizó por los modelos generados, se comprueba la hipótesis de esta tesis. El poder predictivo obtenido es incluso mayor al que se obtiene con modelos que incorporan directamente más información, lo que permite reducir costes en I+D.

## 9 Glosario

$\alpha_i$	: Factor de exposición al solvente de cada aminoácido de clase $i$
$\Gamma$	: Área superficial media asociada a una propiedad
$\varphi_i$	: Componente número $i$ de un vector de propiedad aminoacídica (APV)
$\xi$	: Señal externa en los algoritmos de redes neuronales
$\hat{\alpha}_i^l$	: Fracción del área correspondiente al aminoácido de tipo $i$ dentro de la máxima área superficial accesible (ASA)
$A$	: Conjunto de 20 aminoácidos posibles
<b>Ala</b>	: Aminoácido Alanina
<b>ANNs</b>	: Algoritmos de análisis de datos basados en Redes Neuronales
<b>APVs</b>	: Vectores de propiedades aminoacídicas
<b>Arg</b>	: Aminoácido Arginina
<b>ASA</b>	: Área Superficial Accesible
<b>ASH</b>	: Hidrofobicidad Superficial Media
<b>Asn</b>	: Aminoácido Asparagina
<b>Asp</b>	: Aminoácido Ácido Aspártico
<b>ASP</b>	: Área superficial media asociada a una propiedad
<b>ATPS</b>	: Sistema de dos fases acuosas
<b>Cut off</b>	: Radios o nivel de prioridad de los APV
<b>Cys</b>	: Aminoácido Cisteína
$D_k$	: Valor experimental de una propiedad de la proteína $k$ en un sistema
$\widehat{D}_k$	: Valor predicho de una propiedad de la proteína $k$ en un sistema
$\widehat{D}_k^{-k}$	: Valor predicho de una propiedad de la proteína $k$ en un sistema, sin utilizar esta proteína en el ajuste del modelo.

<b>DRT</b>	: Tiempo de Retención Dimensional
<b>GA</b>	: Algoritmo de Optimización Genética
<b>GG</b>	: Algoritmo de <i>clustering</i> Growing Grid
<b>Gln</b>	: Aminoácido Glutamina
<b>Glu</b>	: Aminoácido Ácido Glutámico
<b>Gly</b>	: Aminoácido Glicina
<b>GNG</b>	: Algoritmo de <i>clustering</i> Growing Neuronal Gas
<b>HA</b>	: Algoritmo de <i>clustering Hierarchical Algorithm</i>
<b>HCA</b>	: Área Hidrofóbica de Contacto
<b>HIC</b>	: Cromatografía de Interacción Hidrofóbica
<b>His</b>	: Aminoácido Histidina
$j^*$	: Conjunto de valores de los parámetros del un algoritmo, para el cual se ha detectado la máxima varianza
<b>Ile</b>	: Aminoácido Isoleucina
<b>K</b>	: Coeficiente de partición de proteínas en ATPS
$\hat{l}$	: Razón entre la longitud de la proteína y el máximo largo observado en la base de datos de trabajo
<b>Leu</b>	: Aminoácido Leucina
$\log\Gamma_0$	: Propiedad intrínseca del sistema
<b>LH</b>	: Hidrofobicidad local
<b>Lys</b>	: Aminoácido Lisina
<b>MCL</b>	: Algoritmo de <i>clustering Markov Clustering Algorithm</i>
$n_i$	: Número de aminoácido de clase $i$ en la proteína
<b>PDB</b>	: <i>Protein Data Base</i>
<b>Phe</b>	: Aminoácido Fenilalanina
$\hat{r}_i$	: Fracción de la superficie proteica ocupado por el aminoácido de clase $i$
<b>RASA</b>	: Razón entre el área superficial media (ASA) de un aminoácido y su máximo ASA
<b>RNS</b>	: Algoritmo de <i>clustering Restricted Neighborhood Search</i>
$s_{ij}^*$	: Raíz de la desviación estándar de los resultados obtenidos con el algoritmo $i$ , para el conjunto de parámetros $j^*$
$s_{ij}$	: Raíz de la desviación estándar de los resultados obtenidos con el algoritmo $i$ , para el conjunto de parámetros $j$
$s_i$	: Raíz de la desviación estándar de los resultados obtenidos con el algoritmo $i$
$s_{inj}^*$	: Raíz de la desviación estándar de los resultados obtenidos con el algoritmo $i$ , para el conjunto de parámetros $j^*$ , luego de $n$ ejecuciones
$S_i$	: Suma del área superficial accesible (ASA) para todo aminoácido de clase $i$
$S_{max,i}$	: Máxima área superficial media posible (ASA), para todos los aminoácidos de clase $i$
<b>Ser</b>	: Aminoácido Serina
<b>SOM</b>	: Algoritmo de <i>clustering Self-Organizing Maps</i>
$t_R$	: Tiempo para el cual se obtiene el <i>peak</i> de la cromatografía
$t_0$	: Tiempo en el que comienza el gradiente de sal
$t_f$	: Tiempo en el que termina el gradiente de sal.
<b>Thr</b>	: Aminoácido Tironiana

<b>Trp</b>	: Aminoácido Triptófano
<b>Tyr</b>	: Aminoácido Tirosina
<b>Val</b>	: Aminoácido Valina
<b>Z</b>	: Solución de un algoritmo
<b>Z<sub>ij</sub></b>	: Conjunto de soluciones obtenidas con el algoritmo <i>i</i> y un conjunto de parámetros fijo <i>j</i>
<b>Z<sub>inj</sub>*</b>	: Conjunto de soluciones obtenidas con el algoritmo <i>i</i> , <i>n</i> ejecuciones, para el conjunto de parámetros fijo <i>j</i> *

## 10 Bibliografía

1. Casper, S. y H. Kettler, *National institutional frameworks and the hybridization of entrepreneurial business models: the german and UK biotechnology sectors*. Industry and Innovation, 2001. **8**(1): p. 5-30.
2. Fausnaugh, J.L. y F.E. Regnier, *Solute and mobile phase contributions to retention in hydrophobic interaction chromatography of proteins*. Journal of chromatography, 1986. **359**: p. 131-46.
3. Mahn, A., M.E. Lienqueo y J.A. Asenjo, *Effect of surface hydrophobicity distribution on retention of ribonucleases in hydrophobic interaction chromatography*. Journal of chromatography. A, 2004. **1043**(1): p. 47-55.
4. Salgado, J.C., B.A. Andrews, M.F. Ortuzar y J.A. Asenjo, *Prediction of the partitioning behaviour of proteins in aqueous two-phase systems using only their amino acid composition*. Journal of chromatography. A, 2008. **1178**(1-2): p. 134-44.
5. Mahn, A. y J.A. Asenjo, *Prediction of protein retention in hydrophobic interaction chromatography*. Biotechnology advances, 2005. **23**(5): p. 359-68.
6. Mahn, A., G. Zapata-Torres y J.A. Asenjo, *A theory of protein-resin interaction in hydrophobic interaction chromatography*. Journal of chromatography. A, 2005. **1066**(1-2): p. 81-8.
7. Albertsson, P.A., *Partition of Cell Particles and Macromolecules*. John Wiley & Sons, 1986. New York.
8. Eiteman, M.A. y J.L. Gainer, *Predicting partition coefficients in polyethylene glycol-potassium phosphate aqueous two-phase systems*. Journal of chromatography, 1991. **586**: p. 341-6.
9. Salgado, J.C., I. Rapaport y J.A. Asenjo, *Prediction of retention times of proteins in hydrophobic interaction chromatography using only their amino acid composition*. Journal of chromatography. A, 2005. **1098**(1-2): p. 44-54.
10. Wolfson, H., *Protein Structure, en Algorithms for Molecular Biology*. Tel Aviv University School of Computer Science, 2002. Disponible en: <http://www.cs.tau.ac.il/~rshamir/algmb/01/algmb01.html>.
11. *Protein data bank*. 1971 [30/03/2011]. Disponible en: <http://www.pdb.org/pdb/home/contactUs.do>.
12. Rutgers y UCSD, *Anual Report Protein data bank, en* 2010. 1-16.
13. Salgado, J.C., I. Rapaport y J.A. Asenjo, *Is it possible to predict the average surface hydrophobicity of a protein using only its amino acid composition?* Journal of chromatography. A, 2005. **1075**(1-2): p. 133-43.
14. Laboratory, K., *Amino acid indices, substitution matrices and pair-wise contact potentials*. Kyoto University, 2008 [28/03/2011]. Disponible en: <http://www.genome.jp/aaindex/>.
15. Jain, A.K., M.N. Murty y P.J. Flynn, *Data Clustering: A Review*. ACM Computing Surveys, 1999. **31**(3): p. 264-323.
16. Fritzke, B., *A growing neural gas network learns topologies*. Advances in Neural Information Processing Systems, 1995: p. 625-32.

17. Elvin, C.M., A.G. Carr, M.G. Huson, J.M. Maxwell, R.D. Pearson, T. Vuocolo, N.E. Liyou, D.C. Wong, D.J. Merritt y N.E. Dixon, *Synthesis and properties of crosslinked recombinant pro-resilin*. Nature, 2005. **437**(7061): p. 999-1002.
18. Andersen, S.O. y T. Weis-Fogh, *Resilin. A rubber-like protein in arthropod cuticle*. Adv. Insect Physiol., 1964. **2**: p. 1-65.
19. Walker, G., F. Cai, P. Shen, C. Reynolds, B. Ward, C. Fone, S. Honda, M. Koganei, M. Oda y E. Reynolds, *Increased remineralization of tooth enamel by milk containing added casein phosphopeptide-amorphous calcium phosphate*. The Journal of dairy research, 2006. **73**(1): p. 74-8.
20. Fletcher, G.L., C.L. Hew y P.L. Davies, *Antifreeze proteins of teleost fishes*. Annual review of physiology, 2001. **63**: p. 359-90.
21. Jorov, A., B.S. Zhorov y D.S. Yang, *Theoretical study of interaction of winter flounder antifreeze protein with ice*. Protein science : a publication of the Protein Society, 2004. **13**(6): p. 1524-37.
22. Andrews, B.A., A.S. Schmidt y J.A. Asenjo, *Correlation for the partition behavior of proteins in aqueous two-phase systems: effect of surface hydrophobicity and charge*. Biotechnology and bioengineering, 2005. **90**(3): p. 380-90.
23. Asenjo, J.A., A.S. Schmidt, F. Hachem y B.A. Andrews, *Model for predicting the partition behaviour of proteins in aqueous two-phase systems*. Journal of chromatography, 1994. **668**: p. 47-54.
24. Schmidt, A.S., *An investigation of the partition behaviour of proteins based on their physico-chemical properties in aqueous two-phase systems*. Faculty of Agriculture and Food, University of Reading, 1994.
25. Hofstee, B.H., *Hydrophobic affinity chromatography of proteins*. Anal Biochem, 1973. **52**(2): p. 430-48.
26. Shaltiel, S. y Z. Er-El, *Hydrophobic chromatography: use for purification of glycogen synthetase*. Proceedings of the National Academy of Sciences of the United States of America, 1973. **70**(3): p. 778-81.
27. Porath, J., L. Sundberg, N. Fornstedt y I. Olsson, *Salting-out in amphiphilic gels as a new approach to hydrophobic adsorption*. Nature, 1973. **245**(5426): p. 465-6.
28. Hjertén, S., *Fractionation of proteins by hydrophobic interaction chromatography, with reference to serum proteins*. Proceedings Intl. Workshop on Technology for Protein Separation & Improvement of Blood Plasma Fractionation, 1977: p. 410-21.
29. biotech, A.p., *Hydrophobic Interaction Chromatography: Principles and Methods*, Edition AB ed. Amersham Biosciences, 1993.
30. Rosengren, J., S. Pahlman, M. Glad y S. Hjerten, *Hydrophobic interaction chromatography on non-charged Sepharose derivatives. Binding of a model protein, related to ionic strength, hydrophobicity of the substituent, and degree of substitution (determined by NMR)*. Biochimica et biophysica acta, 1975. **412**(1): p. 51-61.
31. Laas, T., *Agardervatives for chromatography, electrophoresis and gel-bound enzymes. IV. Benzylated dibromopropanol cross-linked sepharose as an amphiphilic gel for hydrophobic salting-out chromatography of enzymes with special emphasis on denaturing risks*. Journal of chromatography, 1975. **111**(2): p. 373-87.
32. Maisano, F., M. Belew y J. Porath, *Synthesis of new hydrophobic adsorbents based on homologous series of uncharged alkyl sulphide agarose derivatives*. Journal of chromatography, 1985. **321**(2): p. 305-17.
33. Tanford, C., *Contribution of hydrophobic interactions to the stability of globular conformation of proteins*. J Am Chem Soc 1962. **84**: p. 4240-7.
34. Janson, J.C. y T. Låås, *Hydrophobic interaction chromatography on Phenyl- and Octyl-Sepharose CL-4B*. Chromatography of synthetic and biological macromolecules, 1978.

35. Melander, W. y C. Horvath, *Salt effect on hydrophobic interactions in precipitation and chromatography of proteins: an interpretation of the lyotropic series*. Archives of biochemistry and biophysics, 1977. **183**(1): p. 200-15.
36. Srinivasan, R. y E. Ruckenstein, *Role of Physical Forces in Hydrophobic Interaction Chromatography*. Separation & Purification Methods., 1980. **9**: p. 267–370.
37. Hjertén, S., K. Yao, K.O. Eriksson y B. Johansson, *Gradient and isocratic High Performance Hydrophobic Interaction Chromatography of proteins on agarose columns*. J. Chromatog., 1986(359): p. 99–109.
38. Jennissen, H.P., *Multivalent interaction chromatography as exemplified by the adsorption and desorption of skeletal muscle enzymes on hydrophobic alkyl-agaroses*. Journal of chromatography, 1978. **159**(1): p. 71-83.
39. Parsegian, V.A. y B.W. Ninham, *Temperature-dependent van der Waals forces*. Biophysical journal, 1970. **10**(7): p. 664-74.
40. Visser, J. y M. Strating, *Separation of lipoamide dehydrogenase isoenzymes by affinity chromatography*. Biochimica et biophysica acta, 1975. **384**(1): p. 69-80.
41. Xia, F., D. Nagrath y S.M. Cramer, *Modeling of adsorption in hydrophobic interaction chromatography systems using a preferential interaction quadratic isotherm*. Journal of chromatography. A, 2003. **989**(1): p. 47-54.
42. Arakawa, T. y S.N. Timasheff, *Mechanism of poly(ethylene glycol) interaction with proteins*. Biochemistry, 1985. **24**(24): p. 6756-62.
43. Arakawa, T., R. Bhat y S.N. Timasheff, *Preferential interactions determine protein solubility in three-component solutions: the MgCl<sub>2</sub> system*. Biochemistry, 1990. **29**(7): p. 1914-23.
44. Arakawa, T. y S.N. Timasheff, *Preferential interactions of proteins with salts in concentrated solutions*. Biochemistry, 1982. **21**(25): p. 6545-52.
45. Timasheff, S.N. y T. Arakawa, *Mechanism of Protein Precipitation and Stabilization by Co-solvents*. J. Crystal Growth 1988. **90**(39).
46. Fauchere, J.L., K.Q. Do, P.Y. Jow y C. Hansch, *Unusually strong lipophilicity of 'fat' or 'super' amino-acids, including a new reference value for glycine*. Experientia, 1980. **36**(10): p. 1203-4.
47. Jones, D.D., *Amino acid properties and side-chain orientation in proteins: a cross correlation approach*. Journal of theoretical biology, 1975. **50**(1): p. 167-83.
48. Zaslavsky, B.Y., N.M. Mestechkina, L.M. Miheeva y S.V. Rogozhin, J. Chromatogr., 1982. **240**
49. Rekker, R.F., *The hydrophobic fragmental constant*. Elsevier, 1977. New York.
50. Zaslavsky, B.Y., N.M. Mestechkina, L.M. Miheeva, S.V. Rogozhin, G. Bakalkin, G.G. Rjazhsky, E.V. Chetverina, A.A. Asmuko, J.D. Bespalova, N.V. Korobov y O.N. Chichenkov, *Correlation of hydrophobic character of opioid peptides with their biological activity measured in various bioassay systems*. Biochemical pharmacology, 1982. **31**(23): p. 3757-62.
51. Perkins, T.W., D.S. Mak, T.W. Root y E.N. Lightfoot, *Protein retention in hydrophobic interaction chromatography: modeling variation with buffer ionic strength and column hydrophobicity*. Journal of Chromatography A, 1997. **766**: p. 1-14.
52. Eriksson, K., *Hydrophobic interaction chromatography, en Protein purification: principles, high-resolution methods, and applications*. Wiley-Liss, 1998. New York.
53. Berggren, K., A. Wolf, J.A. Asenjo, B.A. Andrews y F. Tjerneld, *The surface exposed amino acid residues of monomeric proteins determine the partitioning in aqueous two-phase systems*. Biochimica et biophysica acta, 2002. **1596**(2): p. 253-68.
54. Lienqueo, M.E., A. Mahn y J.A. Asenjo, *Mathematical correlations for predicting protein retention times in hydrophobic interaction chromatography*. Journal of chromatography. A, 2002. **978**(1-2): p. 71-9.
55. Gautam, M. y L. Simon, *Partitioning of  $\beta$ -glucosidase from Trichoderma reesei in poly(ethylene glycol) and potassium phosphate aqueous two-phase systems: Influence of pH and temperature*. Biochem. Eng. J., 2006. **30**(1): p. 104-8.

56. Sasakawa, S. y H. Walter, *Partition behavior of native proteins in aqueous dextran-poly(ethylene glycol)-phase systems*. Biochemistry, 1972. **11**(15): p. 2760-5.
57. Abelson, J. y M. Simon, *Aqueous Two-phase Systems, en Methods in Enzymology*. Academic Press, 1994. New York.
58. Johansson, G., *Effects of salts on the partition of proteins in aqueous polymeric biphasic systems*. Acta chemica Scandinavica. Series B: Organic chemistry and biochemistry, 1974. **28**(8): p. 873-82.
59. Johansson, G., *Partition of proteins and micro-organisms in aqueous biphasic systems*. Molecular and cellular biochemistry, 1974. **4**(3): p. 169-80.
60. Eiteman, M.A. y J.L. Gainer, *Peptide hydrophobicity and partitioning in poly(ethylene glycol)/magnesium sulfate aqueous two-phase systems*. Biotechnology progress, 1990. **6**(6): p. 479-84.
61. Berggren, K., H.-O. Johansson y F. Tjerneld, J. Chromatogr. A, 1995(718): p. 67.
62. Hachem, F., B. Andrews y J.A. Asenjo, *Hydrophobic partitioning of proteins in aqueous two-phase systems*. Enzyme Microb. Technol., 1996. **19**(7): p. 507-17.
63. Shanbhag, V., *Estimation of Surface Hydrophobicity of Proteins by Partitioning*. Academic Press, 1994. San Diego, California.
64. Shanbhag, V.P. y C.G. Axelsson, Eur. J. Biochem., 1975. **60**: p. 17.
65. Shanbhag, V.P. y G. Johansson, Biochem. Biophys. Res. Commun., 1974. **61**: p. 1141.
66. Carlsson, M., P. Linse y F. Tjerneld, *Temperature-Dependent Protein Partitioning in Two-Phase Aqueous Polymer Systems*. Macromolecules, 1993. **26**: p. 1546-54.
67. Eiteman, M.A. y J.L. Gainer, *Partition of isomeric dipeptides in poly(ethylene glycol)/magnesium sulfate aqueous two-phase systems*. Biochimica et biophysica acta, 1991. **1073**(3): p. 451-5.
68. Eiteman, M.A., *Predicting partition coefficients of multi-charged solutes in aqueous two-phase systems*. Journal of chromatography. A, 1994. **668**: p. 21-30.
69. Johansson, G., P.A. Albertsson y F. Tjerneld, *Aqueous Two-Phase Separations, en Separation Processes in Biotechnology*. CRC Press, 1990.
70. Lin, D.Q., Y.T. Wu, L.H. Mei, Z.Q. Zhu y S.J. Yao, *Modeling the protein partitioning in aqueous polymer two-phase systems: Influence of polymer concentration and molecular weight*. Chemical Engineering Science, 2003. **58**: p. 2963-72.
71. Riveros, N., *Predicción del coeficiente de partición de proteínas en sistemas de dos fases acuosas a partir de la energía de solvatación y la hidrofobicidad*. Departamento de Ingeniería Química y Biotecnología, Universidad de Chile, 2009. Santiago, Chile.
72. Trindade, I.P., M.M. Diogo, D.M. Prazeres y J.C. Marcos, *Purification of plasmid DNA vectors by aqueous two-phase extraction and hydrophobic interaction chromatography*. Journal of chromatography. A, 2005. **1082**(2): p. 176-84.
73. Olivera-Nappa, A., G. Lagomarsino, B.A. Andrews y J.A. Asenjo, *Effect of electrostatic energy on partitioning of proteins in aqueous two-phase systems*. Journal of chromatography. B, Analytical technologies in the biomedical and life sciences, 2004. **807**(1): p. 81-6.
74. Anderberg, M.R., *Cluster Analysis for Applications, en Academic Press*, 1973. New York.
75. Jain, A.K. y P.J. Flynn, *Image segmentation using clustering, en Advances in Image Understanding*. IEEE Press, 1996. Piscataway, New Jersey.
76. Rasmussen, E., *Clustering algorithms, en Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, 1992. Upper Saddle River, New Jersey.
77. Jain, A.K. y R.C. Dubes, *Algorithms for Clustering Data, en Prentice-Hall advanced reference series*. Prentice-Hall, 1988. Upper Saddle River, New Jersey.
78. Rasmussen, M. y G. Karypis, *gCLUTO: An Interactive Clustering, Visualization, and Analysis System, en CSE/UMN Technical Report*. University of Minnesota, Department of Computer Science and Engineering, 2004. Minneapolis.

79. Mao, J. y A.K. Jain, *A self-organizing network for hyperellipsoidal clustering (HEC)*. IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council, 1996. **7**(1): p. 16-29.
80. Watts, D.J. y S.H. Strogatz, *Collective dynamics of 'small-world' networks*. Nature, 1998. **393**(6684): p. 440-2.
81. Demetrius, L. y T. Manke, *Robustness and network evolution—an entropic principle*. Physica A, 2005. **346**: p. 682–96.
82. Dubes, R.C., *How many clusters are best?—an experiment*. Pattern Recogn., 1987. **20**(6): p. 645–63.
83. Cheng, Y., *Mean shift, mode seeking, and clustering*. IEEE Trans. Pattern Anal. Mach. Intell., 1995. **17**(7): p. 790–9.
84. Weise, T., *Global Optimization Algorithms, Theory and Application*. Self-published, 2006 [20/11/2011]. Disponible en: <http://www.it-weise.de/>.
85. Barabasi, A.L. y Z.N. Oltvai, *Network biology: understanding the cell's functional organization*. Nature reviews. Genetics, 2004. **5**(2): p. 101-13.
86. Zhao, Y. y G. Karypis, *Evaluation of Clustering Algorithms for Document Datasets*. 11th Conference of Information and Knowledge Management (CIKM). 2002: p. 515-24.
87. King, B., *Step-wise clustering procedures*. J. Am. Stat. Assoc., 1967. **69**: p. 86–101.
88. Ward, J.H., *Hierarchical grouping to optimize an objective function*. J. Am. Stat. Assoc., 1963. **58**: p. 236–44.
89. Murtagh, F., *A survey of recent advances in hierarchical clustering algorithms which use cluster centers*. Comput. J., 1984. **26**: p. 354–9.
90. Nagy, G., *State of the art in pattern recognition*. Proc. IEEE 1968. **56**: p. 836–62.
91. Baeza-Yates, R.A., *Introduction to data structures and algorithms related to information retrieval, en Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, 1992. Upper Saddle River, New Jersey.
92. McQueen, J. *Some methods for classification and analysis of multivariate observations, en Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. 1967.
93. Bezdek, J.C., *Pattern recognition with fuzzy objective function algorithms, en Advanced applications in pattern recognition*. Plenum Press, 1981. xv, 256 p. New York.
94. Dave, R.N., *Generalized fuzzy C-shells clustering and detection of circular and elliptic boundaries*. Pattern Recogn., 1992. **25**: p. 713–22.
95. Bäck, T., *Evolutionary algorithms, en Theory and practice : evolution strategies, evolutionary programming, genetic algorithms*. Oxford University Press, 1996. xii, 314 p. New York.
96. Back, T., *Handbook of Evolutionary Computation, Ringbound edition ed, en Computational Intelligence Library*. Oxford University Press in cooperation with the Institute of Physics Publishing, 1997. Bristol, New York.
97. Bäck, T., U. Hammel y H.-P. Schwefel, *Evolutionary computation: comments on the history and current state*. IEEE Transactions on Evolutionary Computation, 1997.
98. Russell, S.J., P. Norvig y J. Canny, *Artificial intelligence : a modern approach*, 2nd ed, en *Prentice Hall series in artificial intelligence*. Prentice Hall, 2003. xxviii, 1080 p. Upper Saddle River, New Jersey.
99. Matyas, J., *Random optimization*. Automation and Remote Control, 1965. **26**: p. 244–51.
100. Kirkpatrick, S., C.D. Gelatt, Jr. y M.P. Vecchi, *Optimization by simulated annealing*. Science, 1983. **220**(4598): p. 671-80.
101. Glover, F. y M. Laguna, *Tabu search, en Modern Heuristic Techniques for Combinatorial Problems*. Blackwell Scientific Publishing, Halsted Press, 1993. Oxford, England.
102. Glover, F. y M. Laguna, *Tabu search, en Handbook of Combinatorial Optimization*. Kluwer Academic Publishers, 1998. Springer Netherlands.

103. Hertz, A., E. Taillard y D. De Werra, *A tutorial on tabu search, en Proceedings of Giornate di Lavoro AIRO'95 (Enterprise Systems: Management of Technological and Organizational Changes)*, 1995. Italy.
104. Battiti, R. y G. Tecchiolli, *The reactive tabu search*. ORSA Journal on Computing., 1994. **6**(2): p. 126–40.
105. Cvijovicacute, D. y J. Klinowski, *Taboo search: an approach to the multiple minima problem*. Science, 1995. **267**(5198): p. 664-6.
106. King, A.D., N. Przulj y I. Jurisica, *Protein complex prediction via cost-based clustering*. Bioinformatics, 2004. **20**(17): p. 3013-20.
107. Brohee, S. y J. van Helden, *Evaluation of clustering algorithms for protein-protein interaction networks*. BMC Bioinformatics, 2006. **7**: p. 488.
108. Darwin, C., *The origin of species by means of natural selection; or, The preservation of favoured races in the struggle for life, en World's classics*. Oxford University Press, 1951. xxxii, 592 p. London.
109. Goldberg, D.E., *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley, 1989. xiii, 412 p. Reading, Mass; Wokingham.
110. Holland, J.H., *Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence, en Complex adaptive systems*. MIT Press, 1992. xiv, 211 p. Cambridge, Mass; London.
111. Holland, J.H., *Outline for a Logical Theory of Adaptive Systems*. J. ACM, 1962. **9**(3): p. 297-314.
112. Goldberg, D.E., *Genetic Algorithms in Search*, 1 ed, *en Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., 1989. Boston, MA, USA.
113. Holland, J.H., *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, 1975. Ann Arbor.
114. Holland, J.H., *Outline for a logical theory of adaptive systems*. Journal of the ACM., 1962. **9**(3): p. 297–314.
115. Crosby, J.L., *Computer simulation in genetics*. John Wiley & Sons, 1973. xiii, 477 p. London, New York.
116. Hertz, J., R.G. Palmer y A. Krogh, *Introduction to the theory of neural computation, en Santa Fe Institute studies in the sciences of complexity Lecture notes*. Addison-Wesley Pub. Co., 1991. xxii, 327 p. Redwood City, Calif.
117. Sethi, I.K. y A.K. Jain, *Artificial neural networks and statistical pattern recognition: old and new connections, en Machine intelligence and pattern recognition*. Elsevier Science Pub. Co. distributor, 1991. xiv, 271 p. New Jersey.
118. Jain, A.K. y J. Mao, *Neural networks and pattern recognition, en In Computational Intelligence: Imitating Life*, 1994.
119. Jain, A.K. y J. Mao, *Artificial neural networks: A tutorial*. IEEE Computer 1996. **29**: p. 31– 44.
120. Oja, E., *A simplified neuron model as a principal component analyzer*. Journal of mathematical biology, 1982. **15**(3): p. 267-73.
121. Sanger, T.D., *An optimality principle for unsupervised learning, en Advances in Neural Information Processing Systems*. Morgan Kaufmann, 1989. San Mateo, California.
122. Kohonen, T., *Self-organizing maps*, 3rd ed, *en Springer series in information sciences*. Springer, 2001. xx, 501 p. New York.
123. Fritzsche, B., *Growing cell structures - a self organizing network for unsupervised and supervised learning*. Neural Networks. 7, 1994. **9**: p. 1441-60
124. Martinez, T.M., *Competitive Hebbian Learning Rule Forms Perfectly Topology Preserving Maps, en In the Proceedings of the International Conference on Artificial Neural Networks*, 1993.
125. Martinez, T.M. y K.J. Schulten. *A "neural-gas" network learns topologies, en Proc. International Conference on Artificial Neural Networks*. 1991. 397- 402. Amsterdam, Netherlands, North-Holland.



126. Fritzke, B., *Growing Grid - a self-organizing network with constant neighborhood range and adaptation strength*. Neural Processing Letters, 1995. **2**(5): p. 9-13.
127. Mishra, S.K. y V.V. Raghavan, *An empirical study of the performance of heuristic methods for clustering, en Pattern Recognition in Practice*, 1994. North Holland.
128. Ismail, M.A. y M.S. Kamel, *Multidimensional data clustering utilizing hybrid search strategies*. Pattern Recogn., 1989. **22**(1): p. 75–89.
129. Al-Sultan, K.S. y M.M. Khan, *Computational experience on four algorithms for the hard clustering problem*. Pattern Recogn. Lett., 1996. **17**(3): p. 295–308.
130. Meek, J.L., *Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition*. Proceedings of the National Academy of Sciences of the United States of America, 1980. **77**(3): p. 1632-6.
131. MATLAB® 7.4.1.287(R2007a). 2007 [1/05/2008]. Disponible en: <http://www.mathworks.com>.
132. S, V.D., *Graph clustering by flow simulation*. Centers for mathematics and computer science (CWI), University of Utrecht, 2000.
133. Enright, A.J., S. Van Dongen y C.A. Ouzounis, *An efficient algorithm for large-scale detection of protein families*. Nucleic Acids Research, 2002. **30**(7): p. 1575-84.
134. Brohée, S. y K. Faust, *Network Analysis Tools (NeAT) Tutorial*. Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRE), Université Libre de Bruxelles, 2009 [17/02/2011]. Disponible en: [http://rsat.ulb.ac.be/tutorials/neat\\_tutorial.pdf](http://rsat.ulb.ac.be/tutorials/neat_tutorial.pdf).
135. Brohée, S., *Network analysis tools*. 2008 [28/03/2011]. Disponible en: [http://rsat.bigre.ulb.ac.be/rsat/index\\_neat.html](http://rsat.bigre.ulb.ac.be/rsat/index_neat.html).
136. Cortés, M.P., *Predicción del coeficiente de partición de proteínas en sistemas de dos fases acuosas a través de la caracterización bioinformática de su superficie*. Departamento de Ingeniería Química y Biotecnología, Universidad de Chile, 2008. Santiago, Chile.
137. Frishman, D. y P. Argos, *Knowledge-based protein secondary structure assignment*. Proteins, 1995. **23**(4): p. 566-79.
138. Miller, S., J. Janin, A.M. Lesk y C. Chothia, *Interior and surface of monomeric proteins*. Journal of molecular biology, 1987. **196**(3): p. 641-56.
139. Hobohm, U., M. Scharf, R. Schneider y C. Sander, *Selection of representative protein data sets*. Protein science : a publication of the Protein Society, 1992. **1**(3): p. 409-17.
140. Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov y P.E. Bourne, *The Protein Data Bank*. Nucleic Acids Research, 2000. **28**(1): p. 235-42.
141. Bernstein, F.C., T.F. Koetzle, G.J. Williams, E.F. Meyer, Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi y M. Tasumi, *The Protein Data Bank: a computer-based archival file for macromolecular structures*. Journal of molecular biology, 1977. **112**(3): p. 535-42.
142. Malmquist, G., U.H. Nilsson, M. Norrman, U. Skarp, M. Stromgren y E. Carredano, *Electrostatic calculations and quantitative protein retention models for ion exchange chromatography*. Journal of chromatography. A, 2006. **1115**(1-2): p. 164-86.
143. Fauchere, J.-L. y V. Pliska, *Hydrophobic parameters  $\pi$  of amino acid side chains from partitioning of N-acetyl-amino-acid amides*. Eur J Med Chem, 1983. **18**: p. 369 - 75.
144. Parker, J.M., D. Guo y R.S. Hodges, *New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites*. Biochemistry, 1986. **25**(19): p. 5425-32.
145. Jesior, J.C., *Hydrophilic framework in proteins?* Journal of protein chemistry, 2000. **19**(2): p. 93-103.
146. Chothia, C., *The nature of the accessible and buried surfaces in proteins*. Journal of molecular biology, 1976. **105**(1): p. 1-12.
147. Rose, G.D., A.R. Geselowitz, G.J. Lesser, R.H. Lee y M.H. Zehfus, *Hydrophobicity of amino acid residues in globular proteins*. Science, 1985. **229**(4716): p. 834-8.

148. Deleage, G. y B. Roux, *An algorithm for protein secondary structure prediction based on class prediction*. Protein engineering, 1987. **1**(4): p. 289-94.
149. Lifson, S. y C. Sander, *Antiparallel and parallel beta-strands differ in amino acid residue preferences*. Nature, 1979. **282**(5734): p. 109-11.
150. Chou, P.Y. y G.D. Fasman, *Prediction of the secondary structure of proteins from their amino acid sequence*. Advances in enzymology and related areas of molecular biology, 1978. **47**: p. 45-148.
151. Levitt, M., *Conformational preferences of amino acids in globular proteins*. Biochemistry, 1978. **17**(20): p. 4277-85.
152. Wolfenden, R., L. Andersson, P.M. Cullis y C.C. Southgate, *Affinities of amino acid side chains for solvent water*. Biochemistry, 1981. **20**(4): p. 849-55.
153. Hopp, T.P. y K.R. Woods, *Prediction of protein antigenic determinants from amino acid sequences*. Proceedings of the National Academy of Sciences of the United States of America, 1981. **78**(6): p. 3824-8.
154. Bull, H.B. y K. Breese, *Surface tension of amino acid solutions: a hydrophobicity scale of the amino acid residues*. Archives of biochemistry and biophysics, 1974. **161**(2): p. 665-70.
155. Kyte, J. y R.F. Doolittle, *A simple method for displaying the hydropathic character of a protein*. Journal of molecular biology, 1982. **157**(1): p. 105-32.
156. Browne, C.A., H.P. Bennett y S. Solomon, *The isolation of peptides by high-performance liquid chromatography using predicted elution positions*. Analytical biochemistry, 1982. **124**(1): p. 201-8.
157. Miyazawa, S. y R.L. Jernigan, *Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation*. Macromolecules, 1985. **18**(3): p. 534-52.
158. Cowan, R. y R.G. Whittaker, *Hydrophobicity indices for amino acid residues as determined by high-performance liquid chromatography*. Peptide research, 1990. **3**(2): p. 75-80.
159. Aboderin, A.A., *An empirical hydrophobicity scale for [alpha]-amino-acids and some of its applications*. International Journal of Biochemistry, 1971. **2**(11): p. 537-44.
160. Abraham, D.J. y A.J. Leo, *Extension of the fragment method to calculate amino acid zwitterion and side chain partition coefficients*. Proteins, 1987. **2**(2): p. 130-52.
161. Bairoch, A., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003*. Nucleic Acids Research, 2003. **31**(1): p. 365-70.
162. Bhaskaran, R. y P.K. Ponnuswamy, *Positional flexibilities of amino acid residues in globular proteins*. Int. J. Pept. Protein Res., 1988. **32**: p. 242 -55.
163. Black, S.D. y D.R. Mould, *Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications*. Analytical biochemistry, 1991. **193**(1): p. 72-82.
164. Dayhoff, M.O., *Atlas of Protein Sequence and Structure*, 1978. 470.
165. Eisenberg, D., E. Schwarz, M. Komaromy y R. Wall, *Analysis of membrane and surface protein sequences with the hydrophobic moment plot*. Journal of molecular biology, 1984. **179**(1): p. 125-42.
166. Eriksson, K.O., *Protein Purification: Principles, High-Resolution Methods and Applications*, 2 ed. Wiley-Liss, 1998. New York.
167. Fraga, S., *Theoretical prediction of protein antigenic determinants from amino acid sequences*. Can. J. Chem., 1982. **60**: p. 2606-10.
168. Grantham, R., *Amino acid difference formula to help explain protein evolution*. Science, 1974. **185**(4154): p. 862-4.
169. Guy, H.R., *Amino acid side-chain partition energies and distribution of residues in soluble proteins*. Biophysical journal, 1985. **47**(1): p. 61-70.
170. Hellberg, S., M. Sjoström, B. Skagerberg y S. Wold, *Peptide quantitative structure-activity relationships, a multivariate approach*. Journal of medicinal chemistry, 1987. **30**(7): p. 1126-35.
171. Janin, J., *Surface and inside volumes in globular proteins*. Nature, 1979. **277**(5696): p. 491-2.

172. Jonsson, J., L. Eriksson, S. Hellberg, M. Sjöström y S. Wold, *Multivariate Parametrization of 55 Coded and Non-Coded Amino Acids*. Quantitative Structure-Activity Relationships, 1989. **8**(3): p. 204-9.
173. Manavalan, P. y P.K. Ponnuswamy, *Hydrophobic character of amino acid residues in globular proteins*. Nature, 1978. **275**(5681): p. 673-4.
174. McCaldon, P. y P. Argos, *Oligopeptide biases in protein sequences and their use in predicting protein coding regions in nucleotide sequences*. Proteins, 1988. **4**(2): p. 99-122.
175. Rao, M.J.K. y P. Argos, *A conformational preference parameter to predict helices in integral membrane proteins*. Biochim. Biophys. Acta, 1986. **869**: p. 197- 214.
176. Roseman, M.A., *Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds*. Journal of molecular biology, 1988. **200**(3): p. 513-22.
177. Sandberg, M., L. Eriksson, J. Jonsson, M. Sjoström y S. Wold, *New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids*. Journal of medicinal chemistry, 1998. **41**(14): p. 2481-91.
178. Sweet, R.M. y D. Eisenberg, *Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure*. Journal of molecular biology, 1983. **171**(4): p. 479-88.
179. Welling, G.W., W.J. Weijer, R. van der Zee y S. Welling-Wester, *Prediction of sequential antigenic regions in proteins*. FEBS letters, 1985. **188**(2): p. 215-8.
180. Wertz, D.H. y H.A. Scheraga, *Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule*. Macromolecules, 1978. **11**(1): p. 9-15.
181. Wilson, K.J., A. Honegger, R.P. Stotzel y G.J. Hughes, *The behaviour of peptides on reverse-phase supports during high-pressure liquid chromatography*. The Biochemical journal, 1981. **199**(1): p. 31-41.
182. Zimmerman, J.M., N. Eliezer y R. Simha, *The characterization of amino acid sequences in proteins by statistical methods*. Journal of theoretical biology, 1968. **21**(2): p. 170-201.
183. Chou, K.C. y C.T. Zhang, *Prediction of protein structural classes*. Critical reviews in biochemistry and molecular biology, 1995. **30**(4): p. 275-349.
184. Zhou, G.P., *An intriguing controversy over protein structural class prediction*. Journal of protein chemistry, 1998. **17**(8): p. 729-38.
185. Zhou, G.P. y N. Assa-Munt, *Some insights into protein structural class prediction*. Proteins, 2001. **44**(1): p. 57-9.
186. Fritzsche, B., *Some Competitive Learning Methods*. 1997 [28/03/2011]. Disponible en: <http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/research/gsn/JavaPaper/>.
187. Albert, R. y A.L. Barabási, *Statistical mechanics of complex networks*. Rev. Mod. Phys, 2002. **74**: p. 47-97.
188. Yungler, L.M. y R.D. Cramer, 3rd, *Measurement of correlation of partition coefficients of polar amino acids*. Molecular pharmacology, 1981. **20**(3): p. 602-8.
189. Hansch, C. y A. Leo, *Substituent Constants for Correlation, en Analysis in Chemistry and Biology*. John Wiley & Sons, 1979. 18-43. New York.
190. Ben-Naim, A., *Hydrophobic interactions*. Plenum Press, 1980. xiii, 311 p. New York.
191. Levine, I.N., *Physical chemistry*, 5th ed. McGraw-Hill, 2003. xxi, 986 p. New York ; London.
192. Nozaki, Y. y C. Tanford, *The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale*. The Journal of biological chemistry, 1971. **246**(7): p. 2211-7.
193. Pliska, V., M. Schmidt y J.L. Fauchere, J. Chromatogr., 1981. **216**: p. 79.
194. Chothia, C., *Hydrophobic bonding and accessible surface area in proteins*. Nature, 1974. **248**(446): p. 338-9.
195. Chothia, C., *Structural invariants in protein folding*. Nature, 1975 **254**(5498): p. 304-8.

196. Huyghes-Despointes, B.M.P., T.M. Klinger y R.L. Baldwin, *Measuring the Strength of Side-Chain Hydrogen Bonds in Peptide Helices: The Gln- Asp (i,i+4) interaction*. *Biochem*, 1995. **34**: p. 13267-71.
197. Sofer, G. y L. Hagel, *Handbook of process chromatography: a guide to optimization, scale-up, and validation*. . Academic Press, 1998: p. 387.
198. Jennissen, H.P., *Hydrophobic interaction chromatography*. . *Int J Bio-Chromatogr* 2000. **5**: p. 131–8.
199. S.A., M.d.M., *Plano esquemático de la red de Metro de Madrid*. 2011 [28/3/2011]. Disponible en: <http://www.metromadrid.es/export/sites/metro/comun/documentos/planos/Planoesp2011-04.pdf>.

## Anexo A: definiciones

<b>Algoritmo</b>	: Conjunto ordenado y finito de operaciones que permite hallar la solución de un problema.
<b>Algoritmo de Optimización</b>	: Conjunto ordenado y finito de operaciones que se utiliza para hallar la solución óptima de un problema.
<b>Algoritmos de Análisis Topológico</b>	: Conjunto ordenado y finito de operaciones que utiliza la información topológica de un conjunto de objetos o datos con un objetivo determinado. En esta tesis se refieren a los algoritmos de <i>clustering</i> .
<b>Algoritmos de Clustering</b>	: Conjunto ordenado y finito de operaciones que se utiliza para obtener una partición natural o inherente a un conjunto de objetos o datos.
<b>Clasificación no Supervisada Cluster</b>	: Que utiliza algoritmos de <i>clustering</i> .
	: Conjunto de objetos o datos agrupados en base a un criterio de similitud, que son parte de un grupo mayor de objetos o datos.
<b>Error de Jackknife</b>	: Indica de la capacidad predictiva de un modelo.
<b>Coeficiente de Partición</b>	: Indica la proporción de una molécula en la fase superior de ATPS, con respecto a la fase inferior.
<b>Coeficiente de Pearson</b>	: Indica el nivel de error asociado al ajuste de un modelo a datos experimentales.
<b>Concordancia</b>	: Coincidencia entre los APVs más influyentes y/o cercanos a las mejores soluciones, obtenidas para iguales o distintos sistemas, con iguales o distintos modelos y/o tipos de algoritmos, y que no son producto del azar.
<b>Consistencia</b>	: Relación coherente entre los APVs más cercanos e influyentes a las mejores escalas generadas a partir de algoritmos de <i>clustering</i> o de optimización.
<b>Cromatografía de Interacción Hidrofóbica</b>	: Sistema de separación de moléculas que utiliza una columna con una matriz para retener o ralentizar las moléculas en base a su hidrofobicidad.
<b>Curva Binoidal</b>	: Curva que resume las distintas concentraciones mínimas de los compuestos que caracterizan cada una de las fases en un sistema de dos fases acuosas, a partir de las cuales se genera las dos fases.
<b>Determinísticos</b>	: Certero, que no utiliza el azar.
<b>Escala de Propiedades Aminoacídicas (APV)</b>	: Conjunto de valores que indican el carácter respecto de una propiedad específica que poseen cada uno de los 20 aminoácidos.
<b>Escalas de Hidrofobicidad</b>	: Conjunto de valores que indican el carácter hidrofóbico de los 20 aminoácidos.
<b>Estocástico</b>	: Pertenciente o relativo al azar.
<b>gCluto</b>	: Software altamente especializado para la búsqueda de <i>clusters</i> y visualización de soluciones.

<b>Grafo</b>	: Herramienta de representación de un sistema a partir de los objetos que lo componen y sus relaciones.
<b>Herramientas de Clustering</b>	: Ver Algoritmos de <i>Clustering</i> .
<b>Heurística</b>	: Conjunto ordenado y finito de operaciones que son parte de un algoritmo de optimización y cuyo objeto es decidir o seleccionar el o los siguientes candidatos a evaluar.
<b>Hidrofilicidad</b>	: Propiedad de moléculas y/o sistemas que engloba un conjunto de interacciones físico-químicas, de carácter aparentemente opuesto a la hidrofobicidad, aunque se ha encontrado que no son propiedades excluyentes.
<b>Hidrofobicidad</b>	: Propiedad de moléculas y/o sistemas que engloba un conjunto de interacciones físico-químicas y que deriva de la segunda ley de la termodinámica.
<b>Híper Elipsoidal</b>	: Que se asemeja a una figura geométrica en el híper espacio semejante a una elipse en el plano cartesiano.
<b>Información Topológica</b>	: Información que nace de la relación entre objetos y/o datos.
<b>Matlab</b>	: Software matemático altamente especializado que contiene herramientas matemáticas, ingenieriles, estadísticas, entre otras, y que contiene un lenguaje de programación pre compilado.
<b>Meta Heurística</b>	: Método heurístico que incluye el uso de una función objetivo para resolver problemas.
<b>Paper</b>	: Trabajo publicado por un investigador en una revista científica reconocida.
<b>Propiedades Topológicas</b>	: Propiedades del espacio generado por un conjunto de objetos o datos, y sus interrelaciones o ubicaciones espaciales.
<b>Sistemas Cromatográficos</b>	: Conjunto de herramientas de separación de moléculas que utiliza una columna con una matriz para retener o ralentizar las moléculas en base a una propiedad específica.
<b>Topología</b>	: Espacio o geometría generada por un conjunto de objetos o datos, y sus interrelaciones o ubicaciones espaciales.

## Anexo B: ejemplos de ligandos en HIC

### Distintos Tipos de Ligando en Una Cromatografía de Interacción Hidrofóbica [29]

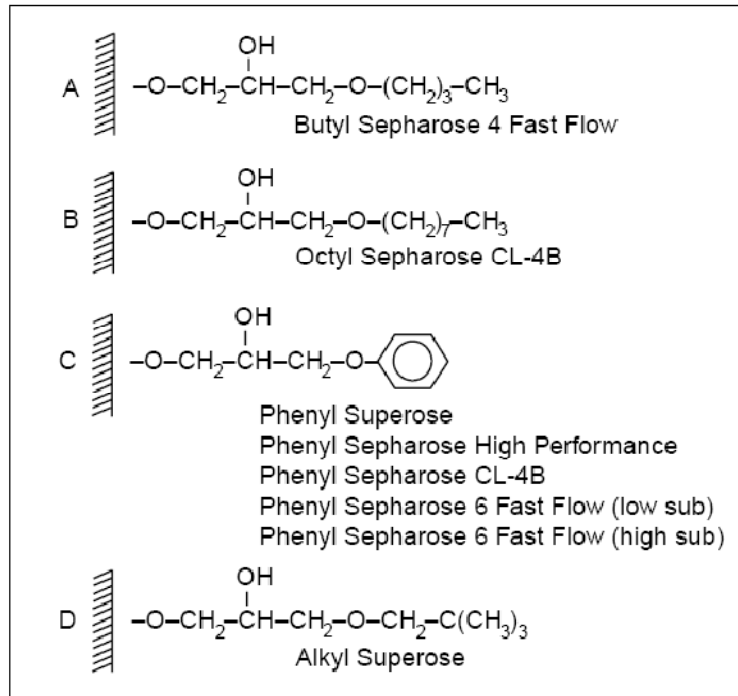


Figura 43: Distintos tipos de ligando en una cromatografía de interacción hidrofóbica [29].

## Anexo C: gráfico efecto de una sal en capacidad en HIC

Gráfico de la Relación Entre la Capacidad de Una Matriz en Función de la Concentración la Sal Sulfato de Amonio,  $(\text{NH}_4)_2\text{SO}_4$ , Para  $\alpha$ -chymotrypsinogen y  $\text{RNA}_{se}$  [29]

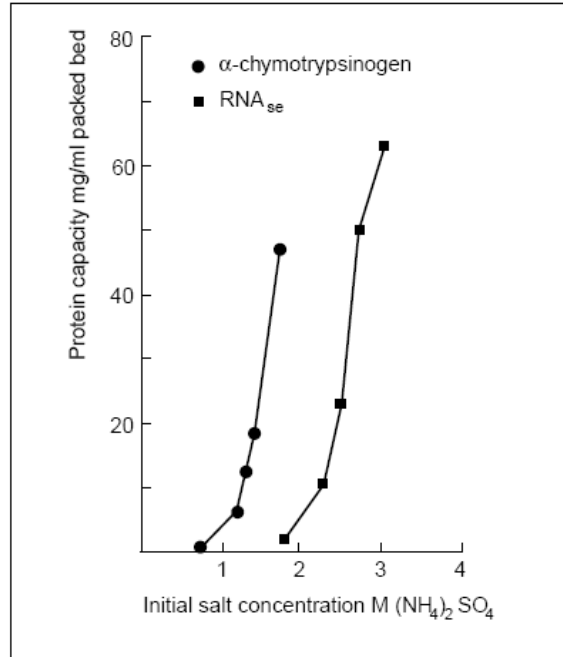


Figura 44: Gráfico de la relación entre la capacidad de una matriz en función de la concentración la sal sulfato de amonio,  $(\text{NH}_4)_2\text{SO}_4$ , para  $\alpha$ -chymotrypsinogen y  $\text{RNA}_{se}$  [29].



## Anexo D: series de Hofmeister [35, 36]

- A. Orden de los aniones y cationes según su contribución a la precipitación de proteínas (*salting out*) o al efecto coatrópico (*salting in*).

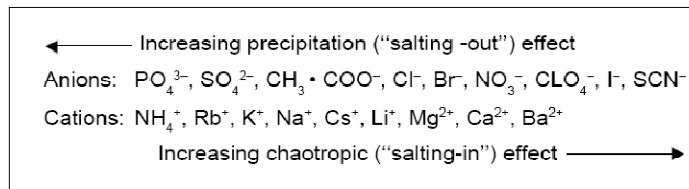


Figura 45: Diagrama de las series de Hofmeister – efecto *salting out* y *salting in*.

- B. Grupo de sales según su contribución a la tensión superficial molal del agua.

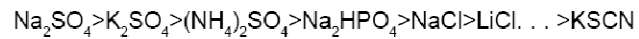


Figura 46: Diagrama de las series de Hofmeister – contribución de sales a la tensión superficial.

# Anexo E: ejemplo de información topológica

## Diagrama del Metro de Madrid: Ejemplo de Información Topológica [199]



Figura 47: Diagrama del Metro de Madrid: Ejemplo de Información Topológica [199].

## Anexo F: pseudocódigo GA

### Pseudocódigo de Algoritmos Evolutivos – *Genetic Algorithm*

1. Inicialmente se genera una población de individuos al azar.
2. Todos los individuos de la población son evaluados, lo cual puede incluir simulaciones y cálculos complejos.
3. Con el resultado de la evaluación, se ha determinado la utilidad de las distintas características de las posibles soluciones (individuos) y ahora se puede asignar un valor de adaptación a cada una de ellas.
4. Ahora se realiza un proceso de filtro estocástico, en el cual los individuos más adaptados tendrán mayor probabilidad de entrar al grupo que se puede reproducir.
5. En la etapa de reproducción se genera descendencia, creada variando o combinando las soluciones del punto 4, siendo integradas estas nuevas soluciones dentro de la población.
6. Si un criterio de término es alcanzado, la evolución para aquí. De lo contrario, se vuelve al paso 2.



Figura 48: Diagrama pseudo código algoritmos evolutivos.

## Anexo G: pseudocódigo GNG

### Pseudocódigo del Algoritmo *Growing Neuronal Gas*

0. Comenzar con dos nodos a y b en posiciones escogidas al azar  $w_a$  y  $w_b$  en  $\mathbb{R}^n$ .
1. Generar una señal  $\xi$  acorde a  $P(\xi)$ , lo que consiste en escoger un dato (APV) al azar.
2. Encontrar los dos nodos más cercanos a la señal:  $s_1$  y  $s_2$ .
3. Incrementar la edad de la conexión (arco) que parte desde  $s_1$ .
4. Sumar el cuadrado de la distancia entre la señal de entrada y el nodo más cercano a la variable contador local:  $\Delta \text{error}(s_1) = \|w_{s_1} - \xi\|^2$ .
5. Mover  $s_1$  y sus dos vecinos topológicos más cercanos conectados hacia  $\xi$  por la fracción  $\epsilon_b$  y  $\epsilon_n$ , respectivamente del total de la distancia:  
$$\Delta w_{s_1} = \epsilon_b (\xi - w_{s_1})$$
$$\Delta w_n = \epsilon_n (\xi - w_n)$$
 para todos los n vecinos directamente conectados de  $s_1$
6. Si  $s_1$  y  $s_2$  están conectados por un arco, establecer la edad de éste igual a 0. Si este arco no existe, crearlo.
7. Remover los arcos con una edad mayor  $a_{\max}$ . Si como resultado de esto existen nodos no conectados a ningún arco, remover éste.

## Anexo H: pseudocódigo GG

### Pseudocódigo del Algoritmo *Growing Grid*

Inicialmente la grilla consiste de cuatro nodos conectados a través de arcos (con valores asociados) en posiciones al azar.

0. Generar una señal  $\xi$  acorde a  $P(\xi)$ , lo que consiste en escoger un dato (APV) al azar.
1. Encontrar el nodo más cercano a la señal,  $s$  cuya posición está dada por  $w_s$ .
2. Incrementar la variable auxiliar asociada al nodo  $\tau_s$  en uno.
3. Adaptar la posición de los nodos de la grilla de tamaño  $k \times m$ , según la ecuación:

$$\Delta w_c = \varepsilon_0 \exp\left(-\frac{d^2(c, s)}{2\sigma^2}\right) (\xi - w_c) \quad (\forall c \in A)$$

Donde  $\varepsilon_0$  es una constante de la tasa de aprendizaje,  $\sigma$  es una constante asociada a la forma de la exponencial,  $A$  es el conjunto de nodos de la grilla ( $A = [a_{ij}]$ ), y  $d(c, s)$  es la distancia *city-block*, la que se define como:

$$d(c_1, c_2) = |i_1 - i_2| + |j_1 - j_2|$$

4. Después de  $k \cdot m \cdot \lambda_g$  procesos de adaptación (paso 3), insertar una fila o columna (según corresponda) entre  $q$  ( $\tau_q \geq \tau_c, \forall c \in A$ ) y su vecino más distante. Las posiciones de los nodos de la nueva fila o columna se obtiene mediante la interpolación de los vecinos de las filas o columnas a las que pertenecen los nodos  $q$  y  $c$ .
5. Luego se resetean las variables  $\tau_c, \forall c \in A$ .
6. El criterio de parada se alcanza cuando se llega al máximo número de nodos permitido, o cuando ningún nodo ha recibido más de  $1/(k \cdot m)$  del total de las señales  $\xi$  emitidas.

## Anexo I: DRT proteínas

### Tiempos de Retención Adimensional obtenidos por Lienqueo *et. al.* [54] para HIC

A continuación se muestran los tiempos de retención de 12 proteínas en una columna de 1 ml fenil-sefarosa de flujo rápido con 25  $\mu\text{mol/ml}$  de ligando. El gradiente utilizado es decreciente con sulfato de amonio 2 M como sal.

**Tabla 38: Tiempos de retención dimensional obtenidos por Lienqueo [54].**

N°	Proteína	PDB ID	DRT [-]
1	Citocromo C	1HRC	0,930
2	Ribonucleasa	1AFU	0,360
3	Mioglobina	1YMB	0,730
4	Conalbumina	1OVT	0,000
5	Ovoalbumina	1OOVA	0,567
6	Liozima	2LYM	0,500
7	Thaumatina	1THV	0,660
8	Qimotripsinogeno A	2CHA	0,370
9	$\beta$ -Lactoglobulina	1CJ5	0,690
10	$\alpha$ - Ailasa	1BLI	0,600
11	$\alpha$ - Qimostripsina	4CHA	0,770
12	$\alpha$ - Lactalbumina	1A4V	0,749

## Anexo J: coeficiente de partición de proteínas utilizadas

### Coeficientes de partición obtenidos por Schmidt [22, 24] para ATPS

A continuación se muestran los coeficientes de partición de 11 proteínas [22, 24] para cuatro sistemas de dos fases con polietilenglicol (PEG) 4000 más: fosfato, sulfato, citrato o dextrano. En cada caso se consideraron tres niveles de concentración de NaCl: sin sal (0,0% p/p), baja concentración de sal (0,6% p/p), y alta concentración de sal (8,8% p/p).

La solución de fosfato es una mezcla de  $K_2HPO_4$  y  $NaH_2PO_4$  a pH 7. En el resto de los sistemas se controló el pH con ácido cítrico (sistema citrato) e hidróxido de sodio (sistemas PEG+Sulfato y PEG+Dextrano). La temperatura durante la elaboración de los sistemas se mantuvo controlada a 20 [°C].

La distancia entre el punto que representa la composición del sistema y la curva binodal es idéntica para todos los sistemas, y el cociente de volúmenes entre la fase superior e inferior utilizado es 1. Las proteínas puras fueron agregadas a los sistemas de manera que la concentración final fuera siempre 1 [g/l] [24].

**Tabla 39: Coeficientes de partición obtenidos por Schmidt normalizados [22, 24].**

N°	Proteína	PDB ID	log K [-]											
			Fosfato			Sulfato			Citrato			Dextran		
			0,0%	0,6%	8,8%	0,0%	0,6%	8,8%	0,0%	0,6%	8,8%	0,0%	0,6%	8,8%
1	$\alpha$ -amilasa	1E40	0,77	0,68	0,32	0,62	0,82	0,45	0,81	0,82	0,55	0,54	0,58	0,44
2	$\alpha$ -chymotripsinogeno A	2CGA	1,00	1,00	0,66	1,00	0,20	0,29	1,00	1,00	0,90	-1,00	-0,97	-0,12
3	$\alpha$ -lactalbumina	1F6S	-1,00	-1,00	-0,73	-1,00	-1,00	-0,42	-0,99	-0,94	-0,30	0,21	0,12	-0,12
4	Amiloglucosidasa	3GLY	-0,48	-0,41	-0,87	0,77	0,45	-0,27	-0,49	-0,78	-0,72	1,00	0,38	0,85
5	Conalbumina	1OVT	-0,51	-0,40	-0,76	0,23	0,31	-1,00	-1,00	-1,00	-0,93	-0,36	-0,38	-0,32
6	Lisosima	2LYM	0,01	0,47	1,00	0,23	1,00	0,60	0,53	0,85	0,94	0,55	0,79	0,92
7	Ovalbumina	1OVA	-0,08	0,52	0,83	-0,13	0,65	0,70	0,50	0,96	0,95	0,74	0,67	0,95
8	Subtilisina	1SBC	-0,10	-0,14	0,85	0,66	0,48	1,00	0,16	0,32	1,00	0,77	0,97	0,76
9	Thaumatina	1THV	-0,48	0,06	0,62	0,30	0,90	0,48	0,03	0,61	0,78	0,71	1,00	1,00
10	Inhibidor de tripsina	1AVU	-0,45	-0,38	-0,76	0,10	0,66	-0,56	-0,29	-0,68	-0,69	-0,30	-0,85	-0,54
11	Suero de bovino	-	-0,51	-0,77	-1,00	0,64	-0,41	-0,60	-0,85	-0,58	-1,00	-0,58	-1,00	-1,00

## Anexo K: listado APVs, clasificación y referencia

**Tabla 40: Escalas de propiedades aminoacídicas utilizadas (APVs) con su numeración y clasificación. La clasificación 1 corresponde a escalas de origen conformacional, la 2 a escalas de hidrofobicidad y la 3 a escalas de origen estadístico.**

Nº Escala	Tipo Escala	Referencia escala	Clasificación
1	A.A.composition.	McCaldon P. Argos P. Proteins: Structure Function and Genetics 4: 99-122(1988).	1
2	A.A.Swiss-Prot.	Bairoch A. Release notes for Swiss-Prot release 41 - February 2003.	1
3	Accessibleresidues.	Janin J. Nature 277:491-492(1979).	1
4	Alpha-helixFasman.	Chou P.Y. Fasman G.D. Adv. Enzym. 47:45-148(1978).	1
5	Alpha-helixLevitt.	Levitt M. Biochemistry 17:4277-4285(1978).	1
6	Alpha-helixRoux.	Deleage G. Roux B. Protein Engineering 1:289-294(1987).	1
7	Antiparallelbeta-strand.	Lifson S. Sander C. Nature 282:109-111(1979).	1
8	Averageburied.	Rose G.D. Geselowitz A.R. Lesser G.J. Lee R.H. Zehfus M.H. Science 229:834-838(1985).	1
9	Averageflexibility.	Bhaskaran R. Ponnuswamy P.K. Int. J. Pept. Protein. Res. 32:242-255(1988).	1
10	Beta-sheetFasman.	Chou P.Y. Fasman G.D. Adv. Enzym. 47:45-148(1978).	1
11	Beta-sheetLevitt.	Levitt M. Biochemistry 17:4277-4285(1978).	1
12	Beta-sheetRoux.	Deleage G. Roux B. Protein Engineering 1:289-294(1987).	1
13	Beta-turnFasman.	Chou P.Y. Fasman G.D. Adv. Enzym. 47:45-148(1978).	1
14	Beta-turnLevitt.	Levitt M. Biochemistry 17:4277-4285(1978).	1
15	Beta-turnRoux.	Deleage G. Roux B. Protein Engineering 1:289-294(1987).	1
16	Bulkiness.	Zimmerman J.M. Eliezer N. Simha R. J. Theor. Biol. 21:170-201(1968).	1
17	Buriedresidues.	Janin J. Nature 277:491-492(1979).	1
18	CoilRoux.	Deleage G. Roux B. Protein Engineering 1:289-294(1987).	1
19	Hphob.Argos.	Rao M.J.K. Argos P. Biochim. Biophys. Acta 869:197-214(1986).	2
20	Hphob.Black.	Black S.D. Mould D.R. Anal. Biochem. 193:72-82(1991).	2
21	Hphob.Breese.	Bull H.B. Breese K. Arch. Biochem. Biophys. 161:665-670(1974).	2
22	Hphob.Chothia.	Chothia C. J. Mol. Biol. 105:1-14(1976).	2
23	Hphob.Doolittle.	Kyte J. Doolittle R.F. J. Mol. Biol. 157:105-132(1982).	2
24	Hphob.Eisenberg.	Eisenberg D. Schwarz E. Komarony M. Wall R. J. Mol. Biol. 179:125-142(1984).	2
25	Hphob.Fauchere.	Fauchere J.-L. Pliska V.E. Eur. J. Med. Chem. 18:369-375(1983).	2
26	Hphob.Guy.	Guy H.R. Biophys J. 47:61-70(1985).	2
27	Hphob.Janin.	Janin J. Nature 277:491-492(1979).	2
28	Hphob.Leo.	Abraham D.J. Leo A.J. Proteins: Structure Function and Genetics 2:130-152(1987).	2
29	Hphob.Manavalan.	Manavalan P. Ponnuswamy P.K. Nature 275:673-674(1978).	2
30	Hphob.Miyazawa.	Miyazawa S. Jernigen R.L. Macromolecules 18:534-552(1985).	2
31	Hphob.mobility.	Aboderin A.A. Int. J. Biochem. 2:537-544(1971).	2
32	Hphob.Parker.	Parker J.M.R. Guo D. Hodges R.S. Biochemistry 25:5425-5431(1986).	2
33	Hphob.pH3,4.	Cowan R. Whittaker R.G. Peptide Research 3:75-80(1990).	2
34	Hphob.pH7,5.	Cowan R. Whittaker R.G. Peptide Research 3:75-80(1990).	2
35	Hphob.Roseman.	Roseman M.A. J. Mol. Biol. 200:513-522(1988).	2
36	Hphob.Rose.	Rose G.D. Geselowitz A.R. Lesser G.J. Lee R.H. Zehfus M.H. Science 229:834-838(1985).	2
37	Hphob.Sweet.	Sweet R.M. Eisenberg D. J. Mol. Biol. 171:479-488(1983).	2
38	Hphob.Welling.	Welling G.W. Weijer W.J. Van der Zee R. Welling-Wester S. FEBS Lett. 188:215-218(1985).	2
39	Hphob.Wilson.	Wilson K.J. Honegger A. Stotzel R.P. Hughes G.J. Biochem. J. 199: 31-41(1981).	2



**Tabla n°39: Escalas de propiedades aminoacídicas utilizadas (APVs) con su numeración y clasificación. La clasificación 1 corresponde a escalas de origen conformacional, la 2 a escalas de hidrofobicidad y la 3 a escalas de origen estadístico (continuación).**

Nº Escala	Tipo Escala	Referencia escala	Clasificación
40	Hphob.Wolfenden.	Wolfenden R.V. Andersson L. Cullis P.M. Southgate C.C.F. Biochemistry 20:849-855(1981).	2
41	Hphob.Woods.	Hopp T.P. Woods K.R. Proc. Natl. Acad. Sci. U.S.A. 78:3824-3828(1981).	2
42	HPLC2,1.	Meek J.L. Proc. Natl. Acad. Sci. USA 77:1632-1636(1980).	2
43	HPLC7,4.	Meek J.L. Proc. Natl. Acad. Sci. USA 77:1632-1636(1980).	2
44	HPLCHFBA.	Browne C.A. Bennett H.P.J. Solomon S. Anal. Biochem. 124:201-208(1982).	2
45	HPLCTFA.	Browne C.A. Bennett H.P.J. Solomon S. Anal. Biochem. 124:201-208(1982).	2
46	Molecularweight.	Most textbooks.	1
47	Numbercodons.	Most textbooks.	1
48	Parallelbeta-strand.	Lifson S. Sander C. Nature 282:109-111(1979).	1
49	PolarityGrantham.	Grantham R. Science 185:862-864(1974).	2
50	PolarityZimmerman.	Zimmerman J.M. Eliezer N. Simha R. J. Theor. Biol. 21:170-201(1968).	2
51	Ratioside.	Grantham R. Science 185:862-864(1974).	1
52	Recognitionfactors.	Fraga S. Can. J. Chem. 60:2606-2610(1982).	1
53	Refractivity.	Jones. D.D. J. Theor. Biol. 50:167-184(1975).	1
54	Relativemutability.	Dayhoff M.O. Schwartz R.M. Orcutt B.C. In "" Vol.5 Suppl.3 (1978).	1
55	Totalbeta-strand.	Lifson S. Sander C. Nature 282:109-111(1979).	1
56	Hphop.JansonII.-aaescondidos	K-O Eriksson, in J-C Janson, L Ryden (Editors) Protein Purification*, 2 Edition, Wiley-Liss, New York, 1998, p 283.	2
57	Hphop.JansonI.-dGetanol_agua	K-O Eriksson, in J-C Janson, L Ryden (Editors) Protein Purification*, 2 Edition, Wiley-Liss, New York, 1998, p 283.	2
58	Hphop.berggren.ATPS	K. Berggren, A. Wolf, J.A. Asenjo, B.A. Andrews, F. Tjerneld , BBA-PROTEIN STRUCT M 1596 (2002) 253.	2
59	Hphop.jesior.4A	J-C. Jesior, Journal of Protein Chemistry 19 (2000) 93	2
60	Hphop.jesior.6A	J-C. Jesior, Journal of Protein Chemistry 19 (2000) 93	2
61	Hphop.jesior.8A	J-C. Jesior, Journal of Protein Chemistry 19 (2000) 93	2
62	Hphob.Leo.2	Abraham D.J. Leo A.J. Proteins: Structure Function and Genetics 2:130-152(1987).	2
63	Hphob.Roseman.2	Roseman M.A. J. Mol. Biol. 200:513-522(1988).	2
64	jonssonMultivariate-Parametrization_Z1	Jonsson J, Eriksson L., Hellberg S., Sjöström M., Wold S. Quant. Struct.-act relat 1989, 8, 204-209	3
65	jonssonMultivariate-Parametrization_Z2	Jonsson J, Eriksson L., Hellberg S., Sjöström M., Wold S. Quant. Struct.-act relat 1989, 8, 204-209	3
66	jonssonMultivariate-Parametrization_Z3	Jonsson J, Eriksson L., Hellberg S., Sjöström M., Wold S. Quant. Struct.-act relat 1989, 8, 204-209	3
67	HellbergPeptide-Cuantitativa_Z1	Hellberg S, Sjöström M, Skaberger B, Wold S, J Med Chem 1987, 30, 1126-1135	3
68	HellbergPeptide-Cuantitativa_Z2	Hellberg S, Sjöström M, Skaberger B, Wold S, J Med Chem 1987, 30, 1126-1135	3
69	HellbergPeptide-Cuantitativa_Z3	Hellberg S, Sjöström M, Skaberger B, Wold S, J Med Chem 1987, 30, 1126-1135	3
70	SandbergMultivariate-87aminoacids_Z1	Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S, J Med Chem 1998, 41, 2481-2491	3

**Tabla n°39: Escalas de propiedades aminoacídicas utilizadas (APVs) con su numeración y clasificación. La clasificación 1 corresponde a escalas de origen conformacional, la 2 a escalas de hidrofobicidad y la 3 a escalas de origen estadístico (continuación).**

<b>Nº Escala</b>	<b>Tipo Escala</b>	<b>Referencia escala</b>	<b>Clasificación</b>
71	SandbergMultivariate-87aminoacids_Z2	Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S, J Med Chem 1998, 41, 2481-2491	3
72	SandbergMultivariate-87aminoacids_Z3	Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S, J Med Chem 1998, 41, 2481-2491	3
73	SandbergMultivariate-87aminoacids_Z4	Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S, J Med Chem 1998, 41, 2481-2491	3
74	SandbergMultivariate-87aminoacids_Z5	Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S, J Med Chem 1998, 41, 2481-2491	3

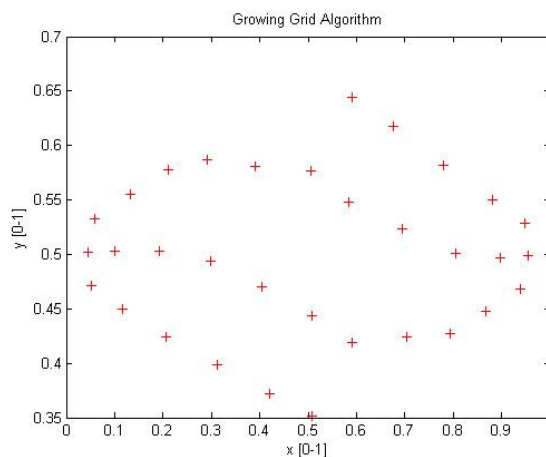
## Anexo L: ajuste visual GNG

### Pre ajuste Visual de los Parámetros – Growing Neuronal Gas

Se diseñaron dos conjuntos de datos bidimensionales (2D) para analizar el comportamiento de los algoritmos durante su ejecución. Adicionalmente se utilizaron las demos disponibles, en base a los cuales reportan los parámetros sugeridos [186]. Estos datos tienen dos ventajas sobre el uso de los APVs: permiten un análisis visual en tiempo real, y el tiempo de ejecución de los algoritmos sobre este conjunto de datos es considerablemente menor. Como resultado del análisis es posible obtener lo siguiente:

- Un rango de análisis a utilizar para cada parámetro;
- Una mayor comprensión sobre cómo afecta la variación de los distintos parámetros en la evolución del análisis topológico; y
- Una evaluación inicial de la multiplicidad de resultados;

En la Figura 49 se muestra, a modo de ejemplo, un gráfico de una solución obtenida con el algoritmo *Growing Grid* al ser utilizado para analizar un conjunto de datos que corresponden a una representación discreta de una circunferencia.



**Figura 49: Solución obtenida con el algoritmo *Growing Grid* al analizar un conjunto de datos que representan de forma discreta una circunferencia.**

Esta es una representación visual de una de las tres soluciones generadas por el Algoritmo *Growing Grid* con un conjunto de parámetros dados (los cuales se mantienen dentro de un cierto rango). Las otras soluciones son una circunferencia y la imagen especular de la Figura 49.

Dado que en la Figura 49 no se representan gráficamente las conexiones de la grilla, en la Figura 50 se presenta un esquema de otra de las soluciones obtenidas con este algoritmo, la cual incorpora las conexiones entre nodos.

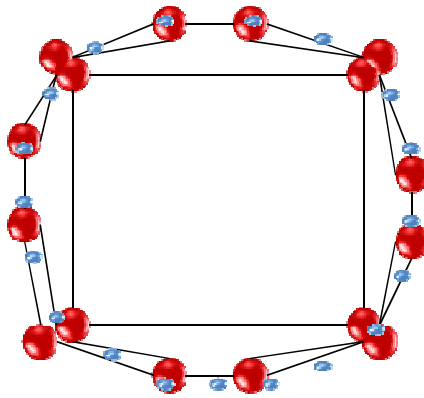


Figura 50: Representación de una solución obtenida por el algoritmo *Growing Grid*, mostrando los datos iniciales en azul, las neuronas en rojo y las conexiones de la grilla de neuronas con líneas negras.

Del análisis anterior, las conclusiones más trascendentes para el análisis de los 74 APVs son:

- El algoritmo *Growing Grid* (GG) presenta una menor cantidad de soluciones distintas que el algoritmo *Growing Neuronal Gas* (GNG), para un mismo conjunto de parámetros. Durante el pre-ajuste visual se presentaron 3 posibles soluciones con el algoritmo GG, e incontables soluciones con el algoritmo GNG
- El algoritmo GG presenta una menor cantidad de soluciones distintas que el algoritmo GNG, para los distintos conjuntos de parámetros. En particular, para el caso de los datos bidimensionales se mantienen las 3 posibles soluciones obtenidas con el algoritmo GG, además de la divergencia. En el caso del algoritmo GNG se aprecian muchas soluciones muy distintas entre sí.
- Para ambos algoritmos se observan soluciones no divergentes, que a la vista representan bien el conjunto de datos iniciales, con un número máximo de nodos mayor ( $n > m$ ) y menor ( $n < m$ ) al número de datos.

Cuando el número de nodos excede el número de datos a analizar ( $n > m$ ), se observan nodos en posiciones espaciales atribuibles a una síntesis de información, es decir, que representan a un conjunto numeroso de datos, por sobre uno o dos datos como ocurre en el caso contrario (un número de nodos menor al de datos).

Por ejemplo, en el análisis de datos bidimensionales que describen una esfera se observa en algunos casos un punto en el centro de ésta, cuya posición es idéntica a la del centroide de los datos, por lo que representa el grupo completo de datos. Cabe destacar que con los algoritmos utilizados el número de nodos no es un parámetro, y depende únicamente de la topología.

Al comparar el conjunto de datos a partir del cual se obtuvieron los parámetros sugeridos en la literatura, y el conjunto de datos bidimensionales generados para el análisis, no fue posible encontrar criterios de ajuste de estos parámetros para el análisis de los 74 APVs. Por lo que éste análisis no es un paso necesario dentro de la metodología utilizada en esta tesis.

## Anexo M: multiplicidad de resultados

### Multiplicidad de Resultados y Determinación del Número Óptimo de Ejecuciones para los algoritmos *Growing Neuronal Gas* y *Growing Grid*

#### *Multiplicidad de Resultados*

Según se vio en la sección 3.3, donde se discuten las características de los algoritmos de *clustering*, los algoritmos se pueden clasificar en determinísticos y estocásticos. Esta clasificación se basa en si el procedimiento principal de los algoritmos utiliza el azar, sin embargo, casi todos los algoritmos de *clustering* lo utilizan, por ejemplo, para determinar el punto de partida o solución inicial.

Si consideramos un algoritmo que no utiliza azar en ninguno de sus pasos (algoritmo  $i = A$ ) y uno que sí lo utiliza (algoritmo  $i = B$ ), al analizar un conjunto de datos fijos  $D$  con un conjunto de parámetros fijos  $P_i$  (que eventualmente puede ser distinto para cada algoritmo  $i$ ), el algoritmo  $A$  siempre obtendrá la misma solución  $Z_A$  al usar una misma solución inicial, sin importar cuántas veces se repita la ejecución del algoritmo. Por otro lado, el algoritmo  $B$  obtendrá un conjunto de soluciones  $Z_{Bj}$ . A esto se le denomina multiplicidad de resultados.

Los algoritmos de redes neuronales utilizados usan el azar en la función que genera la señal externa ( $\xi$ ) que reciben las neuronas, por lo que presentan multiplicidad de resultados. Además, el estado inicial (o vector inicial) se escoge al azar, lo cual es otro factor que contribuye a la multiplicidad de resultados, y por lo tanto, puede generar una mayor varianza sobre el conjunto de soluciones  $Z_{ij}$ .

Cuando se toman decisiones en base a mediciones con un nivel de incertidumbre elevado existe una probabilidad no despreciable de cometer errores. Por ejemplo, si dos muestras de grupos distintos de personas tienen alturas medias de  $\bar{x}_1 = 1$  y  $\bar{x}_2 = 2$  respectivamente, pero estas medidas se generan con tres mediciones cada una y ambas tienen una varianza de  $s^2 = 0,9$ ; el intervalo de confianza con un 95% de probabilidad donde se encuentran las medias de los dos grupos de personas son  $\bar{x}_1 \in [-0,07; 2,07]$  y  $\bar{x}_2 \in [0,93; 3,07]$ , rangos que tienen una intersección entre  $\bar{x}_1 \cap \bar{x}_2 \in [0,93; 2,07]$ , que corresponde a un 53% de cada intervalo de confianza para las medias muestrales. Es claro que no se puede tomar una decisión confiable con estas medias. Sin embargo, si aumentamos el número de mediciones a 12, y conservamos las medias y la varianza, el nuevo intervalo de intersección es de  $\bar{x}_1 \cap \bar{x}_2 \in [1,46; 1,54]$ , y corresponde a un 6,8% del intervalo de confianza para las nuevas medias muestrales ( $\bar{x}_1 \in [0,46; 1,54]$  y  $\bar{x}_2 \in [1,46; 2,54]$ ). En este caso es posible tomar una decisión con un nivel de confianza aceptable, lo cual se establece a través de un test de hipótesis para dos medias con igual varianza.<sup>4</sup>

Del ejemplo se concluye que se genera un sesgo importante si no se considera la varianza producto de la multiplicidad de resultados, y que ello puede llevar a un error

---

<sup>4</sup> En este ejemplo se utiliza la herramienta Intervalo de Confianza por considerarse más intuitiva para un lector sin conocimientos de estadística, sin embargo, la prueba de inferencia estadística más apropiada para determinar si dos medias son distintas o iguales es el test de hipótesis de dos medias (de igual varianza en este ejemplo).

sistemático en la toma de decisiones del análisis de sensibilidad, por lo que es necesario determinar un número adecuado de ejecuciones para cada conjunto de parámetros.

En la introducción de la sección 6.4.1 se discutió sobre las restricciones de cómputo existentes, lo cual obliga a ser muy discretos con la cantidad de cálculos a realizar. Por lo tanto, la solución al problema de la multiplicidad de resultados no pasa por escoger un número arbitrariamente grande de ejecuciones, sino que por encontrar un equilibrio entre la necesidad de reducir la variabilidad, y la necesidad de obtener resultados en un tiempo razonable.

### ***Definición Completa del Problema de Variabilidad***

La multiplicidad de resultados puede contribuir de forma muy importante a la varianza de las soluciones  $Z_{ij}$ , sin embargo, es importante considerar que la variabilidad del sistema está definida por:

- El algoritmo utilizado.
- El conjunto de parámetros utilizado (de los cuales depende el algoritmo).
- El tamaño de la muestra (cantidad de ejecuciones por conjunto de parámetros).
- Variaciones debido al método de medición.

Por lo tanto, no es posible hacer un estudio para determinar el número de ejecuciones adecuado para todos los algoritmos a utilizar, ni tampoco se puede tener certeza sobre la validez general del número de ejecuciones obtenido mediante un conjunto de parámetros fijos.

A continuación se explica para cada ítem cómo se aborda el tema de la variabilidad:

- **El algoritmo a utilizar:** para cada caso se requiere un estudio distinto si se detecta una varianza muy grande.
- **El conjunto de parámetros a utilizar:** como no se puede hacer un estudio independiente para cada conjunto de parámetros, por ser éstos demasiados, se escoge el conjunto de parámetros de mayor variabilidad detectada durante el ajuste preliminar de los parámetros y las primeras inspecciones con el análisis de sensibilidad. El número de ejecuciones resultante del análisis para este conjunto de parámetros se utiliza para cualquier conjunto de parámetros. En términos matemáticos, si  $S_i = \text{Max}\{S_{ij}\} \forall j$ , entonces se considera  $S_i = \text{Max}\{S_{ij^*}\}$ , para un conjunto de parámetros fijos  $j^*$  dentro del conjunto de parámetros analizados.
- **El tamaño de la muestra:** corresponde al conjunto de ejecuciones y, por lo tanto, es la variable más relevante a determinar para controlar la incertidumbre. Esta variable depende del resto de los parámetros y de la varianza deseada.
- **Variaciones debido al método de medición:** el método de medición con que se trabaja, y en base al cual se toman las decisiones, es el Error de Jackknife que se calcula utilizando un *software* [131]. Dentro del procedimiento de cálculo de los ajustes (heurística) el punto de partida se determina al azar, por lo que en la documentación correspondiente se advierte que los resultados pueden presentar variaciones.

Dejando todas las variables fijas, salvo el método de medición, se hizo un estudio sobre el aporte a la variabilidad dado por esta herramienta. La conclusión es que las variaciones debido al método, aunque existen, son despreciables (las variaciones son del orden de magnitud de  $10^{-20}$ ), lo cual era de esperar, dado que independiente del punto de partida, se analizan todos los casos posibles, y por lo tanto la variabilidad proviene de una aproximación numérica durante el cálculo.

Por lo tanto, y bajo las consideraciones expuestas en esta sección, se procede a aproximar el número de ejecuciones necesarias para tener un muestreo adecuado.

### ***Determinación del Número de Ejecuciones de un Algoritmo***

Se ha comentado sobre las dificultades que presenta la multiplicidad de resultados para que una heurística de optimización de los parámetros de los algoritmos pueda sugerir buenas decisiones. En esta sección se explica cómo se determina un número adecuado de ejecuciones para los algoritmos G.G. y G.N.G.

Para determinar un número adecuado de ejecuciones  $n$ , es necesario conocer las condiciones que requiere cumplir el conjunto de soluciones  $Z_{ij^*}$ . Dado que las  $j$  soluciones distintas son las responsables de la variación estándar, una condición a exigir es que el número de ejecuciones a utilizar permita una aproximación razonable de esta. Por lo tanto, el número de ejecuciones ( $n$ ) a exigir es el mínimo valor para el cual el promedio de las varianzas aproxime bien la varianza poblacional (desviación estándar),  $\bar{\sigma}_{ij^*} \approx \bar{S}_{inj^*}$ , considerando que la variación de la varianza entre distintos

grupos de  $n$  ejecuciones no supere un 10% ( $\sqrt{\sum(S_{inj^*} - \bar{S}_{inj^*})^2 / (S_{ij^*} \sqrt{n-1})} - 1 \leq 10\%$ ) (29).

### ***Determinación de una Aproximación de la Varianza Poblacional***

El cálculo de la varianza poblacional (desviación estándar) no se puede obtener mediante un muestreo (solución  $Z_{ij}$ ), a menos que el tamaño del número de ejecuciones tienda a infinito. Sin embargo, como se requiere obtener una aproximación de ésta, se obtuvo el máximo número de ejecuciones que permiten los recursos disponibles durante tres días de cómputo. A partir de esta muestra se determinó que existe una convergencia de la varianza en la medida que aumenta el número de ejecuciones consideradas en la muestra, y se determinó el número de ejecuciones para el cual se cumple el criterio descrito en la sección anterior.

Para demostrar que existe convergencia de la varianza se utilizó la siguiente metodología:

1. Sea un muestreo  $M$  de tamaño  $n$  dado por las soluciones de las primeras  $n$  ejecuciones de un algoritmo  $i$  para un conjunto de parámetros fijo  $j^*$ . Como se explicó anteriormente, se escoge un conjunto de parámetros tal que si  $S_i = \text{Max}\{S_{ij}\} \forall j$ , entonces  $S_i = \text{Max}\{S_{ij^*}\}$ , es decir, se escoge un conjunto de parámetros bajo el cual se aprecia una varianza especialmente grande durante la etapa de pre-ajuste visual y etapas tempranas del análisis de sensibilidad.

2. Sea  $m$  el valor máximo de  $n$  dado por la cantidad de ejecuciones, que se realizó durante los tres días de cómputo, y  $Z_{inj^*}$  la solución obtenida por la ejecución número  $n \in [1, m]$ .

3. La demostración de la convergencia se realiza entonces mediante la construcción del siguiente gráfico, y la apreciación visual de si existe o no convergencia: el eje de las abscisas corresponde al número de ejecuciones consideradas ( $n \in [2, m]$ ), mientras que el eje de las ordenadas corresponde al valor de la varianza promedio de dos mil grupos de soluciones de tamaño  $n$  formados al azar, obtenidas con las  $m$  soluciones individuales ( $Z_{inj^*}$ ). El valor en el eje de las ordenadas simula haber hecho para cada tamaño de muestra (grupo de  $n$  ejecuciones) dos mil repeticiones de cada muestreo de tamaño  $n$ . Al considerar que  $Z_{inj^*} \forall n \in [1, m]$  es un conjunto representativo de las posibles soluciones, y al utilizar la metodología explicada, una convergencia en este gráfico es significativa para determinar una buena aproximación de la varianza poblacional.

A modo de ejemplo, a continuación se muestra en la Figura 51 el gráfico obtenido para el caso del algoritmo GNG.

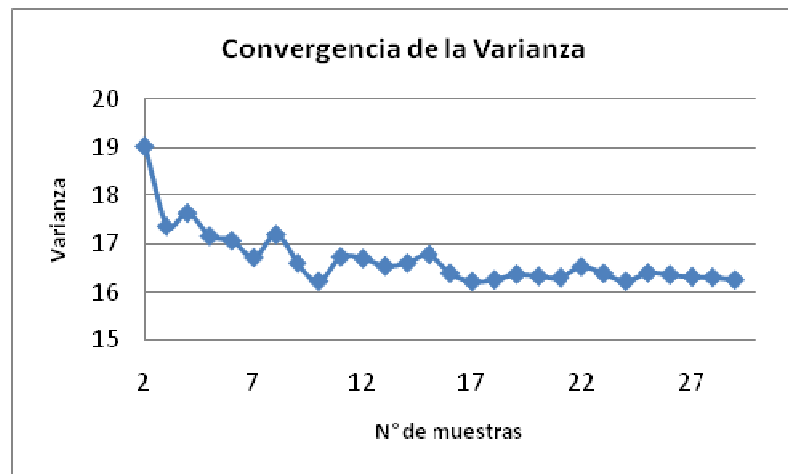


Figura 51: Convergencia de la varianza con el aumento del tamaño de la muestra o ejecuciones del algoritmo GNG para un set determinado de parámetros.

En la Figura 51 se aprecia una convergencia de la varianza al aumentar el número de ejecuciones, el cual se estabiliza entorno a 16,36. El hecho de que se aprecie una convergencia a partir de los muestreos de tamaño 16-17 implica que del muestreo inicial de tamaño  $m$ , utilizado como base para el análisis, después del 41% del total de ejecuciones el número de nuevas soluciones comienza a disminuir significativamente, y/o las nuevas soluciones obtenidas no aportan mayormente al aumento de la varianza. Esto en sí defiende la hipótesis inicial de que  $Z_{inj^*} \forall n \in [1, m]$  es un conjunto representativo de las posibles soluciones  $Z_{inj^*}$ .

Para apreciar la convergencia, en términos de la variación porcentual en torno al valor de convergencia, se presenta el siguiente gráfico.



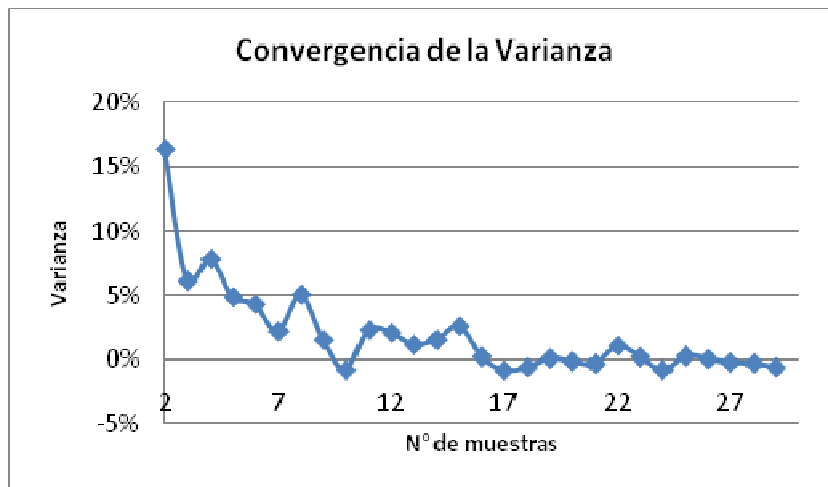


Figura 52: Convergencia de la varianza con el aumento del tamaño de la muestra o ejecuciones del algoritmo GNG para un set determinado de parámetros.

En la Figura 52 se aprecia que para que las mediciones del Error de Jackknife tengan una exactitud del 5% del valor de convergencia, se requieren de 5 o más mediciones (ejecuciones) para cada conjunto de parámetros, valor utilizado en el análisis de sensibilidad para el caso del algoritmo GNG.

De forma similar, se determinó para el caso del algoritmo GG es necesario tres ejecuciones por conjunto de parámetros. Para el resto de los algoritmos utilizados no se apreció una multiplicidad de soluciones significativa, por lo que se utilizó una sola ejecución por cada conjunto de parámetros.

## Anexo N: N° de Lisas APVs

### Aclaración del número de listas de APVs obtenidos a partir de las metodologías de identificación de APVs precursores de las mejores escalas generadas.

Las 78 listas de APVs generadas en la sección 6.5.3 se explican por qué:

- Hay 3 metodologías de obtención de los APVs precursores, descritos en la sección 6.5.1 y 6.5.2, las que se deben al uso de distintos tipos de algoritmos para generar nuevas escalas.
- Los 3 tipos de modelos de Salgado *et. al.* [13] (modelos lineales) y el de Lienqueo *et. al.* [54] (modelo 3D) son muy distintos en su dominio, razón por la cual se trabajaron separadamente. De esta manera el número de tipos de modelos utilizado es 2.
- Son 13 sistemas fisicoquímicos los utilizados, HIC y 4 ATPS, cada una a 3 concentraciones de sal diferentes.

En resumen,  $3 \cdot 2 \cdot 13 = 78$  listados de APVs.

## Anexo O: escalas mejores resultados ATPS

**Escalas generadas a partir de los cuales se obtiene los mejores modelos del coeficiente de partición en los 12 sistemas ATPS.**

**Tabla 41: Escalas generadas a partir de algoritmos de análisis topológico, utilizadas para la construcción de los mejores modelos tipo 3D para los doce sistemas ATPS**

N°	Amino-ácido	PEG-F			PEG-S			PEG-C			PEG-D		
		0,0%	0,6%	8,8%	0,0%	0,6%	8,8%	0,0%	0,6%	8,8%	0,0%	0,6%	8,8%
1	ALA	0,44	0,18	0,31	0,77	0,43	0,23	0,18	0,27	0,42	0,89	0,89	0,31
2	ARG	0,41	0,14	0,33	-0,59	0,27	-0,15	0,14	0,42	0,40	0,52	-0,38	0,33
3	ASN	0,49	0,16	0,42	-0,28	0,06	-0,73	0,16	0,40	0,28	-0,39	0,62	0,42
4	ASP	0,30	0,06	0,50	0,12	0,06	-0,38	0,06	0,31	0,09	-0,62	-0,24	0,50
5	CYS	0,51	0,34	0,68	-0,48	0,35	-0,21	0,34	0,27	0,61	-0,06	-0,13	0,68
6	GLN	0,51	0,15	0,43	-0,47	-0,02	-0,15	0,15	0,38	0,29	-0,29	-0,38	0,43
7	GLU	0,46	0,13	0,50	0,45	0,23	0,49	0,13	0,33	0,07	0,08	-0,61	0,50
8	GLY	0,39	0,21	0,24	-0,60	0,63	0,07	0,21	0,16	0,36	0,62	0,72	0,24
9	HIS	0,41	0,18	0,36	0,52	-0,07	0,84	0,18	0,38	0,68	-0,71	-0,66	0,36
10	ILE	0,78	0,38	0,20	-0,33	0,65	0,33	0,38	0,40	0,54	-0,45	0,84	0,20
11	LEU	0,65	0,26	0,26	0,67	0,46	0,47	0,26	0,41	0,58	0,36	-0,13	0,26
12	LYS	0,42	-0,02	0,32	0,62	-0,07	-0,71	-0,02	0,32	0,33	-0,23	0,56	0,32
13	MET	0,75	0,26	0,32	0,81	0,20	0,60	0,26	0,37	0,67	0,74	0,47	0,32
14	PHE	0,84	0,65	0,22	-0,33	0,22	0,73	0,65	0,66	0,71	-0,61	0,51	0,22
15	PRO	0,73	0,11	0,24	-0,42	0,05	-0,63	0,11	0,36	0,52	-0,38	-0,14	0,24
16	SER	0,48	0,09	0,42	-0,67	0,23	-0,51	0,09	0,34	0,28	-0,37	-0,18	0,42
17	THR	0,42	0,10	0,31	0,84	0,13	-0,46	0,10	0,34	0,29	-0,65	0,31	0,31
18	TRP	0,81	0,83	0,24	0,78	0,10	-0,38	0,83	0,80	0,97	0,45	-0,32	0,24
19	TYR	0,68	0,67	0,18	-0,20	0,12	0,69	0,67	0,69	0,62	-0,34	0,01	0,18
20	VAL	0,68	0,29	0,16	-0,58	0,50	-0,21	0,29	0,34	0,53	-0,22	-0,03	0,16

**Tabla 42: Escalas generadas a partir de algoritmos de análisis topológico, utilizadas para la construcción de los mejores modelos tipo lineal para los doce sistemas ATPS**

N°	Amino-ácido	PEG-F			PEG-S			PEG-C			PEG-D		
		0,0%	0,6%	8,8%	0,0%	0,6%	8,8%	0,0%	0,6%	8,8%	0,0%	0,6%	8,8%
1	ALA	0,02	-0,34	0,22	-0,08	0,34	0,38	-0,34	0,25	0,30	0,32	-0,12	0,23
2	ARG	0,39	0,27	0,36	0,46	-0,45	-0,23	0,27	-0,42	0,38	0,35	0,19	0,11
3	ASN	0,27	0,47	0,13	0,38	-0,26	-0,23	0,46	-0,40	0,19	0,19	0,30	-0,15
4	ASP	-0,29	0,20	-0,30	-0,31	-0,20	-0,31	0,20	-0,18	-0,27	-0,26	0,04	-0,17
5	CYS	0,59	0,60	-0,28	0,72	0,55	0,63	0,59	0,43	-0,13	-0,09	0,70	0,53
6	GLN	-0,13	-0,15	0,01	-0,16	0,20	-0,06	-0,15	0,58	0,06	0,06	-0,16	-0,12
7	GLU	-0,36	-0,25	-0,15	-0,43	-0,17	-0,26	-0,26	-0,63	-0,13	-0,12	-0,26	0,03
8	GLY	0,33	0,58	0,08	0,38	-0,61	0,22	0,58	-0,23	0,17	0,19	0,49	0,45
9	HIS	0,43	0,00	0,56	0,53	-0,31	0,30	-0,01	-0,62	0,62	0,62	0,04	-0,05
10	ILE	0,46	0,38	0,82	0,17	-0,37	0,42	0,39	0,10	0,86	0,87	0,80	0,82
11	LEU	0,08	-0,06	0,02	-0,24	-0,94	0,47	-0,06	0,45	0,01	0,05	0,34	0,46
12	LYS	-0,28	-0,15	-0,29	-0,35	0,51	-0,28	-0,16	0,05	-0,29	-0,30	-0,24	-0,25
13	MET	0,00	-0,17	0,72	-0,23	0,23	0,60	-0,17	0,84	0,74	0,76	0,17	0,39
14	PHE	0,32	-0,12	0,21	0,08	-0,77	0,15	-0,12	0,85	0,17	0,20	0,31	0,50
15	PRO	-0,17	0,38	0,14	-0,28	0,74	0,64	0,39	-0,36	0,13	0,17	0,36	-0,05
16	SER	-0,28	0,17	0,45	-0,34	-0,53	0,50	0,17	0,27	0,52	0,53	0,03	-0,07
17	THR	0,09	-0,08	-0,23	-0,04	0,55	0,67	-0,08	-0,40	-0,18	-0,17	-0,04	0,05
18	TRP	0,26	0,15	0,53	0,13	-0,51	0,86	0,15	-0,30	0,46	0,50	0,46	0,48
19	TYR	0,10	0,29	0,15	-0,11	-0,08	0,56	0,29	-0,39	0,09	0,11	0,49	0,36
20	VAL	0,27	-0,27	0,38	0,00	-0,53	0,10	-0,27	0,07	0,46	0,47	0,10	0,58

**Tabla 43: Escalas generadas a partir de algoritmos de optimización genética, utilizadas para la construcción de los mejores modelos tipo 3D para los doce sistemas ATPS**

N°	Amino-ácido	PEG-F			PEG-S			PEG-C			PEG-D		
		0,0%	0,6%	8,8%	0,0%	0,6%	8,8%	0,0%	0,6%	8,8%	0,0%	0,6%	8,8%
1	ALA	0,61	0,62	0,36	0,34	0,62	0,35	0,35	0,35	0,78	0,35	0,78	0,31
2	ARG	0,45	0,44	0,45	0,38	0,44	0,45	0,31	0,48	0,45	0,42	0,42	0,41
3	ASN	0,62	0,39	0,24	0,52	0,40	0,15	0,24	0,24	0,15	0,15	0,39	0,52
4	ASP	0,72	0,13	0,04	0,72	0,55	0,04	0,03	0,13	0,04	0,13	0,13	0,01
5	CYS	0,00	0,68	0,62	0,00	0,68	0,52	0,12	0,62	0,62	0,62	0,62	0,62
6	GLN	0,35	0,57	0,55	0,46	0,79	0,55	0,25	0,60	0,55	0,30	0,19	0,55
7	GLU	0,51	0,39	0,15	0,00	0,83	0,15	0,32	0,15	0,15	0,00	0,00	0,00
8	GLY	0,62	0,50	0,29	0,62	0,91	0,11	0,91	0,02	0,91	0,19	0,29	0,29
9	HIS	0,17	0,38	0,38	0,45	0,74	0,38	0,57	0,32	0,53	0,32	0,38	0,17
10	ILE	0,20	0,54	0,54	0,20	0,92	0,90	0,59	0,07	0,92	0,92	0,84	0,84
11	LEU	0,41	0,63	0,41	0,90	0,82	0,63	0,63	0,94	0,70	0,70	0,82	0,45
12	LYS	1,00	0,28	0,24	0,67	0,75	0,37	0,28	0,15	0,28	0,28	0,57	0,33
13	MET	0,01	0,74	0,50	0,06	0,74	0,50	0,68	0,11	0,51	0,53	0,90	0,90
14	PHE	0,16	1,00	0,58	0,12	0,12	0,76	0,76	0,60	0,81	0,64	0,58	0,16
15	PRO	0,00	0,71	0,14	1,00	0,00	0,14	0,71	0,71	0,14	0,44	0,44	1,00
16	SER	0,34	0,36	0,21	0,47	0,36	0,29	0,29	0,29	0,29	0,33	0,29	0,29
17	THR	0,66	0,31	0,32	0,50	0,62	0,42	0,27	0,28	0,35	0,29	0,32	0,28
18	TRP	0,05	1,00	0,59	0,05	0,59	1,00	0,88	0,54	0,75	0,14	0,49	0,19
19	TYR	0,43	0,88	0,95	0,45	0,95	0,95	0,83	0,72	0,83	0,57	0,83	0,71
20	VAL	0,38	0,09	0,52	0,09	0,52	0,83	1,00	0,82	1,00	1,00	1,00	0,86

**Tabla 44: Escalas generadas a partir de algoritmos de optimización genética, utilizadas para la construcción de los mejores modelos tipo lineal por los doce sistemas ATPS**

N°	Amino-ácido	PEG-F			PEG-S			PEG-C			PEG-D		
		0,0%	0,6%	8,8%	0,0%	0,6%	8,8%	0,0%	0,6%	8,8%	0,0%	0,6%	8,8%
1	ALA	0,25	0,81	0,36	0,08	0,44	0,17	0,81	0,44	0,99	0,22	0,35	0,62
2	ARG	0,96	0,41	0,45	0,17	0,96	0,41	0,46	0,44	0,42	0,35	0,42	0,46
3	ASN	0,58	0,15	0,37	1,00	0,67	0,82	0,40	0,40	0,16	0,16	0,24	1,00
4	ASP	0,88	0,47	0,04	0,91	0,49	0,01	0,45	0,47	0,47	0,03	0,03	0,03
5	CYS	0,21	0,55	0,14	0,66	0,21	0,45	0,62	0,66	0,62	0,62	0,80	0,68
6	GLN	0,79	0,19	0,79	0,24	0,33	0,16	0,30	0,47	0,66	0,30	0,25	0,33
7	GLU	0,64	0,56	0,15	0,00	0,56	0,25	0,56	1,00	1,00	0,11	0,04	0,25
8	GLY	1,00	0,33	1,00	0,63	1,00	0,82	0,00	0,00	0,29	0,19	0,29	1,00
9	HIS	0,18	0,52	0,08	0,18	0,13	0,50	0,74	0,46	0,38	0,32	0,38	0,38
10	ILE	0,67	0,95	0,67	0,16	0,56	0,65	0,99	0,54	0,90	0,95	0,94	0,94
11	LEU	0,11	0,68	0,29	0,68	0,68	0,27	0,99	0,99	0,68	0,99	0,94	0,70
12	LYS	0,37	0,67	0,50	0,37	0,57	0,37	0,67	0,75	0,37	0,17	0,28	0,24
13	MET	0,00	0,70	0,00	0,00	0,94	0,00	1,00	0,94	0,50	0,90	0,74	0,70
14	PHE	0,34	0,61	0,12	0,12	0,04	0,60	0,69	0,37	0,64	1,00	0,76	1,00
15	PRO	0,73	0,00	0,41	0,77	0,88	0,96	0,00	0,00	0,00	0,14	0,30	0,00
16	SER	0,61	0,47	0,32	0,61	0,88	0,21	0,33	0,90	0,29	0,36	0,36	0,29
17	THR	0,31	0,68	0,28	0,68	0,53	0,42	0,68	0,68	0,62	0,35	0,45	0,62
18	TRP	0,04	0,75	0,04	0,23	0,00	0,61	0,59	1,00	0,59	0,75	0,75	0,49
19	TYR	0,50	0,21	0,50	0,43	0,80	0,34	0,80	0,71	0,13	0,45	0,50	0,80
20	VAL	1,00	0,41	0,01	0,01	0,86	0,86	0,99	0,52	0,01	1,00	0,82	1,00

## Anexo P: parámetros iniciales y finales algoritmos

Listado de valores de los parámetros iniciales y finales utilizados con los algoritmos *Growing Neuronal Gas*, *Growing Grid* y *Bisection Algorithm*.

Tabla 45: Listado de valores de los parámetros iniciales y finales utilizados con los algoritmos *Growing Neuronal Gas*.

Parámetro	Valor Inicial	DRT	ATPS
Lambda	600	450	600
Max. Edge Age	88	89	89;470
Epsilon Winner	0.05	2.00E-04	1.0E-5;2.0E-5; 4.0E-5;8.0E-5;1.0E-4
Epsilon Neighbor	6.00E-04	0.3;0.7	0.01
Alpha	0.5	0.6	0.5
Beta	5.00E-04	5.00E-05	5.00E-06
Max. Nodes	100	100	80;60
Utility	3	3	3
Criterio Parada	-	5000;9000	5000;9000

Tabla 46: Listado de valores de los parámetros iniciales y finales utilizados con los algoritmos *Growing Grid*.

Parámetro	Valor Inicial	DRT	ATPS
Lambda_g	30	30	45
Lambda_f	100	100	100
Epsilon_i	0.1	0,1	0.04
Epsilon_f	0.005	0.005	4.0E-05
Sigma	0.9	0.4	0.6
Max. Nodes	100	74	45;320

Tabla 47: Listado de valores de los parámetros iniciales y finales utilizados con los algoritmos *Bisection Algorithm*.

Parámetros	Valor Inicial	DRT	ATPS
Crfun	l2	l2	l2
Simfun	Cosin	Cosin	Cosin
Colpurne	1	1	1
Clusters	10	45	45
Trials	10	10	10
Iterations	100	100	100

## Anexo Q: lista estudios mejores APVs

**Tabla 48: Identificación de *trabajos asociados a escalas* más relevantes para la generación de los mejores modelos.**

N° Escala	Nombre <i>Paper</i>	Referencia
25	Hydrophobic parameters $\pi$ of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides.	Fauchere J-L. Pliska V.E. Eur. J. Med. Chem. 18:369-375(1983).
32	New Hydrophobicity Scale Derived from High-Performance Liquid Chromatography Peptide Retention Data: Correlation of Predicted Surface Residues with Antigenicity and X-ray-Derived Accessible Sites.	Parker J.M.R. Guo D. Hodges R.S. Biochemistry 25:5425-5431(1986).
60 61	Hydrophilic Framework in Protein?	J-C. Jesior, Journal of Protein Chemistry, Vol 19, No. 2, 2000.
42	Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition.	Meek J.L. Proc. Natl. Acad. Sci. USA 77:1632-1636(1980).
22	The Nature of the Accesible and Bured Surfaces in Proteins.	Chothia C. J. Mol. Biol. 105:1-14(1976).
8	Hydrophobicity of Amino Acid Residues in Globular Proteins.	Rose G.D. Geselowitz A.R. Lesser G.J. Lee R.H. Zehfus M.H. Science 229:834-838(1985).
12	An algorithm for protein secondary structure prediction based on class prediction.	Deleage G. Roux B. Protein Engineering 1:289-294(1987).
7	Antiparallelbeta-strand	Lifson S. Sander C. Nature 282:109-111(1979).
10	Chou-Fasman Prediction of the Secondary Structure of Proteins*	Chou P.Y. Fasman G.D. Adv. Enzym. 47:45-148(1978).
11	Conformational Preferences of Amino Acids in Globular Proteins?	Levitt M. Biochemistry 17:4277-4285(1978).

\* Review del algoritmo Chou-Fasman revisado durante esta tesis, por no ser posible acceder al algoritmo original (Peter Prevelige, Jr., and Gerald D. Fasman).



## Anexo R: resumen estudios mejores APVs

Tabla 49: Resumen de trabajos asociados a escalas más relevantes para la generación de los mejores modelos.

N°	Escala (clasificación propuesta)	Resumen <i>Paper</i>	Consideraciones Relevantes / Comentarios
25	(Hidrofóbica Experimental)	Fouchere y Pliska [143] generan una escala de hidrofobicidad a través del cálculo de la constante $\pi$ (C. Hansch et al, 1979; Fauchere et al, 1930) en octanol/agua, utilizando la estructura aminoácidos N <sup>α</sup> -acetil-amino-acido amida para los 20 aminoácidos.	Se trabaja a pH 7-7.2. No se utilizan aminoácidos libres, y la estructura N-acetil-amino-acido amida no contiene cargas adicionales a las presentes en las cadenas laterales o residuos (-R). La constante $\pi$ se calcula a partir del coeficiente de partición P. El Parámetro $\pi$ eliminar el efecto borde al restar la contribución de la estructura en base a glicina.
32	(Hidrofóbica Experimental)	Parker et. al. [144] genera una escala de hidrofiliidad a través de una HPLC, y correlacionar los resultados con actividad antigena. Para esto se utilizó la estructura básica Ac-Gly-X-X-(Leu) <sub>3</sub> -Lys <sub>2</sub> -Amine, donde X es el aminoácido a ser estudiado. Se calcula el tiempo de retención y se resta el de la estructura Ac-Gly-(Leu) <sub>5</sub> -Lys <sub>2</sub> -Amine. Para eluir se utiliza NaClO <sub>4</sub> , y la escala final es transformada al intervalo [-10,10]. Finalmente utiliza los resultados para predecir las zonas superficiales con actividad antigena en proteínas.	Se estudia el efecto de los extremos a través del péptido Y-Gly-(Leu) <sub>5</sub> -(Lys) <sub>2</sub> -Z, donde Y es N <sup>α</sup> -acetyl o $\alpha$ -amine, y Z es C <sup>α</sup> -amide o Carboxyl. La Cromatografía utilizada es una columna C-18 a pH 7, y la elución con acetonitrilo (acetnitrile) se asocia con el parámetro hidrofiliidad. En realidad esta hidrofiliidad es calculada a través de una competencia hidrofóbica, por lo que la escala corresponde al tipo hidrofóbico.
60	(Estadística Estructural)	Jesior [145] estudia la estructura proteica utilizando como base la vecindad de cada aminoácido (de 0,1 a 20,0 Å con un paso de 0,1 Å) dentro de proteínas con estructura tridimensional conocida. La hipótesis del estudio es la existencia de una red hidrofílica y/o hidrofóbica en el ensamblaje de las proteínas. El cálculo de hidrofobicidad se desarrolla a partir de un proceso iterativo que utiliza la siguiente expresión:	Se utilizan 511 estructuras proteicas no redundantes. Utiliza solo la cadena lateral de las proteínas (-R), y no incluye los hidrógenos. El estudio considera la vecindad entre 4,0 a 8,0 Å. Bajo 3 Å en un 98,7% de las interacciones contienen residuos hidrofílicos y sobre 8 Å la composición de vecinos no presenta diferencias significativas, aunque la importancia relativa de residuos hidrofóbicos sigue en aumento.
61	(Estadística Estructural)	$n^{\circ} \text{ a. a. hidrofóbicos} / n^{\circ} \text{ a. a. hidrofílicos}$ . Independientemente del punto de partida, siempre la convergencia es al mismo resultado.	Se determinaron 41 escalas en el rango mencionado, al utilizar un las secciones entre dos esferas consecutivas al utilizar un paso de 0,1 Å.
42	(Hidrofóbica Experimental Estadística)	El trabajo de Meek [130] consiste en determinar una escala de hidrofobicidad a través del estudio del tiempo de retención de 25 péptidos en una <i>reverse-phase</i> HPLC. Se realizó una optimización del Coeficiente de Correlación entre los valores empíricos y teóricos de hidrofobicidad obtenidos para los 25 péptidos, variando los coeficientes de hidrofobicidad de cada aminoácido utilizando un análisis de sensibilidad como estrategia para examinar el dominio.	La cromatografía se desarrolla a pH 7,4 y 2,1. El uso de péptidos pequeños (~20 a.a.) evita distorsiones conformacionales y da validez al cálculo teórico de la hidrofobicidad. El cálculo de la hidrofobicidad de cada péptido es a través de la suma de las contribuciones individuales de cada aminoácido.

**Tabla 48: Resumen trabajos asociados a escalas más relevantes para la generación de los mejores modelos (continuación).**

N°	Escala (clasificación propuesta)	Resumen <i>Paper</i>	Consideraciones / Comentarios	Relevantes
22	(Estadística Estructural Con Enfoque Hidrofóbico)	Chothia [146] estudia el área superficial accesible de un grupo de 25 proteínas. Para un grupo de seis proteínas incluyendo la estructura primarias (residuos 100% expuesto al solvente), secundaria y terciaria o nativa (caso de las proteínas estudiadas). En esta transición se estudia la proporción de cada tipo de residuo cuya área deja de estar expuesta al solvente. La escala se obtiene considerando el estado global con un 95% de área superficial no expuesta al solvente.	Trabaja con proteínas globulares. Es un estudio estadístico estructural basado en una clasificación hidrofóbica. Concluye que la estructura secundaria se produce principalmente por enlaces no hidrofóbicos, y que en la formación de la estructura terciaria el papel de los residuos hidrofóbicos juega un papel más relevante. También se concluye que el área superficial perdida atribuible a residuos hidrofílicos es aproximadamente constante en relación al área en el estado primario, después de un cierto valor límite.	
8	(Estadístico Estructural Con Enfoque Hidrofóbico)	Rose et. al. [147] estudia el área superficial accesible de un grupo de 12 proteínas. Para cada una de éstas utiliza un estado estándar y la configuración nativa para estimar la pérdida de área. La escala de Rose se construye a través del siguiente cálculo: $f = 1 - A^0 - <A> / A^0$ , Donde $A^0$ es el área superficial en el estado estándar, y $<A>$ es el área superficial en el estado nativo (tridimensional).	El estado estándar utilizado son dos: - Gly-X-Gly; donde X es cada uno de los a.a.; $\varphi=140^\circ$ ; $\psi=135^\circ$ ; $\chi_1=120^\circ$ ; $\chi_2=180^\circ$ . - Los ángulos diédricos varían según probabilidades observadas en proteínas reales. Al graficar el área estándar en función de la pérdida de área se obtiene una clasificación de los aminoácidos en tres categorías con distinto carácter hidrofóbico. Se concluye que la superficie proteica de las estructuras nativas posee una cantidad aproximadamente constante de aminoácidos hidrofóbicos en relación al área en el estado estándar.	
12	(Estructural Estadístico)	El trabajo de Deléage et. al. [148] tiene como objetivo crear un algoritmo de predicción de la estructura secundaria, clasificando cada aminoácido de una proteína dentro de seis posibles estructuras: $\alpha$ , $\beta$ , $\alpha/\beta-\alpha$ , $\alpha/\beta-\beta$ , $\alpha+\beta$ , R. Previamente a la clasificación, para cada aminoácido se calcula el potencial de pertenecer a cuatro configuraciones distintas: $\alpha$ -hélice, <i>sheet</i> , <i>turn</i> y <i>coil</i> , este potencial da origen a la escala de hidrofobicidad.	Tiene un 72% de asertividad en la predicción de la clase y un 61,3 % de los residuos correctamente estructurados.	
10	(Estructural Estadístico)	El trabajo de Chou y Fasman [150] tiene como objetivo crear un algoritmo de predicción de la estructura secundaria. En este contexto la escala generada es el potencial de cada aminoácido de pertenecer a cuatro configuraciones distintas: $\alpha$ -hélice, $\beta$ -sheet, <i>turn</i> , y <i>coil</i> .	El trabajo de Chou y Fasman es idéntico al de Deléage, pero se realizó nueve años antes. Sin embargo, los resultados de Deléage son superiores en el nivel de predicción de estructura secundaria, y más completo (se incluyeron más aminoácidos y estructuras conocidas para el análisis).	

**Tabla 48: Resumen trabajos asociados a escalas más relevantes para la generación de los mejores modelos (continuación).**

N° (clasificación propuesta)	Escala	Resumen <i>Paper</i>	Consideraciones Comentarios	Relevantes	/
11 (Estructural Estadístico)		El estudio de Levitt et. al. [151] es una aplicación de trabajos que siguen la línea del de Chou y Fasman, pero es específico para proteínas globulares. En este caso los aminoácidos se clasifican en solo tres categorías: $\alpha$ -hélice, $\beta$ -sheet y <i>turn</i> . Un posterior análisis teórico sobre la clasificación automatizada permite una clasificación en 9 categorías.	Se concluye que: los aminoácidos voluminosos ( <i>bulky</i> ) se encuentran preferentemente en estructuras $\beta$ -sheet, los <i>polares de cadenas laterales (-R)</i> cortas se encuentran en <i>turns</i> , así como Gly y Pro; todos los demás aminoácidos se encuentran preferentemente en estructuras tipo $\alpha$ -hélice, excepto Arg que no posee una preferencia estadísticamente significativa por ninguna de las estructuras señaladas.		

## Anexo S: área superficial accesible de proteínas con DRT conocido

**Tabla 50: Área superficial accesible (ASA) de los aminoácidos hidrofóbicos y neutros (GLY), según la clasificación de Rose et. al. [145], para las proteínas con DRT conocido (ver anexo I).**

Proteína	ALA [Å <sup>2</sup> ]	ILE [Å <sup>2</sup> ]	LEU [Å <sup>2</sup> ]	CYS [Å <sup>2</sup> ]	MET [Å <sup>2</sup> ]	PHE [Å <sup>2</sup> ]	TRP [Å <sup>2</sup> ]	VAL [Å <sup>2</sup> ]	GLY [Å <sup>2</sup> ]	Promedio [Å <sup>2</sup> ]	Desviación Estándar [Å <sup>2</sup> ]	Desv./Prom. [-]
1HRC	143,5	389,1	848,8	72,1	91,7	116,7	45,7	83	211,8	222,5	256,9	115%
1AFU	409,9	15,8	76,1	54,8	50	32,6	0	282,6	194	124,0	141,3	114%
1YMB	478,1	355,9	485,9	144,8	100,7	97,5	113,5	194,5	107,5	230,9	163,7	71%
1OVT	203,8	215,1	69,1	60,6	63,1	144,9	14,3	141,1	470,5	153,6	137,1	89%
1OVA	878,3	314,9	501	18,6	217,4	131,1	57,9	681,3	446,9	360,8	291,9	81%
2LYM	509	528,5	939,1	192,3	298,1	326,8	190,6	829,3	1148,6	551,4	346,5	63%
1THV	414,1	99,2	239,8	192,1	0,8	221,8	56,9	204,6	619,3	227,6	189,5	83%
2CHA	568,2	190,4	312,8	0	6,8	136,5	19,4	141,7	668,3	227,1	244,6	108%
1CJ5	714	203,1	363,9	175,5	105,8	356,4	221,7	309,3	419,9	318,8	179,8	56%
1BLI	259,7	103,1	208,4	74,4	1,6	89,3	237,8	194,5	546,4	190,6	158,5	83%
4CHA	731	182,9	458,9	156	90,6	305,7	201,8	374	416,3	324,1	197,7	61%
1A4V	744,4	171,8	683,4	0	80,9	310,7	714,2	345,8	1071,3	458,1	361,0	79%
Promedio	504,5	230,8	432,3	95,1	92,3	189,2	156,2	315,1	526,7	251,9	-	-
Desviación Estándar	232,6	142,9	280,7	73,4	87,9	109,8	196,0	225,5	320,7	168,6	-	-
Desv./Prom.	46%	62%	65%	77%	95%	58%	125%	72%	61%	-	-	-

**Tabla 51: Área superficial accesible (ASA) de los aminoácidos medianamente polares, según la clasificación de Rose et. al. [145], para las proteínas con DRT conocido (ver anexo I).**

Proteína	HIS [Å <sup>2</sup> ]	SER [Å <sup>2</sup> ]	THR [Å <sup>2</sup> ]	TYR [Å <sup>2</sup> ]	Promedio [Å <sup>2</sup> ]	Desviación Estándar [Å <sup>2</sup> ]	Desv./Prom. [-]
1HRC	86,9	250,8	275,9	160,9	193,6	86,6	45%
1AFU	149,2	966,1	486,7	424,4	506,6	339,6	67%
1YMB	265,4	249,9	418,6	353,7	321,9	79,0	25%
1OVT	124,6	0	436,3	102,1	165,8	188,3	114%
1OVA	214	2272,9	712,3	195,2	848,6	979,3	115%
2LYM	388,8	1867,4	1478,2	377,2	1027,9	761,4	74%
1THV	0	607,9	911	745,5	566,1	397,2	70%
2CHA	676,1	185,2	384,4	50,7	324,1	271,8	84%
1CJ5	86,6	1453,6	1230	336,1	776,6	666,8	86%
1BLI	29,9	338,3	358,5	124,2	212,7	161,5	76%
4CHA	83,5	1451,7	1185	336,9	764,3	657,2	86%
1A4V	1033,3	1190,9	591,9	960,1	944,1	253,8	27%
Promedio	261,5	902,9	705,7	347,3	554,4	-	-
Desviación Estándar	306,3	743,8	400,9	269,2	310,9	-	-
Desv./Prom.	117%	82%	57%	78%	-	-	-

**Tabla 52: Área superficial accesible (ASA) de los aminoácidos polares, según la clasificación de Rose et. al. [145], para las proteínas con DRT conocido (ver anexo I).**

Proteína	ARG [Å <sup>2</sup> ]	ASN [Å <sup>2</sup> ]	ASP [Å <sup>2</sup> ]	GLN [Å <sup>2</sup> ]	GLU [Å <sup>2</sup> ]	LYS [Å <sup>2</sup> ]	PRO [Å <sup>2</sup> ]	Promedio [Å <sup>2</sup> ]	Desviación Estándar[Å <sup>2</sup> ]	Desv./Prom. [-]
1HRC	206,5	286,9	1023,8	697,9	820,3	1287,1	130,1	636,1	442,4	70%
1AFU	418,1	877,6	292,5	469	390,2	1171,6	250,3	552,8	341,4	62%
1YMB	221,4	410,7	795,9	815,8	1375,6	1358,5	482,3	780,0	452,4	58%
1OVT	80,8	328	192,8	406,4	1011,5	2153,9	215,3	627,0	738,7	118%
1OVA	1172,4	1074	1042	1146,6	2210,5	1686,2	817,6	1307,0	477,5	37%
2LYM	2733	2419,4	3339,3	1765,1	3631,9	4708,7	1390,4	2855,4	1141,0	40%
1THV	1408,1	380,1	792,1	380,3	558,4	1141,4	655,3	759,4	388,7	51%
2CHA	158,1	234,5	729,5	462,3	1216,1	1866,8	283,5	707,3	628,3	89%
1CJ5	353,8	924,9	497,1	752,4	396	1795	198,7	702,6	541,1	77%
1BLI	1636,6	1040,4	519,7	278,2	118,3	492,5	182,6	609,8	547,1	90%
4CHA	333,2	899,3	480,3	694,9	341,6	1762,1	164,2	667,9	541,4	81%
1A4V	1658,9	1445,7	1505,8	1173,1	1519,4	1776	635,3	1387,7	381,0	27%
Promedio	865,1	860,1	934,2	753,5	1132,5	1766,7	450,5	966,1	-	-
Desviación Estándar	842,7	624,8	839,0	428,1	991,5	1026,4	374,0	732,3	-	-
Desv./Prom.	97%	73%	90%	57%	88%	58%	83%	-	-	-

## Anexo T: hidrofobicidad de las proteínas con DRT conocido al modificar la escala de Wertz y Scheraga [178]

**Tabla 53: Hidrofobicidad de las proteínas utilizadas para HIC (ver proteínas en Anexo I), obtenidas a partir de las escalas generadas al variar el valor de hidrofobicidad de un aminoácido sobre la escala de Wertz y Scheraga [178]. La hidrofobicidad se calcula según la ecuación 15, y la modificación de la escala consiste en sumar al valor de hidrofobicidad, del aminoácido señalado, la media de la escala original (0,47). En esta tabla, un perfil de hidrofobicidad se entiende como los valores asociados a cada proteína para un aminoácido modificado.**

Proteína	Aminoácido con Valor de Hidrofobicidad Modificado																			
	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
1HRC	0,31	0,31	0,32	0,36	0,30	0,34	0,35	0,31	0,30	0,32	0,35	0,38	0,30	0,30	0,31	0,31	0,32	0,30	0,31	0,30
1AFU	0,28	0,28	0,31	0,27	0,25	0,28	0,28	0,26	0,26	0,25	0,26	0,33	0,25	0,25	0,27	0,31	0,28	0,25	0,28	0,27
1YMB	0,31	0,29	0,30	0,32	0,29	0,33	0,36	0,29	0,30	0,30	0,31	0,35	0,29	0,29	0,31	0,30	0,30	0,29	0,30	0,29
1OVT	0,21	0,20	0,22	0,21	0,20	0,22	0,27	0,23	0,20	0,21	0,20	0,35	0,20	0,21	0,21	0,19	0,23	0,20	0,20	0,21
1OVA	0,29	0,30	0,30	0,29	0,26	0,30	0,33	0,28	0,27	0,27	0,28	0,31	0,27	0,27	0,29	0,33	0,28	0,27	0,27	0,28
2LYM	0,24	0,28	0,27	0,29	0,24	0,26	0,29	0,25	0,24	0,24	0,25	0,31	0,24	0,24	0,26	0,26	0,26	0,24	0,24	0,25
1THV	0,30	0,34	0,29	0,31	0,29	0,29	0,30	0,31	0,28	0,28	0,29	0,33	0,28	0,29	0,31	0,31	0,32	0,28	0,31	0,29
2CHA	0,27	0,25	0,26	0,28	0,24	0,27	0,31	0,28	0,28	0,25	0,26	0,35	0,24	0,25	0,26	0,25	0,26	0,24	0,24	0,25
1CJ5	0,33	0,31	0,33	0,32	0,30	0,33	0,31	0,31	0,30	0,30	0,31	0,37	0,30	0,31	0,30	0,36	0,35	0,30	0,31	0,31
1BLI	0,32	0,41	0,37	0,34	0,31	0,32	0,31	0,34	0,31	0,31	0,32	0,34	0,30	0,31	0,32	0,33	0,33	0,32	0,31	0,32
4CHA	0,33	0,31	0,34	0,32	0,31	0,33	0,31	0,32	0,30	0,31	0,32	0,38	0,30	0,31	0,31	0,36	0,35	0,31	0,31	0,32
1A4V	0,33	0,36	0,35	0,35	0,32	0,35	0,36	0,34	0,34	0,32	0,33	0,36	0,32	0,32	0,33	0,35	0,33	0,33	0,34	0,32

## Anexo U: tablas de resultados obtenidas a partir de un análisis de Anova

**Tabla 54:** Tabla resultado del análisis de varianza (ANOVA) de la variación de la hidrofobicidad de las proteínas utilizadas para HIC, obtenida a partir de las escalas generadas al variar el valor de hidrofobicidad de un aminoácido sobre la escala de Wertz y Scheraga [178] (se utiliza la matriz traspuesta de la Tabla 54). La hipótesis nula en este caso es que la varianza de la hidrofobicidad no se explica a través de la separación de los datos por aminoácido (la variación de la hidrofobicidad es similar entre distintos aminoácidos).

<i>Origen de las variaciones</i>	<i>Suma de Cuadrados</i>	<i>Grados e Libertad</i>	<i>Promedio de los cuadrados</i>	<i>F</i>	<i>Probabilidad</i>	<i>Valor crítico para F</i>
Entre grupos	1,01507	19	0,0534	12,8027	3,2E-26	1,6340
Dentro de los grupos	0,91804	220	0,0041			
<b>Total</b>	<b>1,93312</b>	<b>239</b>				

**Tabla 55:** Tabla resultado del análisis de varianza (ANOVA) de la variación de la hidrofobicidad de las proteínas utilizadas para HIC, obtenida a partir de las escalas generadas al variar el valor de hidrofobicidad de un aminoácido sobre la escala de Wertz y Scheraga [178] (ver Tabla 54). La hipótesis nula en este caso es que la varianza de la hidrofobicidad no se explica a través de la separación de los datos por proteína (la variación de la hidrofobicidad es similar entre distintas proteínas).

<i>Origen de las variaciones</i>	<i>Suma de Cuadrados</i>	<i>Grados de Libertad</i>	<i>Promedio de los cuadrados</i>	<i>F</i>	<i>Probabilidad</i>	<i>Valor crítico para F</i>
Entre grupos	0,0368720	11	0,0033	0,4030	0,9538	1,8308
Dentro de los grupos	1,8962506	228	0,0083			
<b>Total</b>	<b>1,9331226</b>	<b>239</b>				

**Tabla 56:** Tabla resultado del análisis de varianza (ANOVA) de la variación de la hidrofobicidad de las proteínas utilizadas para HIC [ref], obtenida a partir de las escalas generadas al variar el valor de hidrofobicidad de un aminoácido sobre la escala de Wertz y Scheraga [178] (ver Tabla 54). La hipótesis nula en este caso es que la varianza de la hidrofobicidad no se explica a través de la separación de los datos por tipo de proteína, según el análisis sobre las áreas superficiales que se encuentra en Anexo Z.

<i>Origen de las variaciones</i>	<i>Suma de Cuadrados</i>	<i>Grados de Libertad</i>	<i>Promedio de los cuadrados</i>	<i>F</i>	<i>Probabilidad</i>	<i>Valor crítico para F</i>
Entre grupos	0,01379629	3	0,0045	0,5476	0,6502	2,6463
Dentro de los grupos	1,813869	216	0,0083			
<b>Total</b>	<b>1,82766529</b>	<b>219</b>				



## Anexo V: tablas de resultados obtenidas a partir de un análisis de Anova

Tabla 57: Variación de la hidrofobicidad de las proteínas utilizadas para HIC [ref], obtenida a partir de las escalas generadas al variar el valor de hidrofobicidad de un aminoácido sobre la escala de Wertz y Scheraga [178]. La modificación consiste en sumar al valor de hidrofobicidad, del aminoácido señalado, la media de la escala original (0,47). La hidrofobicidad se calcula según la ecuación 15, y esta tabla está ordenada para realizar un análisis de varianza (ANOVA) para evaluar la significancia estadística de las proteínas como grupos que permitan explicar la varianza de la hidrofobicidad.

Aminoácido Modificado	1BLI	2CHA	1OVT	1HRC	1A4V	1THV	4CHA	1CJ5	1YMB	2LYM	1OVA	1AFU
ALA	3,1%	10,8%	8,9%	7,6%	9,8%	3,5%	7,2%	13,2%	10,3%	5,8%	10,7%	6,2%
ARG	4,5%	11,1%	4,1%	3,0%	13,1%	18,6%	24,6%	3,7%	5,1%	36,8%	4,9%	13,9%
ASN	6,2%	23,2%	7,7%	12,2%	12,0%	16,5%	6,6%	5,4%	13,4%	23,4%	13,1%	12,1%
ASP	22,2%	7,7%	14,8%	7,2%	11,6%	22,7%	13,9%	16,9%	7,2%	11,7%	7,0%	12,6%
CYS	1,6%	1,4%	2,7%	2,2%	0,2%	1,3%	3,4%	0,0%	2,5%	1,7%	2,3%	0,0%
GLN	15,1%	12,4%	15,2%	15,1%	12,8%	12,0%	6,7%	10,7%	10,9%	6,2%	10,1%	9,8%
GLU	17,7%	10,3%	25,6%	37,5%	24,7%	24,7%	9,8%	28,2%	5,7%	2,7%	5,0%	12,7%
GLY	4,6%	5,1%	2,0%	17,5%	5,0%	7,8%	10,8%	15,5%	6,1%	12,3%	6,1%	9,0%
HIS	1,9%	3,9%	4,9%	4,6%	2,4%	2,6%	0,0%	15,7%	1,3%	0,7%	1,2%	8,7%
ILE	8,4%	0,4%	6,6%	8,0%	3,5%	3,6%	1,7%	4,4%	2,9%	2,3%	2,7%	1,4%
LEU	18,4%	2,0%	9,1%	2,6%	5,6%	6,4%	4,2%	7,3%	5,3%	4,7%	6,7%	5,7%
LYS	27,8%	31,0%	25,3%	79,9%	18,8%	32,1%	20,0%	43,3%	26,0%	11,1%	25,7%	14,9%
MET	2,0%	1,3%	1,9%	2,3%	2,4%	2,0%	0,0%	0,2%	1,5%	0,0%	1,3%	0,7%
PHE	2,5%	0,9%	1,8%	5,4%	1,5%	2,2%	3,9%	3,2%	5,2%	2,0%	4,5%	2,6%
PRO	2,8%	6,6%	9,0%	8,0%	9,1%	9,5%	11,5%	6,6%	2,9%	4,1%	2,4%	5,3%
SER	5,4%	25,6%	4,7%	0,0%	25,4%	12,7%	10,6%	4,3%	21,1%	7,6%	21,2%	10,0%
THR	6,0%	12,9%	7,8%	16,2%	7,9%	10,1%	15,9%	8,9%	17,8%	8,1%	17,3%	5,0%
TRP	1,0%	0,0%	2,1%	0,5%	0,6%	1,3%	1,0%	0,4%	3,2%	5,3%	2,9%	6,0%
TYR	3,5%	11,2%	6,6%	3,8%	2,2%	2,6%	13,0%	1,2%	4,9%	2,8%	4,9%	8,0%
VAL	1,8%	7,5%	3,6%	5,2%	7,6%	5,6%	3,6%	3,3%	4,5%	4,4%	5,5%	2,9%

## Anexo Y: tablas de resultados obtenidas a partir de un análisis de Anova

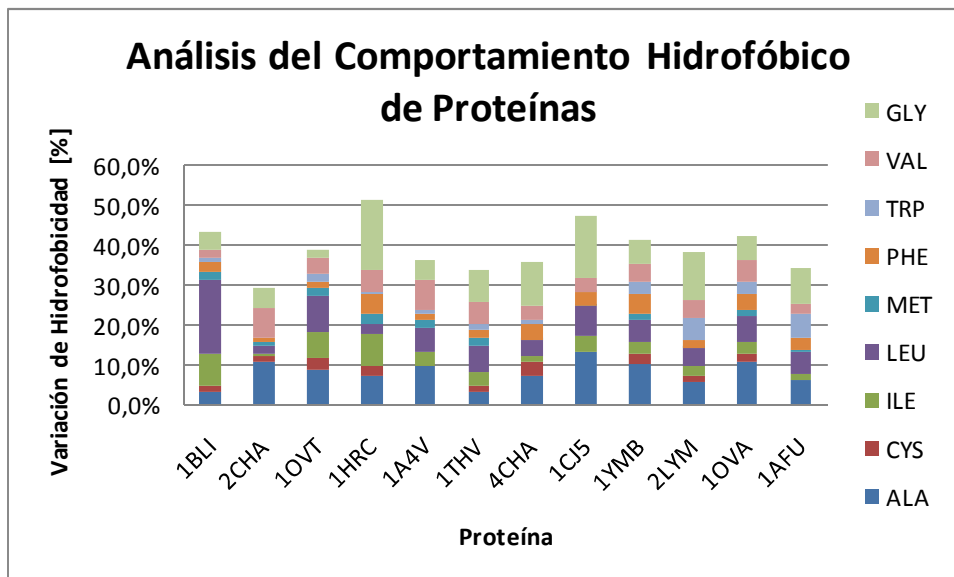


Figura 53: Variación de la hidrofobicidad de las proteínas utilizadas para generar los modelos del DRT en HIC, obtenida al modificar la hidrofobicidad de cada uno de los aminoácidos hidrofóbicos y neutros (según la clasificación de Rose et. al. [145]) por separado en un valor fijo e igual a la media de la escala original sobre la escala de Wertz y Scheraga [178].

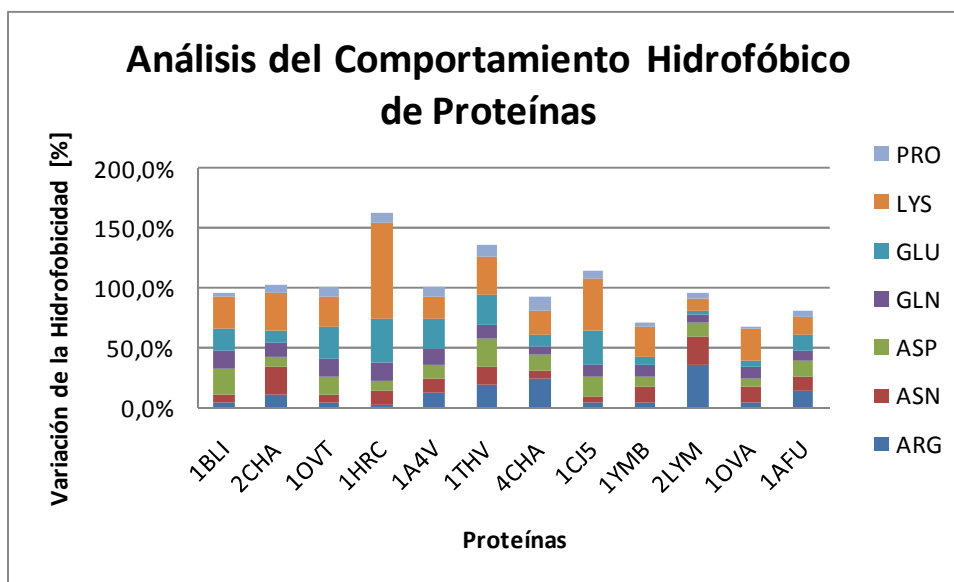


Figura 54: Variación de la hidrofobicidad de las proteínas utilizadas para generar los modelos del DRT en HIC, obtenida al modificar la hidrofobicidad de cada uno de los aminoácidos medianamente polares (según la clasificación de Rose et. al. [145]) por separado en un valor fijo e igual a la media de la escala original sobre la escala de Wertz y Scheraga [178].



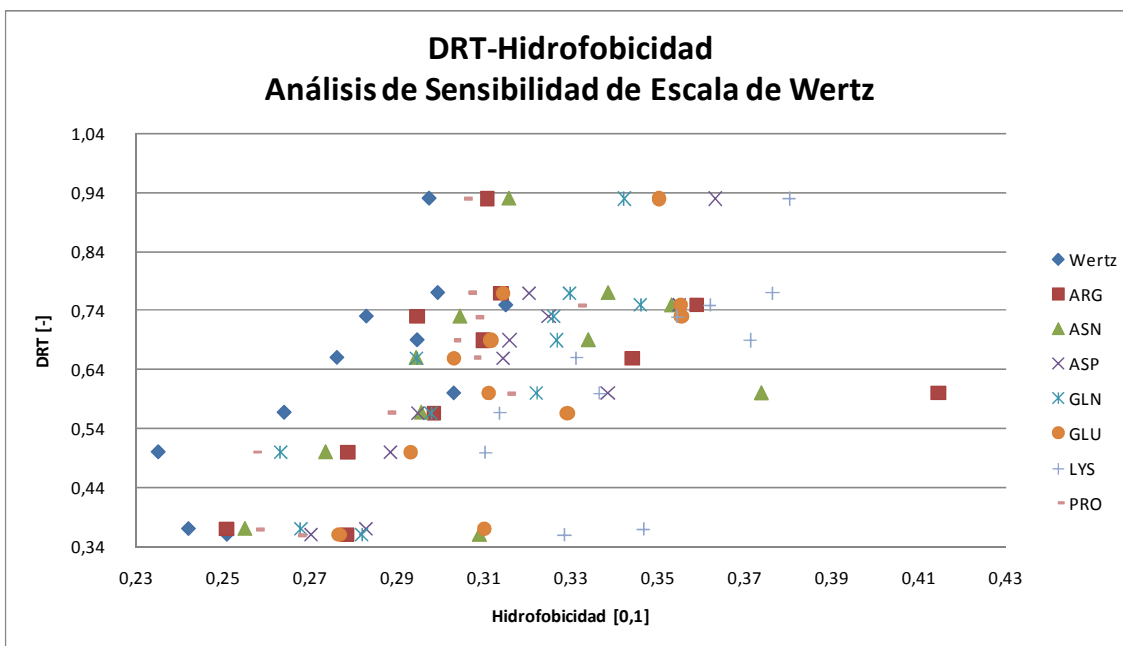


Figura 57: Gráfico del DRT-Hidrofobicidad de las proteínas utilizadas para generar los modelos del DRT en HIC, obtenida al modificar la hidrofobicidad de cada uno de los aminoácidos medianamente polares (según la clasificación de Rose et. al. [145]) por separado en un valor fijo e igual a la media de la escala original sobre la escala de Wertz y Scheraga [178].

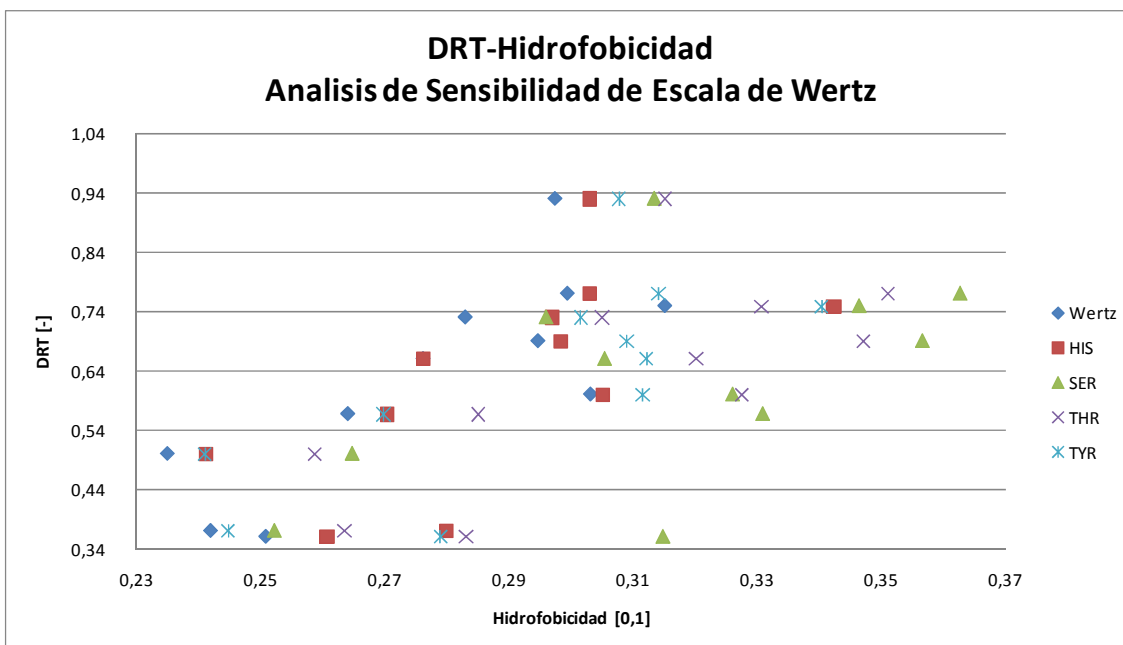


Figura 58: Gráfico del DRT-Hidrofobicidad de las proteínas utilizadas para generar los modelos del DRT en HIC, obtenida al modificar la hidrofobicidad de cada uno de los aminoácidos polares (según la clasificación de Rose et. al. [145]) por separado en un valor fijo e igual a la media de la escala original sobre la escala de Wertz y Scheraga [178].

## Anexo Z: estudio del área Superficial accesible (ASA) de las proteínas con DRT conocido

En esta sección se realiza un análisis con el objetivo de clasificar las proteínas utilizadas para el ajuste de los modelos de HIC, según su comportamiento hidrofóbico-polar esperable.

Según se discutió en la sección 7.3.2, existen tres tipos de aminoácidos: hidrofóbico, medianamente polares, y polares. Adicionalmente el aminoácido Gly podría actuar como un aminoácido neutro. Rose *et. al.* clasifica los aminoácidos de la siguiente manera [145]:

- Aminoácidos hidrofóbicos: (Gly), Ala, Cys, Val, Ile, Leu, Met, Phe y Trp.
- Aminoácidos medianamente polares: (Gly), Ser, Thr, His y Tyr.
- Aminoácidos polares: (Gly), Pro, Asp, Asn, Glu, Gln, Lys, y Arg.

Para clasificar las proteínas utilizadas para obtener los modelos de HIC, se analizó la contribución relativa de cada tipo de aminoácido al área superficial accesible (ASA) de cada proteína. Los valores del ASA fueron calculados previamente por Salgado *et. al.* utilizando el software STRIDE a partir de la estructura tridimensional de las proteínas [13, 137].

Del análisis del ASA, y utilizando la clasificación de Rose *et. al.* [145], es posible obtener el porcentaje de área superficial accesible de cada tipo (hidrofóbico, neutro, medianamente polar y polar) para cada proteína, según se muestra en la Tabla 58 a continuación.

**Tabla 58: Porcentaje del área superficial accesible según la clasificación propuesta por Rose *et. al.* [145], e incluyendo el aminoácido Gly como un aminoácido neutro.**

Proteína	Hidrofóbico	Neutro (Gly)	M. Polar	Polar
1HRC	39%	17%	15%	29%
1AFU	20%	16%	40%	25%
1YMB	38%	8%	23%	31%
1OVT	20%	39%	13%	28%
1OVA	27%	17%	30%	26%
2LYM	21%	25%	21%	33%
1THV	19%	32%	28%	21%
2CHA	21%	39%	18%	22%
1CJ5	29%	19%	34%	18%
1BLI	22%	39%	15%	24%
4CHA	30%	19%	34%	17%
1A4V	23%	30%	25%	21%

Luego, sumando los porcentajes de los aminoácidos hidrofóbicos y neutros (Gly) por un lado, y los medianamente polares y polares, por el otro, se obtiene el siguiente resultado:

**Tabla 59: Se muestran los porcentaje de ASA de cada proteína, según las clasificaciones: hidrofóbico-neutro y medianamente polar-polar.**

N° Proteína	Hidrofóbicos y Neutros	M. Polares y Polares	Clasificación
10	61%	39%	Proteínas Tipo 1
8	60%	40%	
4	59%	41%	
1	56%	44%	
12	53%	47%	Tipo 1-2
7	51%	49%	Proteínas Tipo 2
11	50%	50%	
9	49%	51%	
3	46%	54%	Proteínas Tipo 3
6	46%	54%	
5	44%	56%	
2	36%	64%	Proteína Tipo 4

En la Tabla 59 representa con colores los porcentajes del ASA según la clasificación propuesta. Cuando el color se acerca al azul, el comportamiento de una proteína se espera que sea más hidrofóbico, y mientras el color asociado se acerca al amarillo se espera un comportamiento más polar. En la Tabla 59 se distingue la existencia de 4 grupos de proteínas, para las cuales es razonable esperar un comportamiento similar en un gráfico del DRT-Hidrofobicidad. La variación de la hidrofobicidad a partir de cambios de escalas se estudia a continuación.

## Anexo AA: escala de Wertz y Scheraga

Tabla 60: Escala de hidrofobicidad de Wertz y Scheraga [178].

Nº	Aminoácido	Wertz
1	ALA	0,38
2	ARG	0,32
3	ASN	0,20
4	ASP	0,11
5	CYS	0,93
6	GLN	0,07
7	GLU	0,13
8	GLY	0,18
9	HIS	0,70
10	ILE	0,86
11	LEU	0,82
12	LYS	-
13	MET	0,80
14	PHE	1,00
15	PRO	0,07
16	SER	0,32
17	THR	0,13
18	TRP	0,98
19	TYR	0,59
20	VAL	0,73