Universidad Pablo de
Olavide de Sevilla

Predicción de Estructuras de Proteínas basada en Vecinos más Cercanos

MEMORIA QUE PRESENTA
Gualberto Asencio Cortés

PARA OPTAR AL GRADO DE DOCTOR POR LA
UNIVERSIDAD PABLO DE OLAVIDE DE SEVILLA

DIRECTOR
Jesús S. Aguilar Ruiz

Área de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior

Junio de 2013

Universidad Pablo de Olavide de Sevilla

Área de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior

Predicción de Estructuras de Proteínas basada en Vecinos más Cercanos

Tesis Doctoral

Gualberto Asencio Cortés

Sevilla, Junio de 2013

D. Jesús S. Aguilar Ruiz, profesor Titular de Universidad adscrito al Área de Lenguajes y Sistemas Informáticos de la Universidad Pablo de Olavide de Sevilla,

CERTIFICA QUE:

D. Gualberto Asencio Cortés, Ingeniero Informático por la Universidad de Sevilla, ha realizado bajo su supervisión el trabajo de investigación titulado:

PREDICCIÓN DE ESTRUCTURAS DE PROTEÍNAS BASADA EN VECINOS MÁS CERCANOS

Una vez revisado, autoriza la presentación del mismo como tesis doctoral en la Universidad Pablo de Olavide de Sevilla y estima oportuna su presentación al tribunal que habrá de valorarlo. Dicha tesis ha sido realizada dentro del programa de doctorado *Biotecnología y Tecnología Química,* de la Universidad Pablo de Olavide de Sevilla.

Sevilla, Junio de 2013.

A mi Anitilla

Resumen

Las proteínas son las biomoléculas que tienen mayor diversidad estructural y desempeñan multitud de importantes funciones en todos los organismos vivos. Sin embargo, en la formación de las proteínas se producen anomalías que provocan o facilitan el desarrollo de importantes enfermedades como el cáncer o el Alzheimer, siendo de vital importancia el diseño de fármacos que permitan evitar sus desastrosas consecuencias. En dicho diseño de fármacos se precisa disponer de modelos estructurales de proteínas que, pese a que su secuencia es conocida, en la mayoría de los casos su estructura aún se ignora. Es por ello que la predicción de la estructura de una proteína a partir de su secuencia de aminoácidos resulta clave para la cura de este tipo de enfermedades.

En la presente Tesis se ha analizado profundamente el estado del arte del problema de la predicción de la estructura terciaria y cuaternaria de una proteína, aportando diversos aspectos y puntos de vista de los métodos más actuales y relevantes presentes en la literatura. Por otra parte, se propone un método nuevo para la predicción de mapas de distancias que representan estructuras proteínicas mediante un esquema de vecinos más cercanos empleando propiedades físico-químicas de aminoácidos como entrada. Se ha realizado una exhaustiva experimentación y se han analizado los resultados desde varios puntos de vista y destacando diversos aspectos de interés. Finalmente, se ha aplicado la propuesta metodológica a dos grupos de proteínas de interés biológico: las proteínas de virus y de mitocondrias, obteniéndose resultados muy prometedores en ambos casos.

Índice general

Índice de figuras

Índice de tablas

Parte I

Introducción

Capítulo 1

Introducción

La presente Tesis Doctoral se encuentra enmarcada en el área de la bioinformática. La bioinformática es la ciencia de la investigación, desarrollo y aplicación de herramientas computacionales y aproximaciones para la expansión del uso de datos biológicos y médicos, incluyendo aquellas herramientas que sirven para adquirir, almacenar, organizar, analizar o visualizar tales datos. Existen dos principales áreas de estudio dentro de la bioinformática: la genómica y la proteómica. Mientras que la genómica se centra en el estudio del genoma de los seres vivos, la proteómica lo hace en el estudio del proteoma de los mismos. En concreto, la proteómica estudia las secuencias, estructuras y funciones de las proteínas desde que son sintetizadas a partir del ADN.

Dentro de la bioinformática existe una especialidad dedicada al estudio de las estructuras de las macromoléculas de origen biológico denominada bioinformática estructural [Gu and Bourne, 2003]. Los tres principales tipos de macromoléculas estudiadas en bioinformática estructural son el ácido desoxiribonucléico (ADN), el ácido ribonucléico (ARN) y las proteínas. Entre los diversos problemas que se estudian dentro de la bioinformática estructural aplicada a las proteínas se encuentra la predicción de estructuras de proteínas (PEP), la cual consiste en determinar la estructura de una proteína únicamente a partir de su secuencia.

El problema de la PEP comenzó a tratarse de solucionar aproximadamente en el año 1972 [Anfinsen, 1972], cuando Anfinsen probó que ciertas secuencias proteínicas producían siempre las mismas estructuras, siempre y cuando el entorno fuera el mismo. A partir de estos experimentos, se mantiene que toda la información que determina la estructura de una proteína se encuentra en su secuencia. Desde entonces se han sucedido multitud de diferentes aproximaciones computacionales que pretenden resolver este problema aun sin solución.

1.1. Motivación

1.1.1. ¿Qué son las proteínas y en qué consiste la PEP?

Las proteínas son las macromoléculas más versátiles y diversas que están presentes en todos los organismos vivos. Éstas adquieren estructuras complejas y desempeñan multitud de funciones. Una de las más habituales es la estructural, como la del colágeno, que es un complejo proteínico flexible con gran resistencia a la tracción, especialmente abundante en la piel y en los huesos de los mamíferos. Aparte de la función estructural, las proteínas realizan otras muchas, como la función enzimática (la pepsina o la sacarasa), la inmunológica (como los anticuerpos) o la transmisión de señales (como la rodopsina).

Las proteínas están formadas por una o varias cadenas de aminoácidos, que son pequeñas moléculas de las cuales existen tan sólo veinte tipos diferentes. Los aminoácidos están unidos unos con otros por un enlace químico covalente denominado enlace peptídico. En la figura 1.1 se representa la estructura de la oxi-hemoglobina, que es una proteína de cuatro cadenas de 141 y 146 aminoácidos cuya función principal es la del transporte de oxígeno a través de la sangre.

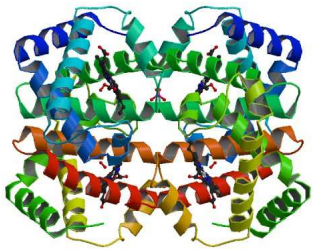


Figura 1.1: Estructura 3D de la proteína 1HHO correspondiente a la oxihemoglobina de humano.

Gracias a la sucesión concreta de aminoácidos de una cadena proteínica en un entorno fisiológico apropiado se obtiene, mediante su plegamiento, una estructura muy concreta que permite a la proteína desempeñar las funciones para las que ha sido diseñada a través de la evolución. Por este motivo, debido a que toda la información que conduce a la estructura de una proteína parece encontrarse en su secuencia [Anfinsen, 1972], surgen métodos que pretenden predecir la estructura de una proteína únicamente a partir de su secuencia de aminoácidos.

La motivación principal para el esfuerzo, tiempo y costes dedicados a la resolución del problema de la predicción de estructuras de proteínas es la

contribución a la cura de enfermedades, mediante el diseño de fármacos. A continuación, se explicará con mayor detalle a qué enfermedades nos referimos y cómo contribuye la PEP a su cura. Por otra parte, la PEP también es de gran importancia para la comprensión de las funciones de las proteínas, ya que son las estructuras proteínicas las que determinan sus funciones.

1.1.2. Enfermedades en las que la PEP es clave

El plegamiento de las proteínas, desde su secuencia hacia su estructura, no siempre produce un resultado satisfactorio. En ocasiones se generan estructuras incapaces de desempeñar sus funciones. Estos plegamientos incorrectos se deben, generalmente, bien a determinadas mutaciones en las secuencias proteínicas, o bien a cambios físico-químicos en el entorno fisiológico. Estas alteraciones modifican la capacidad de las proteínas para plegarse correctamente, afectando a la estabilidad de su conformación nativa. Estas alteraciones también hacen disminuir la cantidad de proteínas funcionales en la región del organismo donde deben actuar y resultan tóxicas para las células.

Estas estructuras ineficaces de proteínas tienden a unirse entre sí en un proceso llamado agregación, cuyo objetivo es el de unir y enterrar las regiones hidrofóbicas que quedan expuestas al solvente entre varias proteínas mal plegadas, formando lo que se conoce como un amiloide. Este es un proceso indeseado pues una vez formados los amiloides, las proteínas desnaturalizadas son irrecuperables.

Los organismos vivos actuales son capaces de detectar las alteraciones en los entornos fisiológicos y actuar en consecuencia, como por ejemplo en los choques térmicos, que consisten en la rotura de determinados materiales, vivos o no, debido a un cambio drástico en la temperatura. Ante tales circunstancias, la célula potencia la creación masiva de proteínas HSP, o proteínas de choque térmico, que proporcionan una respuesta al estrés fisiológico ocasionado. Dentro de este grupo de proteínas HSP, las chaperonas auxilian a las proteínas que se están plegando de forma incorrecta y hacen que se plieguen correctamente. Para ello, las chaperonas dirigen la ruta de plegamiento o paisaje energético hacia una estructura funcional, evitando la formación de agregados o amiloides.

No obstante, en ocasiones, la gran cantidad de proteínas mal plegadas supera lo que los mecanismos auxiliares de plegamiento pueden abordar, provocando determinadas enfermedades. Entre estas enfermedades, encontramos el Alzheimer, las encefa-lopatías espongiformes asociadas a priones, la enfermedad de Creutzfeldt-Jakob en humanos, la encefalopatía espongiforme bovina, la anemia falciforme (por mutaciones de origen genético en la hemoglobina que hemos mostrado anteriormente), la fibrosis cística, el síndrome de McKusick-Kaufman y Bardet-Biedl, el Parkinson

juvenil autosómico-recesivo, el escorbuto y el cáncer, entre otras. En concreto, con respecto al cáncer, existe una proteína denominada p53, supresora de tumores, cuyo objetivo es supervisar la integridad del material genético de la célula e impedir que posibles mutaciones alteren el funcionamiento de las células. Cuando esta proteína sufre a su vez mutaciones, su funcionalidad se pierde y, por ende, este mecanismo de defensa.

1.1.3. Diseño de fármacos

El objetivo es evitar que se produzcan plegamientos incorrectos de proteínas o reducir su número de ocurrencias. De esta forma, por tanto, se persigue impedir la aparición de las enfermedades que hemos citado. Aunque existen múltiples procedimientos para resolver este problema, en esencia todos tratan de conseguir alterar, inhibiendo o potenciando, alguna función molecular de una proteína o impedir la interacción entre varias, en unas o varias rutas de interés biomédico.

Según las características de cada problema, el procedimiento puede ser diferente. En el caso de perseguir la alteración de alguna función molecular de una proteína, lo habitual es atacar al centro activo de interés terapéutico de la proteína. Sin embargo, si el objetivo es impedir la interacción entre varias proteínas, lo habitual es romper las cavidades superficiales que sirven de puntos de unión entre las mismas.

En cualquier caso, generalmente se trata de encontrar al menos una molécula, denominada "molécula líder", de bajo peso molecular, a partir de la cual sea posible crear un compuesto que pueda administrarse como un fármaco. Esta molécula líder debe ser capaz, bien de unirse al centro activo de una proteína y alterar su función; o bien de romper la superficie de interacción de una proteína para impedir que ésta interaccione con otra.

La estrategia de diseño de fármacos utilizada para encontrar dicha molécula líder variará según el problema, aunque habitualmente se comienza por identificar la diana terapéutica a base de análisis funcional y detección de la interacción proteína-proteína. Una vez se ha identificado la diana terapéutica, es preciso caracterizarla, modelando la molécula receptora y localizando la superficie de interacción. Una vez localizada dicha superficie, se realizan modelos de la interacción proteína-proteína mediante simulaciones de docking [Kitchen et al., 2004]. Finalmente, se procede al desarrollo de la molécula líder, generando gigantescas baterías de pruebas a partir de bibliotecas virtuales de compuestos, mediante un proceso conocido como virtual screening [Sánchez-Linares et al., 2012].

Entre todos los pasos de una estrategia de diseño de fármacos existen fundamente dos de ellos en los que se precisa el modelo de la estructura de una proteína a partir de su secuencia de aminoácidos: en el modelado del receptor y en el modelado de la molécula líder. Ambos pasos son cruciales

para el diseño del fármaco y, por ende, para la cura de la enfermedad asociada.

No obstante, en la mayoría de los casos no se conocen las estructuras de las proteínas que se precisan. Esto es así debido al gran coste temporal y de recursos para la obtención experimental de la estructura de una proteína, frente a la relativa gran facilidad con la que se puede obtener su secuencia. De hecho, actualmente se conocen 29.266.939 secuencias [Consortium, 2012] y tan sólo 87.838 estructuras [Berman et al., 2000] están resueltas experimentalmente hasta la fecha. Es por este motivo por el que poder predecir la estructura de una proteína, sin conocerla previamente, es de tan crucial importancia.

1.2. Objetivos

Los objetivos principales de esta Tesis Doctoral son los siguientes:

- Identificar y analizar los datos de entrada disponibles derivados de las secuencias proteínicas y utilizados con mayor frecuencia en los métodos de PEP más relevantes publicados en el área. Identificar y analizar las estructuras de datos predichas más utilizadas, las cuales representan los modelos predichos de estructuras de proteínas. Identificar y analizar las aproximaciones algorítmicas más utilizadas y que mejor resultados han obtenido en el problema de la PEP.

- Desarrollar un método basado en el paradigma de los vecinos más cercanos para la predicción de estructura terciaria de proteínas. Los datos de entrada serán propiedades físico-químicas de aminoácidos y la información de salida será una matriz de distancias.

- Realizar un estudio de predicción con el método propuesto sobre diferentes conjuntos de proteínas, tanto de propósito general como de interés biológico, analizando en detalle el comportamiento predictivo y los resultados obtenidos.

1.3. Organización

En el capítulo 2 se introducen los elementos y nociones fundamentales de carácter biológico necesarios para comprender el problema de la PEP. Estas nociones están relacionadas con las secuencias y estructuras de las proteínas.

El capítulo 3 introduce el paradigma de los vecinos más cercanos, su metodología y la evaluación de los resultados de regresión y clasificación.

En el capítulo 4 se presenta el estado del arte referente a la predicción de estructuras terciarias de proteínas. Se incluye una revisión de los datos de entrada que se han utilizado en la literatura, los tipos de información de

salida y los métodos que existen, clasificados tanto por tipo de aproximación biológica como algorítmica.

El capítulo 5 describe nuestra propuesta de predicción de estructuras de proteínas basada en vecinos más cercanos, mapas de distancias y propiedades físico-químicas.

En el capítulo 6 se presentan los resultados obtenidos en la experimentación principal realizada con nuestro método de predicción. Esta experimentación se realiza sobre cinco conjuntos de proteínas de diferentes tamaños. Se incluyen gráficos que ilustran características relevantes de los datos de proteínas utilizados y diferentes análisis sobre las predicciones obtenidas.

El capítulo 7 muestra dos aplicaciones de nuestra propuesta a proteínas de interés biológico: proteínas de cápsides de virus y proteínas de matriz de mitocondrias. Se presentan los resultados obtenidos.

El capítulo 8 resume las conclusiones principales de la Tesis y los trabajos futuros.

Por último, se incluyen los apéndices, el glosario de términos, una lista de acrónimos y la bibliografía utilizada.

1.4. Contribuciones

Las contribuciones principales obtenidas a partir de resultados de esta Tesis se encuentran clasificadas en relación a la propuesta presentada:

1. La propuesta de predicción de estructuras de proteínas mediante matrices de distancias basada en vecinos más cercanos presentada en esta Tesis, denominada ASPpred, fue publicada en [Asencio-Cortés and Aguilar-Ruiz, 2011]. Se han publicado los siguientes trabajos sobre este sistema predictor:

    - [Asencio-Cortés et al., 2012] Asencio-Cortés, G., Aguilar-Ruiz, J. S., Chamorro, A. E. M., Ruiz, R., Toca, C. E. S. Prediction of Mitochondrial Matrix Protein Structures Based on Feature Selection and Fragment Assembly. In European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBio 2012) Lecture Notes in Computer Science, pp. 156–167, 2012.
    - [Asencio-Cortés et al., 2011b] Asencio-Cortés, G., Aguilar-Ruiz, J. S., Chamorro, A. E. M. A Nearest Neighbour-Based Approach for Viral Protein Structure Prediction. In European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBio 2011) Lecture Notes in Computer Science, pp. 69–76, 2011.

- [Asencio-Cortés and Aguilar-Ruiz, 2011]
  Asencio-Cortés, G., Aguilar-Ruiz, J. S. Predicting protein distance maps according to physicochemical properties. Journal of Integrative Bioinformatics, 8(3):181, 2011.
- [Asencio-Cortés et al., 2011c] Asencio-Cortés, G., Aguilar-Ruiz, J. S. and Chamorro, A. E. M. Prediction of Protein Distance Maps by Assembling Fragments According to Physicochemical Similarities. In 5th International Conference on Practical Applications of Computational Biology and Bioinformatics (PACBB 2011), Advances in Intelligent and Soft Computing, 93, pp. 271–277, 2011.
- [Asencio-Cortés et al., 2011a] Asencio-Cortés, G., Aguilar-Ruiz, J. S., Chamorro, A. E. M. Predicción de mapas de distancias de proteínas basada en vecinos más cercanos. In XIV Conferencia de la Asociacion Española para la Inteligencia Artificial (CAEPIA 2011), 2011.
- [Asencio-Cortés and Aguilar-Ruiz, 2010] Asencio-Cortés, G. and Aguilar-Ruiz, J. S. Importancia de las propiedades físico-químicas de los aminoácidos en la predicción de estructuras de proteínas usando vecinos más cercanos. In Actas del XV Congreso Español sobre Tecnologias y Logica Fuzzy (ESTYLF 2010), pp. 459–464, ISBN: 978-84-92944-02-6, 2010.
- [Asencio-Cortés and Aguilar-Ruiz, 2009] Asencio- Cortés, G., Aguilar- Ruiz, J. S. Predicción de estructuras de proteínas mediante vecinos mas cercanos usando características inherentes a los aminoácidos. In Actas de la XIII Conferencia de la Asociacion Española para la Inteligencia Artificial (CAEPIA 2009), 2009.

2. Otras contribuciones relacionadas con la predicción de estructuras de proteínas basadas en otros esquemas de *soft computing* fueron también publicadas:

   - [Márquez-Chamorro et al., 2012] Márquez-Chamorro, A. E., Asencio-Cortes, G., Divina, F. and Aguilar-Ruiz, J. S. Evolutionary decision rules for predicting protein contact maps. Pattern Analysis and Applications, 1–13, 2012.
   - [Márquez-Chamorro et al., 2012] Márquez-Chamorro, A. E., Divina, F., Aguilar-Ruiz, J. S., Bacardit, J., Asencio-Cortes, G. and Santiesteban- Toca, C. E. A NSGA-II algorithm for the residue-residue contact prediction. In Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBio

2012) number 7246 in Lecture Notes in Computer Science pp. 234–244. Springer, 2012.

- [Santiesteban-Toca et al., 2012] Santiesteban-Toca, C. E., Asencio-Cortés, G., Márquez-Chamorro, A. E. and Aguilar-Ruiz, J. S. Short-Range interactions and decision tree-based protein contact map predictor. In Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBio 2012) number 7246 in Lecture Notes in Computer Science pp. 224–233. Springer, 2012.
- [Márquez-Chamorro et al., 2011a] Márquez-Chamorro, A. E., Divina, F., Aguilar-Ruiz, J. S. and Asencio-Cortés, G. An evolutionary approach for protein contact map prediction. In Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBio 2011) number 6623 in Lecture Notes in Computer Science pp. 101–110. Springer, 2011.
- [Santiesteban-Toca et al., 2011] Santiesteban-Toca, C. E., Chamorro, A. E. M., Asencio-Cortés, G. and Aguilar-Ruiz, J. S. A decision tree-based method for protein contact map prediction. In Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBio 2011) number 6623 in Lecture Notes in Computer Science pp. 153–158. Springer, 2011.
- [Márquez-Chamorro et al., 2011] Márquez-Chamorro, A. E., Divina, F., Aguilar-Ruiz, J. S. and Asencio-Cortés, G. A multiobjective genetic algorithm for the Protein Structure Prediction. In 11th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 1086–1090, IEEE, 2011.
- [Márquez-Chamorro et al., 2011b] Márquez-Chamorro, A. E., Divina, F., Aguilar-Ruiz, J. S. and Asencio-Cortés, G. Residue-Residue Contact Prediction Based on Evolutionary Computation. In 5th International Conference on Practical Applications of Computational Biology and Bioinformatics (PACBB 2011) number 93 in Advances in Intelligent and Soft Computing pp. 279–283, Springer, 2011.
- [Márquez-Chamorro et al., 2010] Márquez-Chamorro, A. E., Divina, F., Ruiz, J. S. A. and Asencio-Cortés, G. Alpha helix prediction based on evolutionary computation. In Pattern Recognition in Bioinformatics number 6282 in Lecture Notes in Computer Science pp. 358–367. Springer, 2010.
- [Márquez-Chamorro et al., 2011] Márquez-Chamorro, A., Divina, F., Aguilar-Ruiz, J. and Asencio-Cortés, G. (2011). Un Algoritmo Genético para la Predicción de Mapas de Contacto Basado en Propiedades de Aminoácidos. In Actas de la XIV Conferencia de

la Asociación Española para la Inteligencia Artificial (CAEPIA 2011), 2011.

1.5. Resumen

En este capítulo se han introducido las características principales del contexto de la presente Tesis. Ésta se encuentra enmarcada dentro de la bioinformática estructural y la proteómica, y se centra en el problema de la predicción de estructuras de proteínas. Se han presentado la motivaciones principales para este problema. Se han descrito los tres objetivos principales de la Tesis, la organización del documento y, finalmente, las contribuciones realizadas.

Capítulo 2

Contexto biológico

El problema de la predicción de estructuras de proteínas se encuentra rodeado de una serie de conceptos clave de carácter biológico que es necesario describir. En este capítulo explicaremos en primer lugar la composición de las proteínas, sus aminoácidos y la síntesis. A continuación nos centraremos en las secuencias proteínicas y las bases de datos más importantes. Una vez explicados los conceptos clave de las secuencias, abordaremos las diferentes estructuras de proteínas, su clasificación y las bases de datos más relevantes. Finalmente, describiremos con más detalle el problema de la predicción de estructuras de proteínas.

2.1. Proteínas y aminoácidos

Las proteínas son macromoléculas compuestas por una o varias cadenas de elementos llamados aminoácidos. En la figura 2.1 se muestra el patrón molecular de un aminoácido. Como se puede apreciar en dicha figura, los aminoácidos están formados por un grupo amino (NH_2), un grupo carboxilo ($COOH$) y un carbono central llamado carbono alfa enlazado a un átomo de hidrógeno y a un grupo R. El grupo R, también llamado residuo o cadena lateral, es la única parte que varía de un aminoácido a otro. Existen tan sólo 20 aminoácidos naturales distintos y en la tabla 2.1 se muestran sus nombres y sus símbolos de una y tres letras.

Los aminoácidos están enlazados unos con otros para formar una secuencia de proteína. Este enlace se denomina enlace peptídico, es de tipo covalente y se forma entre el carbono del grupo carboxilo de un aminoácido y el nitrógeno del grupo amino de otro aminoácido, tal como se ilustra en la figura 2.2.

Se denomina esqueleto o *backbone* al conjunto de todos los átomos de una proteína salvo los de los grupos R de sus aminoácidos. A los ángulos ϕ y ψ que aparecen en la figura 2.2 se les denomina ángulos de torsión o ángulos dihédricos, los cuales determinan la estructura de la proteína.

Nombre	Símb.3	Símb.1	Nombre	Símb.3	Símb.1
Alanina	Ala	A	Leucina	Leu	L
Arginina	Arg	R	Lisina	Lys	K
Asparagina	Asn	N	Metionina	Met	M
Ácido aspártico	Asp	D	Fenilalanina	Phe	F
Cisteína	Cys	C	Prolina	Pro	P
Glutamina	Gln	Q	Serina	Ser	S
Ácido glutámico	Glu	E	Treonina	Thr	T
Glicina	Gly	G	Triptófano	Trp	W
Histidina	His	H	Tirosina	Tyr	Y
Isoleucina	Ile	I	Valina	Val	V

Tabla 2.1: Los 20 aminoácidos naturales y sus símbolos.

En la figura 2.3 se muestra la secuencia de una proteína real con todos sus átomos. Las regiones sombreadas en color verde corresponden a la parte fija del patrón molecular de cada aminoácido y conforman el esqueleto de la proteína. El resto de átomos de color blanco constituyen las cadenas laterales o residuos.

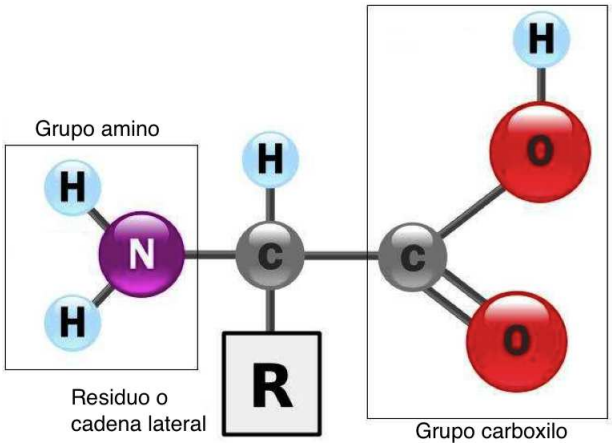


Figura 2.1: Patrón molecular de un aminoácido.

A partir de la secuencia de aminoácidos de una proteína se origina la estructura de la misma en el espacio, mediante un proceso dinámico de atracciones y repulsiones físico-químicas denominado plegamiento o *protein folding*. Cada secuencia de aminoácidos genera una única estructura tridimensional en el espacio, la cual otorga a la misma unas funciones concretas. La disposición espacial estable en entorno fisiológico se denomina conformación nativa de la proteína.

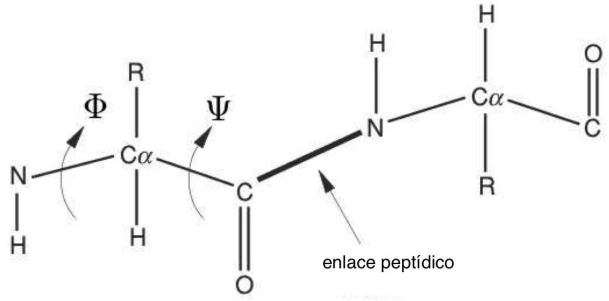


Figura 2.2: Enlace peptídico y ángulos de torsión.

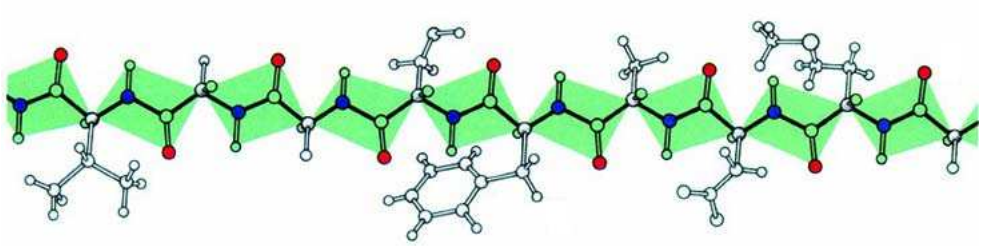


Figura 2.3: Secuencia de aminoácidos con todos sus átomos.

Una propiedad físico-química de un aminoácido es una característica que procede directa o indirectamente de la naturaleza del aminoácido, de su composición molecular. Cada aminoácido posee un valor concreto de cada propiedad físico-química. Se han identificado numerosas propiedades físico-químicas de aminoácidos. El mayor repositorio de propiedades de aminoácidos que se encuentra dispuesto actualmente para la comunidad científica es AAindex [Kawashima et al., 2008].

AAindex contiene fundamentalmente dos grupos de propiedades de aminoácidos: propiedades físico-químicas y propensiones estadísticas. Las primeras se derivan de la naturaleza molecular de los aminoácidos, las segundas se calculan en función del comportamiento de los mismos en diferentes entornos.

En la figura 2.4 se muestran las propiedades físico-químicas elementales de los aminoácidos en un diagrama de Venn. En dicho diagrama los aminoácidos son representados en su notación de símbolo de una letra. Como se puede apreciar en el diagrama, por ejemplo, los aminoácidos que tienen carga positiva son H, K y R, de los cuales H y K son hidrofóbicos (repelen el agua) y K es además aromático (contiene un anillo aromático). Mientras que el diagrama de Venn de la figura 2.4 asigna un valor lógico para cada propiedad a cada aminoácido, las propiedades del repositorio AAindex son

numéricas y, por tanto, aportan más información.

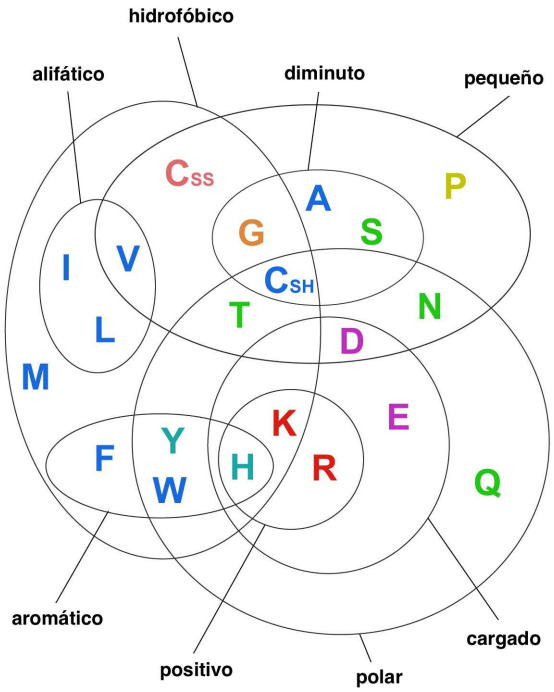


Figura 2.4: Propiedades físico-químicas elementales de los aminoácidos.

Las propensiones estadísticas reflejan el comportamiento observable más frecuente de cada aminoácido en un entorno determinado. Por ejemplo, la propiedad denominada propensión de pertenencia a hélice alfa refleja, para cada aminoácido, la frecuencia relativa de que un aminoácido forme parte de una hélice alfa dentro de una proteína. Veremos con mayor detalle las hélices alfa y otros motivos estructurales en el epígrafe 2.3.

2.2. Síntesis de proteínas

Una vez analizadas las características principales de las proteínas, nos centramos en el proceso de síntesis o producción de una proteína, el cual es llevado a cabo en la célula en varias fases. En la figura 2.5 se muestra esquemáticamente dicho proceso.

Tal como se ilustra en la figura 2.5, la síntesis comienza a partir de una secuencia de nucleótidos procedente del ADN (a) en el núcleo de una célula. A través de un proceso denominado transcripción (h), la parte de ADN que codifica una proteína es replicada en una hebra de ARN mensajero (b). Esta hebra de ARN (b) atraviesa el ribosoma (c), el cual es un complejo proteínico

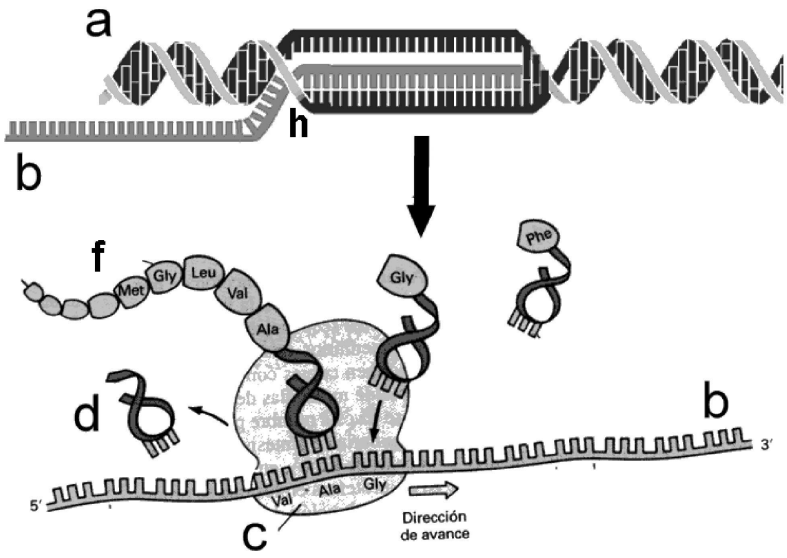


Figura 2.5: Síntesis de proteínas.

encargado de la síntesis de proteínas. Pequeñas hebras de ARN transferente (d) se enlazan en un extremo con tripletas de nucleótidos (c) para formar un aminoácido, el cual se encuentra sujeto en el otro extremo. Finalmente, los aminoácidos de las hebras de ARN transferente se unen mediante enlaces peptídicos formando la cadena de aminoácidos de la proteína (f).

2.3. Estructuras de proteínas

Una vez las proteínas son sintetizadas en el ribosoma, éstas adquieren, de forma espontánea en la mayoría de los casos, determinadas estructuras en el espacio dependiendo de su secuencia de aminoácidos y de las condiciones del entorno. Esta estructura concreta, su conformación nativa, le confiere la posibilidad de realizar determinadas funciones moleculares y participar en determinados procesos biológicos.

Una proteína puede presentar cuatro fases de formación de su estructura: primaria, secundaria, terciaria y cuaternaria. La estructura primaria consiste únicamente en su secuencia de aminoácidos, es decir, una secuencia lineal de aminoácidos unidos mediante enlaces peptídicos. La estructura secundaria es aquella en la que pueden haberse formado motivos estructurales, los cuales son regularidades estructurales en ciertas regiones de la proteína. Existen dos tipos de motivos estructurales: las hélices alfa y las láminas beta. En la figura 2.6 se representan ambos tipos.

Las hélices alfa son estructuras helicoidales dextrógiras, con unos 3,6

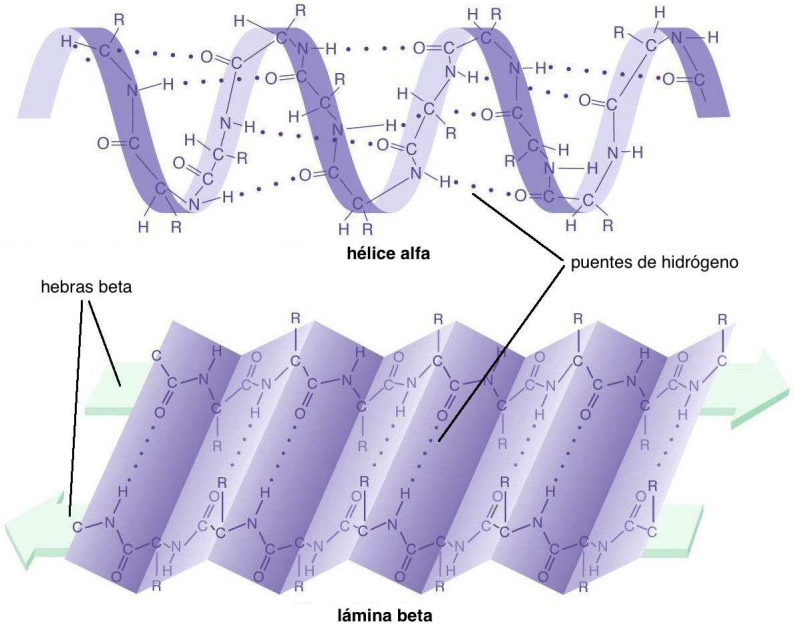


Figura 2.6: Motivos de estructura secundaria: hélice alfa y lámina beta.

aminoácidos por vuelta. Éstas se mantienen estables gracias a interacciones físicas llamadas puentes de hidrógeno (representados mediante líneas punteadas en la figura 2.6) entre el grupo amino de un aminoácido y el grupo carboxilo del aminoácido situado cuatro lugares después en la secuencia.

Las láminas beta se caracterizan por presentarse en forma aplanada y extendida. Al igual que las hélices alfa, los puentes de hidrógeno le confieren la estabilidad a la estructura. Las láminas beta constan de varias cadenas aminoacídicas, denominadas hebras beta, que permanecen enfrentadas y se mantienen juntas con puentes de hidrógeno en forma de zig-zag (líneas punteadas en la 2.6). La estructura laminar formada le confiere flexibilidad pero no elasticidad.

La estructura terciaria es la estructura completa tridimensional de una cadena de aminoácidos. Es la forma en que dicha cadena se pliega en el espacio. Esta forma queda determinada y estabilizada por medio de enlaces químicos y fuerzas físicas, tales como puentes de hidrógeno, fuerzas de Van der Vaals, enlaces iónicos, enlaces disulfuro, interacciones electrostáticas e hidrofóbicas.

Algunas proteínas se componen de dos o más secuencias de aminoácidos. En estos casos, la estructura cuaternaria queda descrita por la posición de todas sus secuencias en el espacio. Las fuerzas de estabilización que sostienen cada secuencia son las mismas fuerzas responsables de la estabilización de la

estructura terciaria. Un ejemplo de una proteína con estructura cuaternaria es la hemoglobina, ilustrada anteriormente en la figura 1.1.

2.3.1. Bases de datos de estructuras

La base de datos de estructuras de proteínas por excelencia es el Protein Data Bank (PDB) [Berman et al., 2000]. Se trata de un repositorio internacional para el procesamiento y distribución de estructuras macromoleculares 3D determinadas experimentalmente por cristalografía de Rayos X y resonancia magnético-nuclear. Todas las estructuras (públicas) de proteínas resueltas se encuentran en esta base de datos. Incluye tanto proteínas como ácidos nucléicos y otros complejos macromoleculares. Toda la información que se obtuvo en la determinación experimental de las estructuras de proteínas se encuentra almacenada en un fichero para cada proteína con un formato propio denominado formato PDB.

Un fichero en formato PDB se refiere a una única proteína y contiene varias secciones en las que se especifican las diferentes características de su estructura. La sección de estructura primaria viene indicada por la palabra reservada SEQRES y contiene una lista de símbolos de aminoácidos de tres letras que conforman la/s secuencia/s de aminoácidos de la proteína.

La sección de estructura secundaria está indicada por líneas con las palabras reservadas HELIX y SHEET y especifican los lugares de la proteína donde se encuentran las hélices alfa y las láminas beta, respectivamente. La sección de estructura terciaria (o cuaternaria para proteínas con varias secuencias) está indicada con la palabra reservada ATOM. Cada línea de texto que comienza con la palabra ATOM en un fichero PDB especifica las coordenadas (x, y, z) espaciales de un átomo de la proteína en su conformación nativa.

Habitualmente en los métodos de PEP se utiliza tan sólo un átomo por aminoácido, no todos los átomos de la proteína. El átomo de referencia comunmente utilizado en la literatura es el carbono beta. Este carbono beta es el carbono que pertenece al grupo R de un aminoácido y que está enlazado con el carbono central (ver figura 2.1).

2.4. Predicción de estructuras de proteínas

Como se ha explicado anteriormente, la PEP consiste en un proceso computacional que, a partir de la secuencia de aminoácidos de una proteína (estructura primaria), genera un modelo predicho de estructura para la misma. Esta estructura predicha puede ser secundaria, terciaria o cuaternaria, y según este tipo de estructura el método de PEP será de un tipo u otro. En esta Tesis se ha propuesto un método de predicción de estructura terciaria, el cual se explica con detalle en el capítulo 5.

Actualmente no existe ningún método que consiga predecir la estructura terciaria (o cuaternaria) con la suficiente precisión como para que el modelo generado tenga realmente utilidad. La precisión obtenida por las propuestas más relevantes y recientes de la literatura se encuentra en torno al 30 %, como veremos en el capítulo 4. Esta precisión debe ser mejorada para que los modelos de estructuras predichas puedan servir para la inferencia de sus funciones o para el diseño de fármacos, entre otros propósitos.

Con el objetivo de promover el desarrollo de nuevos métodos de predicción que mejoren los resultados existentes, la Universidad de California, patrocinada por el US National Institute of General Medical Sciences, lanzaron en 1994 la Critical Assessment of techniques for protein Structure Prediction (CASP) [Moult et al., 2011]. CASP es un campeonato bianual para la competición de métodos de predicción de estructuras de proteínas. Existen diferentes modalidades de competición, entre ellas la predicción de contactos entre residuos. Explicaremos con detalle este tipo de predicción en el epígrafe 4.3.

2.5. Resumen

En este capítulo se han resumido los conceptos básicos de carácter biológico que rodean al problema de la predicción de estructuras de proteínas. Se ha explicado la composición química de las proteínas y de sus aminoácidos y los distintos niveles de su estructura. Se ha indicado la base datos por excelencia de estructuras de proteínas y explicado el contenido de los ficheros de datos de proteínas. En último lugar, se ha puesto en relieve la baja precisión de los métodos de predicción actuales y la existencia del campeonato oficial de PEP.

Capítulo 3

Vecinos más cercanos

3.1. Introducción

En este capítulo definimos formalmente el paradigma de los vecinos más cercanos como técnica algorítmica para la clasificación o regresión de ejemplos de una base de datos en el marco de un aprendizaje automático supervisado. En primer lugar, definiremos los elementos necesarios para la formalización del algoritmo. A continuación describiremos el propio algoritmo y el procedimiento de evaluación de los resultados. Posteriormente, se justifica la elección de este tipo de esquema algorítmico para el método de PEP propuesto en esta Tesis.

3.2. Elementos

Definimos el conjunto A de n atributos como $\{a_1, \ldots, a_n\} \in \Re^n$. Sea R una matriz $\Re^{m \times n}$ denominada ejemplos de entrenamiento o training y el vector C denominado clases de entrenamiento, tal como se definen en la ecuación 3.1.

$$R = \begin{pmatrix} r_{1,1} & \cdots & r_{1,n} \\ \vdots & \ddots & \vdots \\ r_{m,1} & \cdots & r_{m,n} \end{pmatrix} C = \begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix} \tag{3.1}$$

Sea T una matriz $\Re^{s \times n}$ denominada ejemplos de test, el vector Q denominado clases de test y el vector P denominado predicciones de test, tal como se definen en la ecuación 3.2.

$$T = \begin{pmatrix} t_{1,1} & \cdots & t_{1,n} \\ \vdots & \ddots & \vdots \\ t_{s,1} & \cdots & t_{s,n} \end{pmatrix} Q = \begin{pmatrix} q_1 \\ \vdots \\ q_s \end{pmatrix} P = \begin{pmatrix} p_1 \\ \vdots \\ p_s \end{pmatrix} \tag{3.2}$$

Tanto C como Q y P pueden ser vectores de números reales o de valores discretos. Si lo son de números reales, entonces $C \in \Re^m$, $Q \in \Re^s$ y $P \in \Re^s$. Si son vectores de valores discretos, definimos $\{C_1, \ldots, C_d\}$ como su conjunto de valores discretos.

3.3. Método

El objetivo del algoritmo de vecinos más cercanos es producir valores para el vector P a partir de los ejemplos de test T, los ejemplos de entrenamiento R y las clases de entrenamiento C.

El procedimiento llevado a cabo por el algoritmo consiste en buscar, para cada fila i de la matriz T, las k filas de la matriz R cuyas distancias con t_i sean mínimas. La distancia entre dos vectores x e y es euclídea, se define en la ecuación 3.3 y es aplicable dado que R y T se encuentran en un espacio métrico.

$$distancia(x, y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2} \tag{3.3}$$

El algoritmo necesita un único parámetro k el cual indica el número de ejemplos (o vecinos) más cercanos a encontrar para cada ejemplo de test. Definimos en Algoritmo 1 dicho algoritmo.

Algoritmo 1 VECINOS MÁS CERCANOS

ENTRADA T, R, C, k
SALIDA P

```
inicio
  para i := 1 hasta s hacer
    clases := ∅, distancias := ∅
    para j := 1 hasta m hacer
      d := distancia(t_i,r_j)
      si d < máximo de distancias entonces
        si tamaño de distancias < k entonces
          h := Agregar d a distancias y devolver su posición
        si no
          h := Sustituir en distancias el elemento máximo por d y devolver su posición
        fin si
        Agregar c_j a clases en la posición h
      fin si
    fin para
    si clase discreta entonces
      p_i := moda de clases
    si no
      p_i := media aritmética de clases
    fin si
  fin para
  devolver P
fin
```

Como se puede apreciar, el algoritmo es cuadrático, en concreto de orden $O^{m \times s}$, pues debe calcular las distancias entre todos los elementos de R y T, sin excepción.

3.4. Evaluación

La evaluación consiste en determinar el grado de acierto de las predicciones realizadas por un algoritmo de predicción. Existen diversas métricas para cuantificar este grado de acierto según sea la naturaleza del problema de predicción. Para problemas de regresión (clase continua) las dos medidas más utilizadas son la raíz del error cuadrático medio y la raíz del error relativo cuadrático. Por el contrario, para problemas de clasificación (clase discreta) dos de las medidas más empleadas son la sensibilidad y la precisión.

En el problema de la PEP en particular, como veremos en el epígrafe 4.3, prácticamente todos los métodos resuelven un problema de clasificación (en concreto con clase binaria) al predecir estructuras terciarias o cuaternarias de proteínas. Por tanto, las medidas de evaluación que se utilizarán en el método que se propone en esta Tesis son las propias de un problema de clasificación.

A continuación describiremos ambos grupos de métricas en función del tipo de predicción (regresión o clasificación), usando los elementos formales descritos anteriormente. En concreto, se trata de cuantificar la diferencia entre el vector P de predicciones de test y el vector Q de clases de test.

3.4.1. Regresión

La raíz del error cuadrático medio (RMSE, *Root Mean Squared Error*) se define tal como se muestra en la ecuación 3.4.

$$RMSE(P,Q) = \sqrt{\frac{\sum_{i=1}^{s}(p_i - q_i)^2}{s}} \tag{3.4}$$

La medida RMSE tiene el inconveniente de la escala numérica de los valores de los atributos, ya que es una medida de error absoluta. Por el contrario, la raíz del error relativo cuadrático (RRSE, *Root Relative Squared Error*) es una medida relativa entre 0 y 1, tal como se define en la ecuación 3.5. $\overline{Q}$ es la media aritmética de los valores del vector Q.

$$RRSE(P,Q) = \sqrt{\frac{\sum_{i=1}^{s}(p_i - q_i)^2}{\sum_{i=1}^{s}(q_i - \overline{Q})^2}} \tag{3.5}$$

3.4.2. Clasificación

Definiremos las medidas de sensibilidad y precisión, además de la especificidad, la exactitud y el coeficiente de correlación de Mathews. No obstante, en primer lugar debemos definir cuatro medidas básicas (TP, TN, FP y FN), ya que las primeras están basadas en éstas. Todas estas 9 medidas se calcularían para cada valor C_x $(1 \leq x \leq d)$ de la clase discreta.

Por ejemplo, en el problema de la PEP se suele utilizar una clase que puede tomar 2 valores distintos (clase binaria), por tanto, se podrían calcular hasta $9 \times 2 = 18$ medidas. No obstante, en el problema que nos ocupa sólo tienen interés las medidas referentes a uno de los dos valores de la clase binaria, por lo que las medidas utilizadas en los capítulos 6 y 7 son sólo 5 (se omiten las cuatro medidas básicas).

Definimos las siguientes cuatro medidas básicas para un valor discreto C_x $(1 \leq x \leq d)$ de la clase:

- True positives (TP): Número de predicciones en las que $p_i = q_i \wedge q_i = C_x$ con $1 \leq i \leq s$.
- True negatives (TN): Número de predicciones en las que $p_i = q_i \wedge q_i \neq C_x$ con $1 \leq i \leq s$.
- False positives (FP): Número de predicciones en las que $p_i \neq q_i \wedge p_i = C_x$ con $1 \leq i \leq s$.
- False negatives (FN): Número de predicciones en las que $p_i \neq q_i \wedge q_i = C_x$ con $1 \leq i \leq s$.

A continuación definimos las medidas de sensibilidad (también llamada recall o coverage), precisión (también llamada accuracy), especificidad, exactitud y coeficiente de correlación de Mathews (MCC):

$$Sensibilidad = \frac{TP}{TP + FN} \tag{3.6}$$

$$Precisión = \frac{TP}{TP + FP} \tag{3.7}$$

$$Especificidad = \frac{TN}{TN + FP} \tag{3.8}$$

$$Exactitud = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.9}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.10}$$

3.5. Justificación en el problema de la PEP

Para poder abordar el problema de la predicción de la estructura de una proteína es necesario disponer de un procedimiento computacional, y todo procedimiento computacional tiene una entrada, un proceso de datos y una salida. Analizaremos en el siguiente capítulo los datos de entrada e información de salida más habituales utilizadas en la literatura referente al problema de la PEP, así como los métodos de predicción más relevantes que existen.

De entre todos los métodos que se han utilizado para resolver este problema, las técnicas de minería de datos han demostrado ser el mejor medio para abordar e intentar resolver el problema de la PEP, como veremos en el capítulo 4. Dentro de las técnicas de minería de datos aplicables a este problema, se encuentra el paradigma de los vecinos más cercanos. Como hemos visto, este paradigma se basa en la idea de que ejemplos similares en sus características presentan también un comportamiento similar en su atributo de predicción (clase).

Este paradigma ha sido el que se ha escogido para la propuesta realizada en esta Tesis por dos razones. Por una parte, existen multitud de características observables de las secuencias de proteínas y, aunque se desconoce cuáles son las que realmente determinan su estructura en todos los casos, se ha probado que proteínas con similares regiones estructurales comparten también, en la mayoría de los casos, ciertas características en sus secuencias [Gupta et al., 2005b]. Este comportamiento está en consonancia con el procedimiento de predicción efectuado por el algoritmo de vecinos más cercanos, ya que los atributos de entrada serán características de las secuencias y la clase, una región de su estructura.

Por otra parte, debido a la aparente complejidad del patrón natural que gobierna el plegamiento de las proteínas, las diversas técnicas de clasificación empleadas en la literatura han producido modelos de conocimiento realmente complejos [Bacardit et al., 2012, Li et al., 2011] y difícilmente interpretables. El algoritmo de vecinos más cercanos se caracteriza por carecer de un modelo de conocimiento propio, estando éste formado por todos los ejemplos de entrenamiento, eliminando el coste de generación del mismo. Además, el algoritmo de vecinos más cercanos, a diferencia de otros métodos como las redes neuronales artificiales, las máquinas de soporte vectorial o los algoritmos evolutivos, apenas posee parametrización, únicamente el número K de vecinos.

3.6. Resumen

En este capítulo hemos definido formalmente el paradigma de los vecinos más cercanos como técnica algorítmica tanto para la clasificación como la

regresión de ejemplos. Hemos mostrado los elementos necesarios para la formalización del algoritmo, así como el propio algoritmo. También hemos definido las métricas de evaluación que se utilizan tanto en clasificación como en regresión. En último lugar, se ha proporcionado una justificación a la elección de este esquema algorítmico para el método de PEP propuesto en esta Tesis.

Parte II

Estado del arte

Capítulo 4

Estado del arte

4.1. Introducción

En este capítulo se abordan las diferentes propuestas existentes en la literatura referente a la predicción de estructura terciaria de proteínas. Se han analizado 65 propuestas publicadas en el periodo 2003-2013, de ellas 58 aparecen en revistas con impacto, 4 en Lecture Notes y 3 en actas de congresos de IEEE. Para explicar dichas propuestas, se ha dividido el capítulo en varias secciones.

En la sección 4.2 se indican cuáles son los datos de entrada que más se han utilizado para predecir la estructura de una proteína. Clasificaremos estos datos en dos grupos, ya que pueden estar derivados únicamente de la secuencia de aminoácidos o bien pueden contener rasgos estructurales de la proteína. En último lugar se resume gráficamente la frecuencia de aparición de los datos de entrada en los diferentes métodos, detallando en una tabla cuáles son los datos que utilizan cada uno de ellos.

En la sección 4.3 analizaremos los cuatro tipos de información que los métodos de PEP generan para representar una estructura predicha de una proteína. Resumiremos la frecuencia con la que se utilizan en la literatura y detallaremos qué información genera cada método analizado.

En la sección 4.4 se explican las medidas de evaluación que habitualmente se utilizan en la literatura según el tipo de información de salida. Se explican también conceptos fundamentales en la evaluación de los métodos de PEP, tales como el umbral de contacto, la mínima separación y el ranking Top L/x. Se indican asimismo cuáles son los esquemas de validación más utilizados. Por último se resume gráficamente la frecuencia de uso de cada tipo de evaluación y validación en la literatura.

Abordaremos en la sección 4.5 los tres enfoques clásicos desde el punto de vista biológico para resolver el problema de la PEP, indicando, para cada método analizado, a qué enfoque pertenece.

En la sección 4.6 se explican los distintos métodos analizados,

clasificándolos por su tipo de aproximación algorítmica. Se han incluido cinco tipos de aproximaciones, que son las más empleadas en la literatura, y una categoría adicional para otros tipos de algoritmos. Dentro de cada tipo de aproximación, se detalla, para cada método, el tipo de evaluación y validación que emplea, así como qué medida utiliza y qué valor tiene.

Al final del capítulo se resumen gráficamente las tendencias principales de los métodos analizados, siendo un reflejo del comportamiento y evolución de la literatura referente al problema que nos ocupa.

4.2. Datos de entrada

Los datos de entrada que se utilizan en los métodos de PEP parten todos de la secuencia de aminoácidos de la proteína, tal como se constató en el epígrafe 1.1. No obstante, también se utilizan datos que contienen rasgos estructurales (de estructura secundaria o terciaria). Sin embargo, es importante señalar que estos datos estructurales son predicciones, no son datos reales, pues se derivan únicamente de estructuras de proteínas de entrenamiento y se infieren para secuencias de test.

Los datos de entrada, al ser derivados de la secuencia, es posible obtener siempre un dato para cada aminoácido. Existen numerosos métodos que, para realizar cada predicción parcial, utilizan datos de aminoácidos individuales de la secuencia. Éstos métodos los etiquetaremos como INDV, para indicar que utilizan aminoácidos individuales de la secuencia proteínica.

Otra forma de recopilar datos que se utiliza habitualmente es por intervalos, entornos o ventanas. Esto es, se recuperan datos de los aminoácidos que rodean a uno dado. Se suelen utilizar dos ventanas de un determinado tamaño, por ejemplo ± 3 aminoácidos, en torno a dos determinados (i y j) para predecir si éstos se encuentran cercanos en el espacio de la estructura de la proteína. Es decir, para dos ventanas de ± 3 aminoácidos centradas en i y j se recuperarían los aminoácidos $i-3, i-2, i-1, i, i+1, i+2, i+3$ y $j-3, j-2, j-1, j, j+1, j+2, j+3$. Éstos métodos los etiquetaremos como ENV, para indicar que utilizan todos los aminoácidos en uno o varios entornos de la secuencia proteínica.

Por último, existen métodos que utilizan datos que engloban la secuencia completa de aminoácidos, tales como su número de aminoácidos, la frecuencia de aparición de un tipo de aminoácido, etcétera. Etiquetaremos como GLOB a estos métodos, para indicar que utilizan todos los aminoácidos de la secuencia de forma global.

A continuación abordamos los diferentes datos de entrada que se utilizan en la literatura, clasificándolos en datos derivados de la secuencia de aminoácidos y en datos que contienen rasgos estructurales.

4.2.1. Datos derivados de la secuencia

Propiedades físico-químicas (PCP)

Las propiedades físico-químicas de aminoácidos (PCP, por sus siglas en inglés), como se explicó en el epígrafe 2.1, son características inherentes a la molécula que forma el grupo R del aminoácido. Estas características confieren propiedades únicas a cada uno de los 20 aminoácidos que existen en la naturaleza.

Las propiedades que son observables a veces se cuantifican numéricamente y a veces se dice si un aminoácido posee una propiedad o no la posee, es decir, se utiliza un valor lógico, como vimos en la figura 2.4. Lo habitual en los métodos de la literatura es utilizar una cuantificación numérica para las propiedades físico-químicas y el repositorio más completo y utilizado es AAindex, como se mencionó en el epígrafe 2.1.

Las seis propiedades físico-químicas más habituales se resumen en la tabla 4.1. Junto al nombre de la propiedad se indica la escala numérica más reciente de carácter general y su referencia. Aunque sólo indicamos una escala por propiedad, existen varias más que pueden consultarse en [Kawashima et al., 2008].

Propiedad	Escala	Referencia
Hidrofobicidad	KUHL950101	[Kuhn et al., 1995]
Polaridad	RADA880108	[Radzicka and Wolfenden, 1988]
Carga	KLEP840101	[Klein et al., 1984]
Volumen	PONJ960101	[Pontius et al., 1996]
Peso molecular	FASG760101	[Atsushi, 1980]
Área accesible	RADA880106	[Radzicka and Wolfenden, 1988]

Tabla 4.1: Propiedades físico-químicas de aminoácidos más utilizadas.

Las propiedades físico-químicas son datos muy valiosos a la hora de predecir la estructura de una proteína debido a que las fuerzas que gobiernan el proceso de plegamiento de la proteína son de tipo físico (fuerzas de Van der Waals, interacciones electrostáticas, puentes de hidrógeno) y químico (enlaces disulfuro, interacciones hidrofóbicas) y dependen de los aminoácidos de la secuencia. Etiquetaremos con PCP a los métodos analizados que utilizan propiedades físico-químicas de aminoácidos.

Mutaciones correlacionadas (CMU)

Las mutaciones correlacionadas (CMU, por sus siglas en inglés) son evidencias de coevolución entre aminoácidos de una secuencia. Esto quiere decir que ciertos grupos de aminoácidos han mantenido una colaboración a través de la evolución de una secuencia proteínica. La evidencia de ello

se puede obtener si analizamos su traza evolutiva, es decir, su alineamiento múltiple de secuencias, en busca de aminoácidos que cambian mutuamente. Es decir, cuando un aminoácido ha mutado en algún momento de la evolución, otros aminoácidos relacionados con él lo han hecho también para adaptarse y conseguir que el conjunto de la molécula se mantenga estable en estructura y, por ende, en funciones.

Se ha comprobado que aminoácidos que se encuentran próximos en el espacio de la conformación nativa de una proteína (es decir, que se encuentran haciendo contacto) presentan las evidencias anteriormente citadas. Esto es, los cambios evolutivos en un aminoácido se compensan con cambios en el aminoácido con el que se encuentra en contacto, tal como se muestra en la figura 4.1.

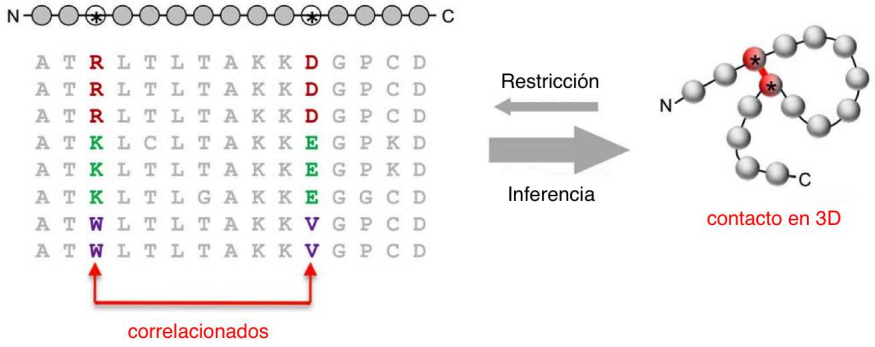


Figura 4.1: Mutaciones correlacionadas (adaptada de [Marks et al., 2011]).

La lectura interesante de este comportamiento, desde el punto de vista de la PEP, es, como sugiere la figura 4.1, inferir que dos aminoácidos hacen contacto en el espacio a partir de que ambos están correlacionados en el alineamiento múltiple de la secuencia en la que se encuentran. Muchos métodos usan esta técnica con éxito para predecir contactos entre aminoácidos.

No obstante, existe un problema fundamental en esta técnica de predicción y es que si bien cuando existe contacto en 3D, existe mutación correlacionada, esta implicación no se tiene en sentido inverso. Es decir, una mutación correlacionada no implica necesariamente contacto en 3D. Esto es así debido a la existencia de correlaciones indirectas o transitivas. Si un aminoácido A está en contacto 3D con uno B (contacto entre A y B) y éste está en contacto con uno C (contacto entre B y C), tanto A y B como B y C están correlacionados y por tanto también lo están siempre A y C. Sin embargo A y C no hacen contacto en 3D. Este problema, diagnosticado en 1999 por Lapedes et al. [Lapedes et al., 1999], se conoce como análisis de emparejamiento directo (DCA: Direct Coupling Analysis).

Existen actualmente tres formas principales de resolver este problema que se han llevado a la práctica: mediante la maximización de la entropía [Burkoff et al., 2013, Miyazawa, 2013, Morcos et al., 2011], resolviendo un modelo Potts (generalización del modelo de Ising) [Ekeberg et al., 2013] y mediante la covarianza inversa y dispersa [Savojardo et al., 2013, Jones et al., 2012, Nugent and Jones, 2012]. No obstante, existen numerosos trabajos, por lo general de principios del año 2011 y anteriores, que utilizan mutaciones correlacionadas para predecir contactos y no hacen ningún DCA, incurriendo por tanto en un gran número de falsos positivos [Wang et al., 2011, Shackelford and Karplus, 2007].

Se haga o no un DCA, las técnicas de predicción de contactos basadas en mutaciones correlacionadas adolecen de otro problema, y es que dependen de la cantidad de secuencias conocidas. Con un alineamiento múltiple que no tenga al menos 1000 secuencias no redundantes, no hay información evolutiva suficiente para que las mutaciones correlacionadas tengan significancia [Sułkowska et al., 2012]. Por ejemplo, el número de secuencias alineadas, en el caso de las proteínas del campeonato CASP, es insuficiente para alcanzar la precisión obtenida por los mejores métodos.

Etiquetaremos con CMU a los métodos analizados que utilizan mutaciones correlacionadas y con DCA a los que, usando CMU, realizan un análisis DCA para reducir falsos positivos.

Matriz de puntuación de posición específica (PSSM)

La matriz de puntuación de posición específica (PSSM) es una matriz que representa motivos en una secuencia proteínica, indicando la relación evolutiva entre los aminoácidos de una secuencia y su posición en la misma con los de otras secuencias alineadas.

En concreto, PSSM es una matriz que tiene 20 filas (debido a que existen 20 posibles aminoácidos) y tantas columnas como aminoácidos tenga la secuencia a analizar. El elemento (a, b) de la matriz PSSM indica la tendencia evolutiva del aminoácido situado en la posición b de la secuencia a mutar hacia el tipo de aminoácido a. Las matrices PSSM se generan automáticamente a partir de la herramienta de alineamiento múltiple PSI-BLAST [Altschul et al., 1997].

En los métodos de predicción de contactos entre aminoácidos habitualmente se utilizan las columnas i y j (de 20 elementos cada vector-columna) para predecir si los aminoácidos i y j de la secuencia se encuentran en contacto en la estructura. Etiquetaremos con PSSM a los métodos analizados que utilizan esta matriz.

Composición de aminoácidos (AAC)

Uno de los datos de entrada más frecuentes es la frecuencia, absoluta o relativa, de ocurrencia de cada uno de los 20 tipos de aminoácidos dentro de una secuencia o fragmento de la misma. Este dato se suele denominar composición de aminoácidos (AAC, por sus siglas en inglés). Por ejemplo, el fragmento $PRRGAPRC$ tiene la siguiente composición de aminoácidos: $\{A = 1, R = 3, N = 0, D = 0, C = 1, Q = 0, E = 0, G = 1, H = 0, I = 0, L = 0, K = 0, M = 0, F = 0, P = 2, S = 0, T = 0, W = 0, Y = 0, V = 0\}$.

La composición de aminoácidos ha sido usada habitualmente en combinación con otros datos de entrada para caracterizar estadísticamente un fragmento de secuencia, encontrar fragmentos similares y descubrir posibles reglas de comportamiento en fragmentos de secuencias y su implicación en la estructura real. No obstante, la AAC por sí sola ha demostrado ser claramente insuficiente para predecir la estructura de una proteína. Etiquetaremos con AAC a los métodos analizados que utilizan como entrada la composición de aminoácidos.

Propensiones estadísticas (STPR)

Las propensiones estadísticas reflejan el comportamiento observable más frecuente de cada aminoácido en un entorno determinado. Por ejemplo, la propiedad denominada propensión de pertenencia a hélice alfa refleja, para cada aminoácido, la frecuencia relativa de que un aminoácido forme parte de una hélice alfa dentro de una proteína.

Como se indicó en el epígrafe 2.1, el repositorio AAindex, además de propiedades físico-químicas, contiene un buen número de propensiones estadísticas. No obstante, la gran mayoría de los métodos calculan sus propias propensiones estadísticas utilizando sus propios conjuntos de proteínas, que utilizan como entrenamiento, lo cual refleja que las estadísticas obtenidas de un conjunto de proteínas no aportan valor predictivo universal.

Esto se debe a la compleja naturaleza del problema que nos ocupa, donde los más ínfimos detalles de la secuencia, en unas ocasiones unos y en otras ocasiones otros, observables o no, determinan la estructura de una proteína. Etiquetaremos con STPR a los métodos analizados que utilizan como entrada alguna propensión estadística.

Funciones de energía (ENER)

La conformación nativa de una proteína es la conformación de más baja energía libre de Gibbs. El proceso de plegamiento, como cualquier otro proceso biológico, se encuentra bajo control termodinámico y cinético, y es un proceso que está claramente favorecido en condiciones fisiológicas.

Desde el punto de vista entrópico, el proceso de plegamiento supone una disminución de entropía desde la estructura primaria a la conformación nativa. Este descenso de entropía, denominada entropía conformacional, supone un incremento positivo de energía libre en el proceso de plegamiento ($\triangle G = \triangle H – T \triangle S$).

Sin embargo, para que la energía libre ($\triangle G$) aumente es necesario que el incremento de entalpía ($\triangle S$) sea negativo o que existan otros aumentos de entropía ($\triangle H$). La principal contribución entálpica al proceso de plegamiento la constituyen la formación de interacciones no covalentes que estabilizan la estructura nativa y las interacciones hidrofóbicas entre las cadenas laterales apolares que generalmente quedan localizados en el interior de la estructura nativa.

Alguna estructura plegada corresponde con un mínimo de energía libre de la proteína en condiciones fisiológicas. Éste es el motivo por el que las cadenas de aminoácidos se pliegan espontáneamente.

Los dos principales problemas son el cálculo de la energía libre de la proteína y la resolución del mínimo global de esta energía. Un método de predicción basado en energías debe explorar el espacio de posibles estructuras proteínicas que, aunque es astronómicamente inmenso, se utilizan diversas heurísticas y métodos estocásticos de búsqueda. Para cada estructura candidata, debe evaluar una función de energía y encontrar la conformación que minimice dicha función.

Una de las funciones de energía más utilizadas es la AMBER99 [Cornell et al., 1995]. Existen, no obstante, numerosas funciones de energía disponibles y muchos métodos diseñan las suyas propias. Incluso existen tesis doctorales dedicadas a la generación automática de funciones de energía [Widera, 2010]. Etiquetaremos con ENER a los métodos analizados que utilizan como entrada alguna función de energía.

4.2.2. Datos estructurales

Estructura secundaria (SS)

Como se vio en el epígrafe 2.1, la estructura secundaria (SS, por sus siglas en inglés) de una proteína contiene tres elementos: hélices alfa, láminas beta y regiones sin hélices ni láminas denominadas *random coils*. El conocimiento de esta estructura ha demostrado ser de gran utilidad para la predicción de la estructura terciaria y cuaternaria, debido a que estas últimas contienen los mismos elementos de la estructura secundaria, sólo que más compactados.

Sin embargo, la estructura secundaria no es un dato que se pueda derivar de forma determinista a partir de la secuencia de aminoácidos, como ocurre con todas las características que se han visto en el pasado epígrafe 4.2.1. En lugar de ello, existen algoritmos de predicción de estructura secundaria que ponen a disposición datos predichos de estructura secundaria para ser

utilizados en la predicción de estructura terciaria o cuaternaria.

Existen numerosos métodos de predicción de estructura secundaria, no obstante el más utilizado en todos los métodos de predicción de estructura terciaria y, en concreto, en todos los métodos analizados en este capítulo, es PSI-PRED [McGuffin et al., 2000]. Su tasa de acierto se sitúa en torno al 80 %, lo cual introduce un grado de error en los datos de SS que en la mayoría de los casos es aceptable. Etiquetaremos con SS a los métodos analizados que utilizan como entrada predicciones de estructura secundaria.

Accesibilidad al solvente (SA)

La accesibilidad al solvente (SA) mide el grado de exposición de un aminoácido al solvente de la proteína, habitualmente el agua. Las proteínas son en su gran mayoría de morfología globular y compactas. Los aminoácidos que se encuentran en el interior o núcleo de la proteína tienen una accesibilidad al solvente mínima o nula, mientras que los aminoácidos que se encuentran en la superficie de la proteína tienen una accesibilidad al solvente máxima.

Por una cuestión geométrica sencilla, todos los aminoácidos del interior de la proteína se encuentran más cerca de sí unos de otros que todos los aminoácidos de la superficie entre sí, ya que el espacio que ocupan es mucho menor y, por tanto, la densidad es mayor. De ello se deduce que existen muchos más contactos entre aminoácidos del interior que entre aminoácidos de la superficie. Este resultado es extensamente utilizado en los métodos de predicción de estructura terciaria, asignando mayor probabilidad de contacto a los aminoácidos del interior de la proteína.

Además, también es conocido que la SA se encuentra inversamente correlacionada con la hidrofobicidad. Esto es, los aminoácidos menos expuestos suelen ser los más hidrofóbicos, y los más expuestos, los más hidrofílicos.

No obstante, la accesibilidad al solvente, al igual que la estructura secundaria, no es un dato derivado de la secuencia, sino de la estructura. Por ello, al igual que para la SS, los métodos analizados utilizan predictores de accesibilidad al solvente. Aunque existen muchos, uno de los más utilizados es ACCpro [Pollastri et al., 2002]. Etiquetaremos con SA a los métodos analizados que utilizan como entrada predicciones de accesibilidad al solvente.

Perfil de conectividad efectivo (ECP)

Uno de los rasgos estructurales emergentes más interesantes es el denominado perfil de conectividad efectivo (ECP), introducido en el año 2005 por Ugo Bastolla et al. [Bastolla et al., 2005] y revisado en 2008 [Bastolla et al., 2008].

El ECP es una combinación lineal de los autovalores y autovectores de la matriz de contactos de una proteína. Abordaremos las matrices de contactos en el epígrafe 4.3.4. El ECP es capaz de capturar la parte constante y característica de la estructura de una proteína. Además se ha demostrado que el ECP está altamente correlacionado con la distribución de hidrofobicidad de la secuencia de aminoácidos [Bastolla et al., 2005, Bastolla et al., 2008]. Esto está en consonancia con lo que hemos comentado anteriormente: una de las fuerzas más importantes que actúan en el plegamiento de un proteína son las interacciones hidrofóbicas.

Como todo rasgo estructural, el ECP no puede ser derivado directamente de una secuencia de aminoácidos y debe ser predicho. Se han incluido al análisis dos métodos de PEP que utilizan el perfil de conectividad efectivo [Wolff et al., 2008, Wolff et al., 2010] y se han etiquetado con ECP.

4.2.3. Selección de atributos (FSEL)

La selección de atributos es una tarea clásica en minería de datos y consiste en encontrar las características más relevantes de un conjunto de datos para la predicción de su clase.

Como se ha podido comprobar, existen multitud de posibles datos de entrada en este problema y, además, unido al hecho de que todos los datos de entrada pueden ser obtenidos para cada aminoácido de la secuencia.

Es habitual que muchos métodos de la literatura utilicen perfiles con centenares e incluso más de mil atributos (1287 atributos en [Li et al., 2011]). Además, con tal número de atributos es necesario disponer de suficientes ejemplos para cubrir mínimamente el espacio de búsqueda, lo cual supone decenas de millones de instancias. Esto supone un coste computacional muy alto y, con menor número de atributos y bien escogidos, se ganaría ampliamente en eficiencia y, posiblemente, también en eficacia.

Sin embargo, la selección de atributos se ha utilizado escasamente en la PEP. De hecho, sólo 5 de los 65 métodos analizados (el 7.7 %) incorporan una selección de atributos previa [Abu-Doleh et al., 2012, Wang et al., 2011, Lo et al., 2009, Cheng and Baldi, 2007, Shi et al., 2004]. Es cierto que algunos algoritmos de clasificación y regresión llevan implícita una selección de atributos, como en los algoritmos evolutivos o en las máquinas de soporte vectorial. No obstante, el gran número inicial de atributos hace muy costosa su ejecución. Se han etiquetado con FSEL a los métodos analizados que utilizan una selección de atributos previa al proceso de predicción principal.

4.2.4. Resumen

Con el objetivo de ofrecer una visión más completa y detallada del empleo de los datos de entrada en los métodos analizados se han incluido dos histogramas y tres tablas que describimos a continuación.

En la figura 4.2 se muestra un histograma con las frecuencias absolutas de uso de los datos de entrada en los 65 métodos analizados. Como se puede apreciar, los datos más usados son de SS, seguido de PSSM y PCP.

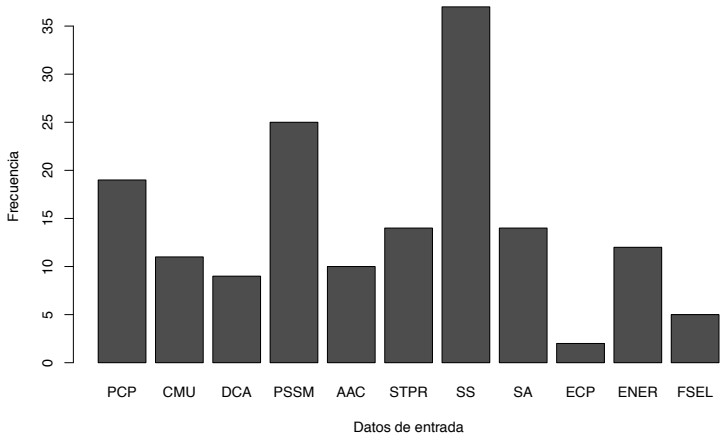


Figura 4.2: Histograma de uso de los datos de entrada.

En la figura 4.3 se muestra la frecuencia del alcance utilizado en la obtención de los datos a partir de la secuencia, como se explicó anteriormente en el epígrafe 4.2. Como se puede observar, lo más habitual en los métodos que se han estudiado es utilizar datos obtenidos de aminoácidos individuales.

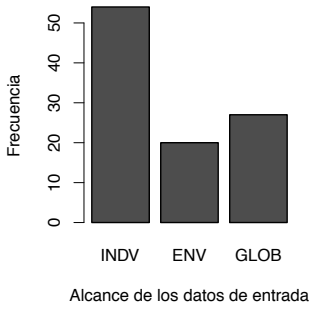


Figura 4.3: Histograma del alcance utilizado en los datos de entrada.

Para concluir el resumen de los datos de entrada, en las tablas 4.2, 4.3 y 4.4 se muestran todos los datos de entrada y su alcance utilizados por cada uno de los 65 métodos analizados, en orden cronológico descendente.

Método	PCP	CMU	DCA	PSSM	AAC	STPR	SS	SA	ECP	ENER	FSEL	INDV	ENV	GLOB
[Burkoff et al., 2013]		•	•									•		
[Ekeberg et al., 2013]		•	•									•		
[Miyazawa, 2013]	•	•	•									•		
[Savojardo et al., 2013]		•	•	•	•							•	•	•
[Abu-Doleh et al., 2012]				•		•	•	•			•	•	•	•
[Aydin et al., 2012]				•		•						•		
[Bacardit et al., 2012]				•	•	•	•	•				•	•	•
[Di Lena et al., 2012]				•	•		•	•				•	•	•
[Eickholt and Cheng, 2012]				•		•	•	•				•	•	•
[Jones et al., 2012]		•	•									•		
[Nugent and Jones, 2012]		•	•				•					•		
[Sułkowska et al., 2012]		•	•			•	•					•		
[Ashkenazy et al., 2011]														
[Calvo et al., 2011]							•					•		
[Eickholt et al., 2011]														
[Li et al., 2011]	•			•	•	•	•	•				•	•	•
[Marks et al., 2011]		•	•									•		
[Morcos et al., 2011]		•	•											•
[Savojardo et al., 2011]				•									•	•
[Wang et al., 2011]		•		•							•	•	•	
[Wei and Floudas, 2011]							•					•		
[Wu et al., 2011]										•		•	•	•
[Yang and Chen, 2011]														
[Chen and Li, 2010]				•									•	
[Dorn and N. de Souza, 2010]						•	•					•	•	
[Hoque et al., 2010]	•											•		
[Rajgaria et al., 2010]	•						•					•		
[Wang et al., 2010]				•								•		

Tabla 4.2: Datos de entrada y alcance utilizado por los métodos analizados (1/3).

Método	PCP	CMU	DCA	PSSM	AAC	STPR	SS	SA	ECP	ENER	FSEL	INDV	ENV	GLOB
[Wolff et al., 2010]				•			•		•			•		
[Zhang et al., 2010a]				•								•		
[Zhang et al., 2010b]										•		•		•
[Björkholm et al., 2009]				•			•					•	•	•
[Gao et al., 2009]														
[Islam and Chetty, 2009]	•											•		
[Judy et al., 2009]							•			•		•		
[Karplus, 2009]				•			•			•		•		•
[Lippi and Frasconi, 2009]	•				•	•	•	•				•	•	•
[Lo et al., 2009]				•		•	•	•			•	•	•	
[Rajgaria et al., 2009]	•					•	•			•		•		•
[Shell et al., 2009]										•		•		
[Tegge et al., 2009]				•			•	•				•		
[Walsh et al., 2009]				•		•	•	•		•				
[Xue et al., 2009]	•			•	•		•	•				•	•	•
[Zhang, 2009]	•			•	•		•	•		•		•	•	•
[Shi et al., 2008]				•								•		
[Wolff et al., 2008]								•		•				•
[Cheng and Baldi, 2007]	•				•	•	•	•			•	•	•	•
[Shackelford and Karplus, 2007]		•			•	•	•	•				•	•	
[Zhang and Han, 2007]	•						•					•		•
[Colubri et al., 2006]						•	•			•		•		•
[Cutello et al., 2006]							•			•		•		•
[Davies et al., 2006]							•					•		•
[Glasgow et al., 2006]							•					•		•
[Liu et al., 2006]	•						•					•		
[Sander et al., 2006]	•			•								•		
[Vullo et al., 2006]	•						•					•		

Tabla 4.3: Datos de entrada y alcance utilizado por los métodos analizados (2/3).

Método	PCP	CMU	DCA	PSSM	AAC	STPR	SS	SA	ECP	ENER	FSEL	INDV	ENV	GLOB
[Gupta et al., 2005a]	•											•		
[Han et al., 2005]				•			•					•		•
[Punta and Rost, 2005]	•			•			•	•					•	•
[Zhang et al., 2005]	•						•					•		•
[MacCallum, 2004]				•									•	
[Shi et al., 2004]	•						•			•		•		
[Zhang and Huang, 2004a]					•		•					•		•
[Zhang and Huang, 2004b]							•					•		
[Cotta, 2003]	•									•		•		

Tabla 4.4: Datos de entrada y alcance utilizado por los métodos analizados (3/3).

4.3. Información de salida

4.3.1. Modelo tridimensional (3D)

El modelo de predicción más completo de la estructura de una proteína es el modelo 3D. En él se incluyen las coordenadas espaciales de cada uno de los átomos de la proteína. Es también el modelo más difícil de predecir, pues es el más complejo y el que más detalles tiene. En última instancia éste es el modelo que tiene realmente utilidad para la finalidad con la que se hace PEP: comprensión de las funciones proteínicas, diseño de fármacos, detección de lugares de interacción entre proteínas, etcétera.

Lo habitual cuando un método de PEP produce un modelo 3D es generar previamente un modelo simplificado basado únicamente en un átomo por aminoácido, habitualmente el carbono alfa (carbono que enlaza con el grupo R) o el beta (el carbono del grupo R que enlaza con el carbono alfa).

Aunque existen diversas técnicas para construir este modelo simplificado, lo habitual es utilizar modelos de rejillas [Hoque et al., 2010], ensamblar fragmentos de estructuras conocidas [Dorn and N. de Souza, 2010], o bien predecir previamente contactos [Sułkowska et al., 2012] o ángulos de torsión [Judy et al., 2009], como veremos a continuación en el siguiente epígrafe.

Se han etiquetado con 3D a los métodos analizados que producen un modelo tridimensional como resultado de la predicción de estructura proteínica.

4.3.2. Ángulos de torsión (TANG)

Los ángulos de torsión, ángulos dihédricos o ángulos ϕ y ψ son los formados entre un aminoácido y el siguiente en la secuencia y que permiten el movimiento de la molécula, tal como se ilustró en la figura 2.2.

Dado que la distancia entre dos aminoácidos consecutivos es prácticamente constante ($\simeq 3{,}8\mathring{A}$), si se conocen los ángulos de torsión, se puede inferir con relativa facilidad un modelo 3D.

Se han etiquetado con TANG a los métodos analizados que predicen los ángulos de torsión de una proteína, bien sea éste su modelo final o construyan un modelo 3D en torno a dichos ángulos. En este último caso, los métodos son marcados con ambas etiquetas (TANG y 3D).

4.3.3. Mapa de distancias (DMAP)

El mapa o matriz de distancias de una secuencia de proteína es una matriz cuadrada $DM \in \Re^{L \times L}$, siendo L la longitud de dicha secuencia. Los elementos de la triangular superior de la matriz DM contienen las distancias reales en el espacio entre todos los pares de aminoácidos de la secuencia. En concreto, el elemento $dm_{i,j}$ con $i < j$ contiene la distancia

entre el aminoácido i-ésimo y el j-ésimo de la secuencia, en el espacio de la conformación nativa de la proteína.

Los elementos de la triangular inferior de la matriz DM contienen las distancias predichas entre todos los pares de aminoácidos de la secuencia, obtenidas mediante algún algoritmo de predicción de distancias.

Los mapas de distancias pueden colorearse asignando un color a cada elemento de la matriz, como se ilustra en la figura 4.4. Este color depende del valor de la celda y suele emplearse una escala continua de colores desde el valor mínimo (que corresponde al color rojo) y el valor máximo (que corresponde al color azul). De este modo, los mapas de distancias son visualmente útiles, pues pueden apreciarse las regiones de la molécula donde se producen acercamientos entre aminoácidos.

La mayoría de estos acercamientos (regiones de color rojo en la figura 4.4) se deben a dos motivos. El primero de ellos, por acercamiento en la secuencia: los aminoácidos cercanos en la secuencia estarán cercanos en el espacio, porque todos los aminoácidos están unidos entre sí formando una cadena. Este tipo de proximidades se localizan en torno a la diagonal principal de la matriz.

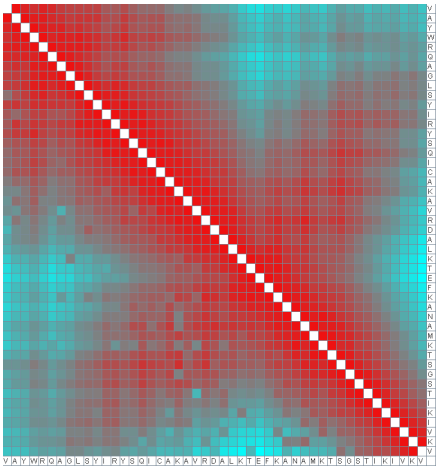


Figura 4.4: Mapa de distancia de una proteína (PDBID: 1E79).

El segundo motivo principal de proximidades entre aminoácidos es la formación de la estructura secundaria. Como se explicó en el epígrafe 2.3, tanto las hélices alfa como las láminas beta generan proximidades entre aminoácidos. Estas correspondencias pueden visualizarse en la figura 4.5 referente a los mapas de contactos, en el siguiente epígrafe.

Al igual que con los ángulos de torsión, a partir de un mapa de distancias es posible reconstruir un modelo tridimensional de la proteína. Para ello es necesario resolver el problema geométrico discreto de distancias moleculares

[Lavor et al., 2012, Mucherino et al., 2012].

Tras una exhaustiva búsqueda en la literatura, tan sólo se ha encontrado un método que predice mapas de distancias de proteínas [Zhang and Huang, 2004b]. Dicho método ha sido etiquetado con DMAP para un posterior análisis.

4.3.4. Mapa de contactos (CMAP)

Los mapas o matrices de contactos son un caso concreto de matrices de distancias donde sus valores han sido discretizados utilizando un valor umbral. En concreto, el mapa de distancias de una secuencia de proteína es una matriz cuadrada $CM \in \{0,1\}^{L \times L}$, siendo L la longitud de la secuencia.

Los elementos de la triangular superior de la matriz CM contienen valores binarios de proximidad real entre todos los pares de aminoácidos de la secuencia. El elemento $cm_{i,j}$ con $i < j$ tiene valor 1 si la distancia en el espacio entre el aminoácido i-ésimo y el j-ésimo es menor o igual a un umbral, denominado umbral de contacto. El elemento $cm_{i,j}$ con $i < j$ tiene valor 0 si la distancia entre los aminoácidos i y j es mayor que dicho umbral.

Los elementos de la triangular inferior de la matriz CM contienen las predicciones de contactos entre todos los pares de aminoácidos de la secuencia, obtenidas mediante algún algoritmo de predicción de contactos.

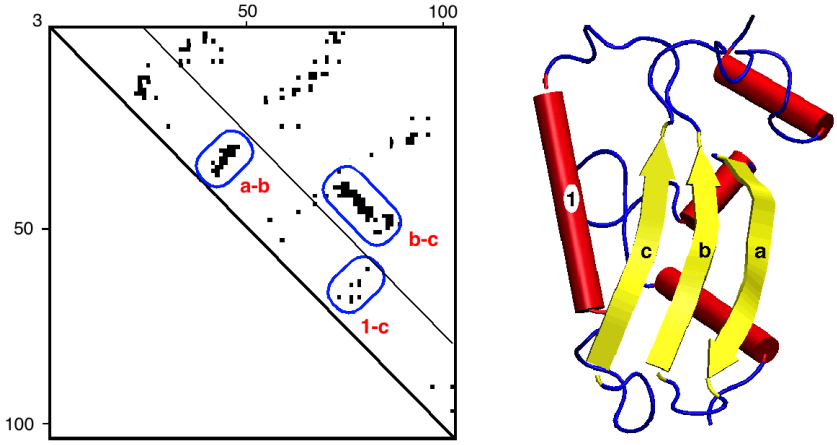


Figura 4.5: Correspondencia entre el mapa de contactos y la estructura secundaria (adaptado a partir de [Punta and Rost, 2005]).

En la figura 4.5 se ilustra un mapa de contactos y el modelo 3D de la proteína que representa. Los elementos de la matriz CM que tienen valor 1 se han coloreado en negro, mientras que los que tienen valor 0 en blanco. Además se han señalado las correspondencias entre la estructura secundaria

(hélices alfa y láminas beta) y los contactos en la matriz CM. En concreto, se han destacado los contactos entre los aminoácidos de la hebra beta a y la b ($a - b$), los de la hebra b y la c ($b - c$), y los de la hebra c y la hélice 1 ($1 - c$).

Definimos separación entre dos aminoácidos i y j al número de aminoácidos que se encuentran entre ambos dentro de la secuencia, es decir, $j - i$. Los contactos entre aminoácidos separados por menos de 6 en la secuencia son muy frecuentes y pueden predecirse con gran facilidad a partir de un dato de entrada de tipo SS (predicción de estructura secundaria).

Sin embargo, los contactos entre aminoácidos con separación 24 o superior, denominados contactos de largo alcance, son más escasos, más informativos y más difíciles de predecir. De hecho, el campeonato CASP, mencionado en el epígrafe 2.4, centra su evaluación en los contactos de largo alcance predichos por los métodos de PEP que compiten.

Por otra parte, existen algunos métodos que discretizan las distancias entre aminoácidos utilizando varios intervalos discretos, proporcionando una representación intermedia entre los mapas de distancias y los de contactos. Por ejemplo, el método de Walsh et al. [Walsh et al., 2009] utiliza matrices de distancias discretizadas en 4 intervalos.

Al igual que los ángulos de torsión y los mapas de distancias, a partir de un mapa de contactos es posible también reconstruir un modelo tridimensional de la proteína [Vassura et al., 2011]. Se han etiquetado con CMAP a los métodos analizados que predicen mapas de contactos de proteínas.

4.3.5. Resumen

En la figura 4.6 se ha resumido el tipo de información de salida generado por los métodos que se han analizado. Como se puede apreciar, 48 de los 65 métodos (el 73.8 %) predicen contactos entre aminoácidos, mientras que tan sólo uno predice mapas de distancias.

Sin embargo, desde un punto de vista práctico, el mapa de distancias es más ventajoso que el mapa de contactos por tres motivos principales:

- El mapa de distancias aporta mayor información que el mapa de contactos acerca de la estructura de una proteína, lo cual es beneficioso para la reconstrucción de modelos tridimensionales de mayor calidad.

- Un mapa de distancias puede convertirse en un mapa de contactos con tan sólo aplicar el valor de umbral deseado. De este modo, cualquier aplicación que utilice mapas de contactos puede ser empleada a partir de un mapa de distancias.

- ¿Por qué utilizar 8 angstroms como umbral estándar? Se ha observado que las interacciones entre aminoácidos cercanos en la proteína se

producen a diferentes distancias, según sea el tipo de interacción y el lugar donde se produzca. Por tanto, fijar cualquier umbral de contacto puede despreciar un buen número de interacciones que se producen a distancias superiores al umbral, al mismo tiempo que puede reconocer interacciones inferiores al umbral que en realidad no lo son. Los mapas de distancias no adolecen de este problema, pues carecen de umbral.

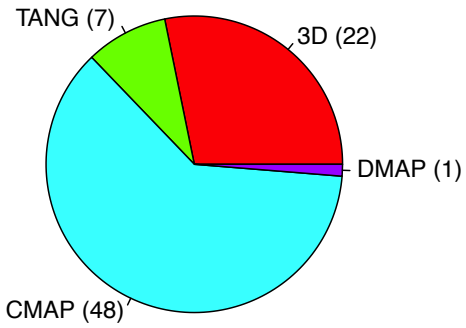


Figura 4.6: Información de salida producida por los métodos analizados.

Para concluir el resumen de la información de salida, en las tablas 4.5, 4.6 y 4.7 se muestra la información de salida (en las cuatro primeras columnas: 3D, TANG, DMAP y CMAP) producida por cada uno de los 65 métodos analizados, en orden cronológico descendente.

Método	3D	TANG	DMAP	CMAP	HOM	ABI	THR	STAT	ANN	SVM	EC	CBR	OAP
[Burkoff et al., 2013]				•		•		•					
[Ekeberg et al., 2013]				•		•		•					
[Miyazawa, 2013]				•		•		•					
[Savojardo et al., 2013]				•		•				•			
[Abu-Doleh et al., 2012]				•		•						•	•
[Aydin et al., 2012]		•				•			•				•
[Bacardit et al., 2012]				•		•					•		
[Di Lena et al., 2012]				•		•			•				
[Eickholt and Cheng, 2012]				•		•			•				
[Jones et al., 2012]				•		•							•
[Nugent and Jones, 2012]	•					•							•
[Sułkowska et al., 2012]	•	•		•		•							•
[Ashkenazy et al., 2011]				•	•	•							•
[Calvo et al., 2011]	•	•			•						•		
[Eickholt et al., 2011]				•	•			•					
[Li et al., 2011]				•		•							•
[Marks et al., 2011]	•			•		•							•
[Morcos et al., 2011]				•		•							•
[Savojardo et al., 2011]				•		•				•			
[Wang et al., 2011]				•		•							•
[Wei and Floudas, 2011]				•		•							•
[Wu et al., 2011]				•		•				•			
[Yang and Chen, 2011]				•		•							•
[Chen and Li, 2010]				•		•					•		
[Dorn and N. de Souza, 2010]	•	•			•				•				
[Hoque et al., 2010]	•					•					•		
[Rajgaria et al., 2010]				•		•							•
[Wang et al., 2010]	•				•								•

Tabla 4.5: Información de salida, tipo de aproximación biológica y algorítmica de los métodos analizados (1/3).

Método	3D	TANG	DMAP	CMAP	HOM	ABI	THR	STAT	ANN	SVM	EC	CBR	OAP
[Wolff et al., 2010]	•					•	•		•				
[Zhang et al., 2010a]	•				•	•	•						
[Zhang et al., 2010b]		•				•					•		
[Björkholm et al., 2009]				•		•							•
[Gao et al., 2009]				•		•							•
[Islam and Chetty, 2009]	•					•					•		
[Judy et al., 2009]	•	•				•					•		
[Karplus, 2009]	•			•	•	•		•	•				
[Lippi and Frasconi, 2009]				•		•			•				
[Lo et al., 2009]				•		•				•			
[Rajgaria et al., 2009]				•		•							•
[Shell et al., 2009]	•					•							•
[Tegge et al., 2009]				•		•			•				
[Walsh et al., 2009]	•			•	•	•			•				
[Xue et al., 2009]				•		•			•				
[Zhang, 2009]	•			•	•	•	•	•		•			
[Shi et al., 2008]				•		•		•					
[Wolff et al., 2008]	•			•			•						•
[Cheng and Baldi, 2007]				•		•				•			
[Shackelford and Karplus, 2007]				•		•			•				
[Zhang and Han, 2007]				•		•					•		
[Colubri et al., 2006]	•				•							•	
[Cutello et al., 2006]	•	•				•					•		
[Davies et al., 2006]				•	•							•	
[Glasgow et al., 2006]	•			•	•							•	
[Liu et al., 2006]				•		•			•				
[Sander et al., 2006]	•			•	•					•			•
[Vullo et al., 2006]				•		•			•				

Tabla 4.6: Información de salida, tipo de aproximación biológica y algorítmica de los métodos analizados (2/3).

Método	3D	TANG	DMAP	CMAP	HOM	ABI	THR	STAT	ANN	SVM	EC	CBR	OAP
[Gupta et al., 2005a]				•		•					•		
[Han et al., 2005]	•				•					•			
[Punta and Rost, 2005]				•		•			•				
[Zhang et al., 2005]				•		•			•				
[MacCallum, 2004]				•		•					•		
[Shi et al., 2004]				•		•					•		
[Zhang and Huang, 2004a]				•		•			•		•		
[Zhang and Huang, 2004b]			•			•			•		•		
[Cotta, 2003]	•					•					•		

Tabla 4.7: Información de salida, tipo de aproximación biológica y algorítmica de los métodos analizados (3/3).

4.4. Evaluación

En esta sección explicaremos las medidas de evaluación utilizadas según el tipo de información producida, tres conceptos fundamentales en la evaluación de mapas de contactos y terminaremos con los esquemas clásicos de validación empleados en los métodos de PEP.

4.4.1. Evaluación de 3D, TANG y DMAP

Aunque existen múltiples medidas que permiten evaluar la calidad de los modelos 3D de proteínas, dos de ellas son las más utilizadas en la literatura: RMSD y GDT_TS. En las ecuaciones 4.1 y 4.2 se muestran las definiciones de ambas medidas.

$$RMSD = \sqrt{\frac{1}{s}\sum_{i=1}^{n}(v_i^x - w_i^x)^2 + (v_i^y - w_i^y)^2 + (v_i^z - w_i^z)^2} \tag{4.1}$$

$$GDT\_TS = 100\frac{\sum_{d_i}\frac{GDT_{d_i}}{NT}}{4}, d_i \in \{1, 2, 4, 8\} \tag{4.2}$$

La raíz de la desviación cuadrática media (RMSD) representa la desviación absoluta (medida en angstroms) de los átomos C_α (carbono alfa, explicado en el epígrafe 2.1) entre el modelo predicho y la estructura real. En la ecuación 4.1, v^x, v^y y v^z representan las coordenadas de los átomos predichos, mientras que w^x, w^y y w^z representan las coordenadas de los átomos en la molécula real.

GDT_TS es la puntuación total del test de distancia global [Zemla, 2003] y es el criterio de evaluación que más se utiliza en el campeonato CASP. En la ecuación 4.2, GDT_{d_i} es el número de átomos C_α predichos que no se desvían más de d_i angstroms del átomo C_α real. NT es el número total de átomos C_α de la molécula. Para realizar esta evaluación, en primer lugar se encuentra el ajuste óptimo entre el modelo y la estructura real, a través de rotaciones y traslaciones en el espacio.

En lo que se refiere a la evaluación de los ángulos de torsión que resultan de una predicción, no se ha introducido en la literatura ninguna medida propia. En su lugar, los métodos de predicción de TANG construyen primero un modelo tridimensional a partir de los ángulos predichos y entonces calculan las medidas de evaluación de modelos 3D (principalmente RMSD y GDT_TS, como se ha comentado anteriormente).

En cuanto a la evaluación de mapas de distancias, tampoco se ha introducido en la literatura ninguna métrica específica. Existen dos posibilidades para evaluar un mapa de distancias: construir un modelo 3D a partir del mapa de distancias y evaluar el modelo 3D con las métricas vistas anteriormente; o bien, convertir el mapa de distancias a un mapa de

contactos aplicando un umbral y utilizar las métricas propias de evaluación de mapas de contactos, que veremos en el siguiente epígrafe.

El método de predicción de DMAP analizado [Zhang and Huang, 2004b] evalúa sus predicciones utilizando la segunda posibilidad que se ha descrito, pero con una cierta variación: utiliza dos umbrales en lugar de uno sólo. Es decir, discretiza los valores de distancias del DMAP en tres intervalos, y luego aplica medidas clásicas de evaluación de mapas de contactos. Esta discretización es inusual en la literatura e impide que el método pueda compararse con facilidad con otras propuestas.

En el método que se propone en esta Tesis, el cual se abordará en el siguiente capítulo, se predicen mapas de distancias y la evaluación que se realiza (capítulos 6 y 7) es la propia de los mapas de contactos, convirtiendo los mapas de distancias a contactos y utilizando el umbral más empleado en la literatura: 8 angstroms.

4.4.2. Evaluación de CMAP

La predicción de un mapa de contactos implica la predicción de los contactos entre cada par de aminoácidos de la secuencia proteínica. Por tanto, al ser un contacto un valor binario, desde el punto de vista de la minería de datos, la predicción de mapas de contactos es una tarea de clasificación sobre una clase binaria. Por consiguiente, todas las medidas de evaluación definidas en el epígrafe 3.4.2 son aplicables aquí. De entre todas ellas, las más utilizadas en la literatura son la precisión (accuracy) y la sensibilidad (recall o coverage).

La precisión indica el tanto por uno de contactos predichos que son reales sobre el total de contactos predichos. Ésta es la medida que más importancia tiene en la literatura referente a la predicción de mapas de contactos de proteínas, incluido el campeonato CASP. La sensibilidad indica el tanto por uno de contactos predichos reales sobre el total de contactos reales de la proteína.

Existen tres conceptos fundamentales en la evaluación de mapas de contactos que deben ser explicados para comprender la evaluación realizada en los métodos que se proponen en la literatura. Éstos son el umbral de contacto, la mínima separación en la secuencia y el ranking Top L/x.

Umbral de contacto

El umbral de contacto fue definido en el epígrafe 4.3.4 y caracteriza a un mapa de contactos de tal manera que dos mapas de contactos con diferente umbral no son comparables, y tampoco las medidas de evaluación que se derivan de ellos.

No todos los métodos de CMAP que se han analizado utilizan el mismo umbral de contacto, con lo que la comparación entre ellos no es posible.

En la figura 4.7 se muestra un histograma con los umbrales de contacto utilizados en los métodos que se han analizado. Como se puede comprobar, el más utilizado es 8 angstroms (en 60 % de los métodos de CMAP).

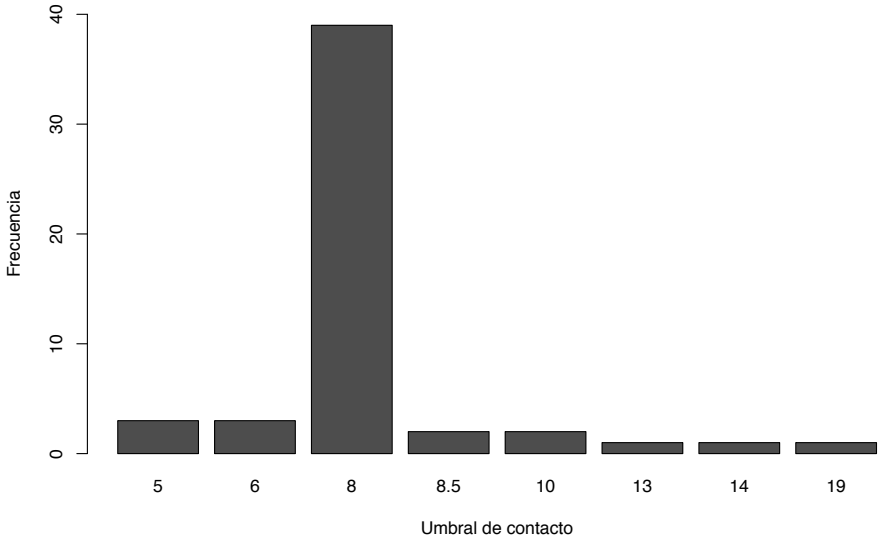


Figura 4.7: Umbrales de contactos utilizados en los métodos analizados.

Mínima separación en la secuencia

Como se ha explicado en el epígrafe 4.3.4, los contactos entre aminoácidos cuya separación es pequeña (5 aminoácidos o menos) son realmente sencillos de predecir utilizando datos de estructura secundaria. Por este motivo, la evaluación de los métodos de predicción de mapas de contactos suele realizarse teniendo en cuenta únicamente las predicciones entre aminoácidos cuya separación es mayor o igual a un determinado umbral. Este umbral se denomina mínima separación en la secuencia y, aunque existen algunos métodos que establecen su propia mínima separación, la que se establece en CASP es de 24 residuos [Moult et al., 2011].

En la figura 4.8 se muestra un histograma apilado cronológico de las mínimas separaciones en la secuencia utilizadas por los 65 métodos que se han analizado. En la leyenda de la figura, MS0 indica mínima separación 0, es decir, se contemplan las predicciones entre todos los pares de aminoácidos. Como se puede apreciar en la figura, la tendencia es utilizar cada vez más 24 como mínima separación, especialmente desde el año 2009.

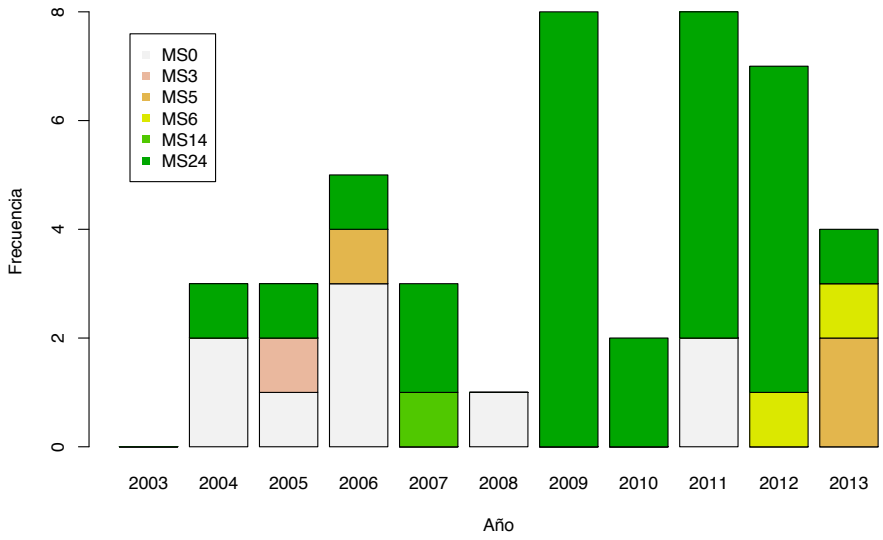


Figura 4.8: Mínimas separaciones que se han utilizado en cada año en los métodos analizados.

Ranking Top L/x

La evaluación de los métodos de predicción de mapas de contactos se sofistica aun más gracias a un requisito que impone el campeonato CASP, denominado ranking Top L/x. Este requisito consiste en ordenar y tomar sólo las L/x mejores predicciones de contacto realizadas en función de las probabilidades que el predictor asigna a cada una de ellas. L es la longitud de la proteína donde se encuentra el par de aminoácidos y x suele valer 1, 2, 5 ó 10. En concreto, para un determinado valor de x, se requiere llevar a cabo los siguientes pasos:

1. Asignar una probabilidad a cada predicción positiva de contacto entre pares de aminoácidos.

2. Ordenar descendentemente la lista de predicciones de contactos según su probabilidad.

3. Tomar los L/x primeras predicciones de la lista, donde L es la longitud de la proteína donde se encuentra el par de aminoácidos.

4. Calcular las medidas de precisión, sensibilidad, etcétera. sólo para las primeras predicciones.

Esta forma de evaluación premia no sólo la bondad de las predicciones de contactos, sino la capacidad del método de conocer la probabilidad con la que pueden existir contactos entre aminoácidos. En la figura 4.9 se muestran los tipos de ranking que se han utilizado en los métodos analizados. Se han etiquetado los métodos con LX1 (Top L/1), LX2 (Top L/2), LX5 (Top L/5), LX10 (Top L/10) y NOLX (sin ranking Top L/x). El tipo de ranking más utilizado, cuando se utiliza ranking, y el que además CASP recomienda, es el Top L/5 (LX5).

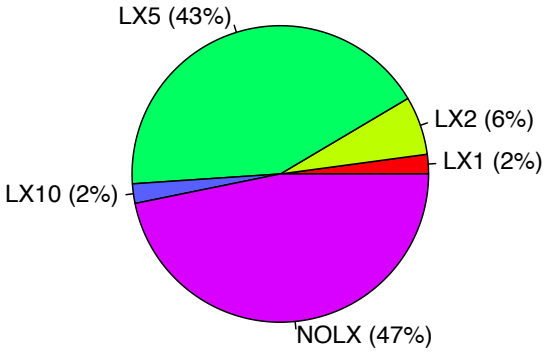


Figura 4.9: Tipos de ranking utilizados en los métodos analizados.

4.4.3. Validación

La validación de los métodos de PEP en la literatura está basada, salvo en casos excepcionales, en las técnicas de validación clásicas de la minería de datos, aunque en muchos casos no se realiza ninguna validación. Los tipos de validaciones que se han empleado han sido: *hold-out* (HOUT), validación cruzada con bolsas (CVFOLD) y validación cruzada *leave-one-out* (CVLOU). En la figura 4.10 se resumen los tipos de validación empleados en los métodos analizados, indicando con NOVAL a los métodos que no realizan ningún tipo de validación de sus resultados.

Como se puede apreciar en la figura 4.10, sólo el 65 % de los métodos analizados realizan algún tipo de validación, siendo el *hold-out* el más utilizado. Para conocer con mayor detalle las validaciones cruzadas con bolsas que se han realizado, en la figura 4.11 se muestra el número de bolsas que se han empleado. Como se puede apreciar, la validación cruzada con 10 bolsas es la que más se utiliza.

En el epígrafe 4.6 se detallará, para cada método, tanto la validación que utiliza, como todos los aspectos de la evaluación que se han tratado en esta sección.

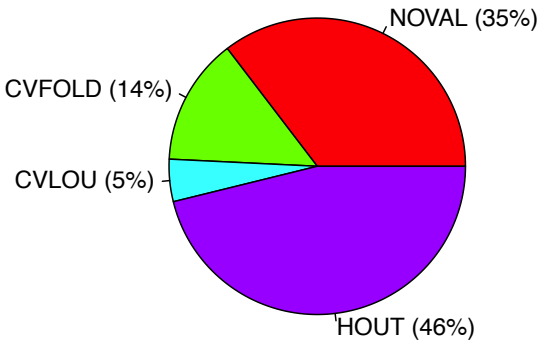


Figura 4.10: Tipos de validación utilizados en los métodos analizados.

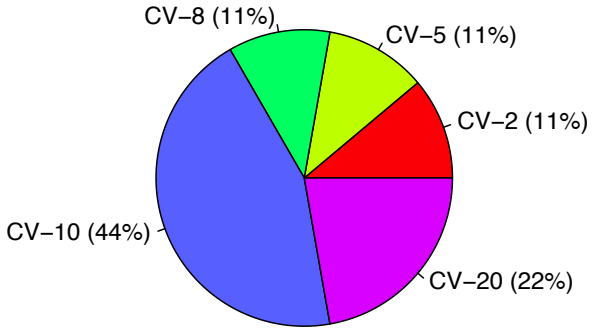


Figura 4.11: Número de bolsas empleadas en la validación CVFOLD.

4.5. Métodos según aproximación biológica

Existen tres grandes tipos de aproximaciones al problema de la PEP, desde un punto de vista biológico: *ab initio*, métodos de homologías y *threading*. En las siguientes subsecciones se detallan sus principales características.

4.5.1. Métodos *ab initio* (ABI)

Los métodos *ab initio* (o *de novo*) tratan de construir modelos de estructuras proteínicas utilizando únicamente datos de entrada procedentes de secuencias de proteínas, sin emplear como plantillas proteínas con

estructuras conocidas. Los métodos *ab initio* se basan generalmente en principios físico-químicos, energéticos o evolutivos, en lugar de utilizar directamente estructuras resueltas previamente [Zhou et al., 2011]. Todos los datos de entrada descritos en el epígrafe 4.2 han sido utilizados por los métodos *ab initio* que se han estudiado, incluidos los datos que contienen rasgos estructurales.

Este tipo de métodos suelen requerir amplios recursos computacionales y, por lo tanto, sólo han sido llevados a la práctica para pequeños conjuntos de proteínas de longitud relativamente corta. Su fiabilidad disminuye con el tamaño de la proteína y tradicionalmente han funcionado bien con secuencias menores a 150 aminoácidos [Zhang, 2008].

La principal ventaja que tienen los métodos *ab initio* es que sólo se necesita la secuencia como información de partida, de modo que, en principio, es posible modelar proteínas que corresponden a plegamientos no conocidos. Se han etiquetado con ABI los métodos analizados cuya aproximación biológica es *ab initio*.

4.5.2. Métodos de homologías (HOM)

La idea principal de los métodos de homologías (modelado por homología, *comparative modelling*) descansa en el hecho de que todas las parejas de proteínas que presentan una identidad de secuencia mayor al 30 % tienen estructura tridimensional similar [Sander and Schneider, 1991]. De este modo se puede construir el modelo tridimensional de una proteína de estructura desconocida, partiendo de la semejanza de secuencia con proteínas de estructura conocida [Lee, 1992, Blundell et al., 1987].

Esto se debe a que las estructuras proteínicas están evolutivamente más conservadas que sus secuencias de aminoácidos. Una secuencia objetivo puede ser modelada con una precisión razonable sobre una plantilla relacionada muy distante, siempre que la relación entre objetivo y plantilla sea perceptible en el alineamiento de sus secuencias [Ginalski, 2006].

La principal dificultad en el modelado por homología proviene más de las dificultades en el alineamiento que de los errores en la predicción de la estructura, dado un buen alineamiento [Zhang and Skolnick, 2005]. Por consiguiente, el modelado por homología es más preciso cuando el objetivo y la plantilla tienen secuencias similares. Se han etiquetado con HOM los métodos analizados que son de homologías.

4.5.3. Métodos de *threading* (THR)

Cuando la similitud entre la secuencia objetivo y la plantilla es demasiado baja no es posible realizar un buen alineamiento y no se puede aplicar con éxito el modelado por homología. En estos casos es posible obtener información estructural de la proteína empleando las técnicas de *threading*.

Los métodos de *threading*, reconocimiento del plegamiento o plegamiento inverso consisten en situar la secuencia problema en diferentes plegamientos conocidos y evaluar cómo se adapta o encaja en cada uno de ellos [Cao et al., 2004, Rost et al., 1997]. Esta adaptación o encaje es definido de diferentes formas, según el algoritmo de *threading*, por ejemplo, coincidencia de estructura secundaria o de ambientes parecidos a como se encuentran en las estructuras reales, etcétera.

El primer paso llevado a cabo es contrastar la secuencia de aminoácidos de una estructura desconocida contra una base de datos de estructuras resueltas. En cada caso, se utiliza una función de puntuación para evaluar la compatibilidad de la secuencia a la estructura, obteniéndose así posibles modelos tridimensionales. Se han etiquetado con THR los métodos analizados cuya aproximación biológica es *threading*.

4.5.4. Resumen

En la figura 4.12 se resumen los tipos de aproximación biológica de los métodos que se han analizado. Como se puede apreciar, la gran mayoría son de tipo *ab initio*. Además, en las tablas 4.5, 4.6 y 4.7 se muestra el tipo de aproximación biológica (en las columnas: ABI, HOM, y THR) de cada uno de los 65 métodos analizados, en orden cronológico descendente.

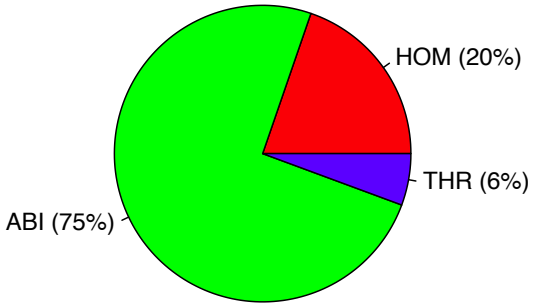


Figura 4.12: Tipos de aproximación biológica de los métodos analizados.

4.6. Métodos según aproximación algorítmica

A continuación se describen los métodos que se han analizado en el presente capítulo, clasificándolos por tipo de aproximación algorítmica. Se

han incluido cinco tipos de aproximaciones, que son las más empleadas en la literatura, y una categoría adicional para otros tipos de algoritmos.

Dentro de cada tipo de aproximación, se detalla, para cada método, el tipo de evaluación y validación que emplea, así como qué medida utiliza y qué valor tiene. En concreto, en las tablas 4.8, 4.9, 4.10, 4.11, 4.12 y 4.13 se consignan los siguientes datos:

- Validación: tipo de validación llevada a cabo por el método, tal como se explicó en el epígrafe 4.4.3.
- Ranking: tipo de ranking que efectúa el método, correspondiente a lo explicado en el apartado 4.4.2, aplicable sólo en métodos de tipo CMAP.
- MinSep: mínima separación en la secuencia que utiliza el método, tal como se explicó en el epígrafe 4.4.2, aplicable también sólo en métodos de tipo CMAP.
- Umbral: valor del umbral de contacto que utiliza el método, según lo visto en 4.4.2, aplicable igualmente sólo en métodos de tipo CMAP.
- Fuente: base de datos de donde el método ha extraído las proteínas que utiliza:
  - PDB [Berman et al., 2000]
  - PFAM [Punta et al., 2012]
  - CULLPDB [Wang and Dunbrack, 2003]
  - PDBSELECT [Griep and Hobohm, 2010]
  - EVA [Rost and Eyrich, 2001]
  - PDBTM [Kozma et al., 2013, Raman et al., 2006]
  - REPRDB [Noguchi and Akiyama, 2003]
  - SCOP [Conte et al., 2000]
  - ASTRAL [Chandonia et al., 2004]
  - CASP [Moult et al., 2011]
- NumProts: número de proteínas que utiliza el método. Cuando en el artículo que describe el método se especifica el tamaño de entrenamiento y de test, se ha consignado la suma de ambos tamaños. Cuando el artículo sólo especifica el número de proteínas de test, es éste el que se indica aquí.
- Medida: tipo de medida de evaluación que utiliza el método: ACC se refiere a la precisión definida en la ecuación 3.7, COV se refiere

a la sensibilidad definida en la ecuación 3.6, RMSD se definió en la ecuación 4.1, GDT_TS fue definida en la ecuación 4.2, SIML se refiere a un criterio de similitud estructural propio definido y utilizado únicamente en el método [Gupta et al., 2005a], F1 (F-measure) es la media armónica de la precisión y la sensibilidad y CORR es el coeficiente de correlación lineal de Pearson. Cuando el artículo proporciona varias medidas, se ha consignado preferentemente ACC o, en su defecto, RMSD o COV, que son las más utilizadas, para favorecer la comparación entre los métodos.

- Valor: valor de la medida de evaluación, que representa el resultado del método.

Se ha indicado un guión (-) en dichas tablas cuando el artículo que describe el método no proporciona dicha información.

4.6.1. Métodos estadísticos (STAT)

Los métodos estadísticos están basados en el análisis de comportamientos recurrentes en secuencias y estructuras y en la comparación de los resultados de este análisis con las secuencias objetivo. Generalmente, cuando se desean obtener datos estadísticos de secuencias, los análisis se realizan mediante alineamientos múltiples de secuencias. Cuando se pretenden obtener datos estadísticos de estructuras, se recuperan fragmentos estructurales de proteínas resueltas experimentalmente.

También tienen cabida en esta categoría aquellos métodos que están basados en el consenso estadístico de las predicciones de estructura terciaria o cuaternaria de otros métodos externos. En total se han analizado 7 métodos estadísticos, los cuales se describen a continuación.

El método [Burkoff et al., 2013] predice contactos entre las hebras β de una proteína, tanto paralelas como antiparalelas. Para ello, utiliza una medida de correlación entre las distribuciones de aminoácidos procedentes del MSA de la proteína. Para evitar los falsos positivos debidos a las conexiones indirectas entre aminoácidos correlados que no se encuentran en contacto, este método resuelve un problema de optimización consistente en maximizar la entropía en las distribuciones de aminoácidos del MSA.

[Ekeberg et al., 2013] predice contactos entre todos los aminoácidos de una proteína. Para ello, utiliza una medida de correlación entre las distribuciones de aminoácidos procedentes del MSA de la proteína. Para evitar los falsos positivos debidos a las conexiones indirectas entre aminoácidos correlados que no se encuentran en contacto, este método resuelve un problema de búsqueda de máxima verosimilitud aproximada (*pseudolikelihood maximization*) en el marco de un modelo Potts, el cual es un caso específico del modelo de Ising.

El método [Miyazawa, 2013] predice también contactos entre todos los aminoácidos de una proteína. Para ello, utiliza varias medidas de correlación entre las distribuciones de aminoácidos procedentes del MSA de la proteína. Para evitar los falsos positivos debidos a las conexiones indirectas entre aminoácidos correlados que no se encuentran en contacto, este método resuelve el problema de maximización de la entropía. La novedad que aporta este método radica en que, aparte de la correlación procedente de las sustituciones de aminoácidos en el MSA, utiliza correlaciones parciales a partir de cambios de propiedades físico-químicas de dichos aminoácidos. En concreto, utiliza volumen, carga, hidrofobicidad, capacidad de formación de puentes de hidrógeno, propensiones de participar en láminas β y giros, capacidad de interacción aromática, cadena lateral ramificada y capacidad de enlace cruzado.

[Eickholt et al., 2011] realiza un consenso entre varias aproximaciones externas. El método recupera varios modelos de otros predictores (SVMCon, TASSER and ROSSETA) y complementa la información a partir de varios procedimientos que mejoran la predicción de contactos entre residuos.

[Zhang, 2009] desarrolla un servidor de predicción denominado I-TASSER, el cual está basado en un predictor basado en homologías con estructuras de proteínas conocidas. Este método estuvo en la primera posición del ranking de métodos en las ediciones de CASP7, CASP8 y CASP9. Realiza un alineamientos de perfiles de características y dispone de un programa de refinado. Utiliza modelos ocultos de Markov (HMM) y simulaciones Monte Carlo durante el proceso de predicción.

El método [Karplus, 2009], denominado SAM-T08, utiliza HMM y proporciona al modelo tridimensional otra información como los alineamientos múltiples de secuencias, predicción de características estructurales locales, listas de plantillas candidatas de estructura conocida, alineamientos con dichas plantillas y predicciones de contactos entre residuos.

[Shi et al., 2008] propone varias aproximaciones para predecir grados de contactos a partir de secuencias de aminoácidos. Su primera aproximación está basada en una combinación lineal de datos de predicción de estructuras secundarias y composición de aminoácidos. Una segunda aproximación está basada en la similitud de secuencias cuyas estructuras se encuentran resueltas.

En la tabla 4.8 se muestra el tipo de validación, los criterios de evaluación y el resultado obtenido en cada uno de los métodos estadísticos que se han explicado.

Método	Validación	Ranking	MinSep	Umbral	Fuente	NumProts	Medida	Valor
[Burkoff et al., 2013]	CV10F	LX2	5	8	PDB	916	ACC	0.61
[Ekeberg et al., 2013]	NOVAL	NOLX	5	8.5	PFAM	17	COV	0.78
[Miyazawa, 2013]	NOVAL	NOLX	6	5	PFAM	15	ACC	0.66
[Eickholt et al., 2011]	HOUT	LX5	24	8	CASP9	26	ACC	0.30
[Karplus, 2009]	HOUT	LX5	24	8	PDB	21	RMSD	3.82
[Zhang, 2009]	HOUT	-	-	-	CASP8	164	RMSD	4.24
[Shi et al., 2008]	NOVAL	NOLX	0	6	PDB	933	CORR	0.87

Tabla 4.8: Métodos estadísticos.

4.6.2. Métodos de redes neuronales artificiales (ANN)

Las redes neuronales aportan un alto grado de flexibilidad al sistema de predicción. Además de la codificación de los datos de entrada procedentes de los pares de aminoácidos, se pueden incluir neuronas con información adicional. En contrapunto, las redes neuronales tienen ciertas limitaciones. Por ejemplo, la limitación en la codificación de los datos de entrada, el uso de parámetros adecuados en la red neuronal y el sobreaprendizaje. En total se han analizado 16 métodos en esta categoría, los cuales se describen a continuación.

[Aydin et al., 2012] realiza una predicción de los ángulos de torsión de la proteína usando perfiles PSSM. Discretiza los ángulos de torsión en 5 intervalos, con lo que el problema que resuelve es de clasificación. La información de salida es un vector de valores discretos para cada aminoácido de la secuencia objetivo. Incorpora al sistema ROSSETA su predicción de ángulos de torsión y mejora los resultados que ofrecía ROSSETA.

[Eickholt and Cheng, 2012] utiliza una red neuronal profunda y una técnica propia de Boosting. Dado que los perfiles de características que genera son extensos, hace uso de los recursos de la GPU y CUDA. Crea perfiles especiales para tres diferentes rangos de separación entre residuos a predecir: corto, medio y largo alcance.

[Di Lena et al., 2012] utiliza tres fases en cascada. Cada fase refina los contactos predichos en la fase anterior. La primera fase utiliza una red neuronal 2D para predecir contactos de forma somera basándose en estructura secundaria. La segunda etapa del método utiliza una segunda red neuronal basada en funciones de energía, afinando en los pares de aminoácidos de hélices α y láminas β. Finalmente, en la última fase se emplea una pila de redes neuronales profundas con diferente parametrización, teniendo en cuenta criterios de espacio y tiempo.

El método [Dorn and N. de Souza, 2010] realiza una predicción de ángulos de torsión para luego formar un modelo 3D de proteína. Utiliza una red neuronal con varias capas intermedias cuyos datos de entrada están formados por subcadenas de caracteres (n-gramas). Estas subcadenas se corresponden con fragmentos de secuencias de estructuras conocidas, a las cuales se les asignan valores de estructura secundaria y ángulos de torsión reales. La salida de la red neuronal son los ángulos de torsión predichos con los que se compone un modelo 3D final.

[Wolff et al., 2010] es un método que combina características ab-initio con procedimientos clásicos de threading. Genera modelos 3D a partir de proteínas con estructuras conocidas. Para seleccionar las estructuras que mejor se adaptan a la secuencia objetivo, se utiliza el perfil de conectividad efectiva (ECP), que es una caracterísitca estructural basada en los autovectores y autovalores de las matrices de contactos de las proteínas.

[Walsh et al., 2009] introduce una nueva clase de restricciones de

distancias: los mapas de contactos multi-clase. Desarrolla dos predictores de mapas con cuatro clases. Uno de ellos basado en técnicas *ab initio*, que incluye información evolutiva. El segundo está basado en plantillas, el cual proporciona nuevos datos de entrada basados en información de homología con estructuras conocidas.

El método [Xue et al., 2009], denominado SPINE-2D, está formado por dos redes neuronales compuestas por una y dos capas, respectivamente. Estas redes usan 34 características de entrada, incluyendo PSSM, 7 propiedades físico-químicas que incluyen hidrofobicidad, volumen y polarizabilidad. También emplea datos de estructura secundaria.

En el método [Lippi and Frasconi, 2009] se propone una arquitectura híbrida basada en redes neuronales y lógica de Markov empleando ponderaciones específicas del dominio, con el objetivo de predecir contactos entre residuos pertenecientes a láminas β. Como entrada se han usado secuencias alineadas, estructura secundaria y accesibilidad al solvente en dos estados.

[Tegge et al., 2009] fue uno de los métodos más precisos según el ranking de CASP8. Este método realiza su cometido en dos fases. En la primera, una red neuronal recursiva 2D-RNN predice un mapa de contactos. En la segunda fase, una red neuronal predice la conformación especial de láminas β.

[Shackelford and Karplus, 2007] predice contactos entre residuos utilizando una red neuronal. Emplea información de mutaciones correlacionadas, pero no hace un DCA, con lo que incurre en numerosos falsos positivos. Realiza un muestreo aleatorio de ejemplos de la clase mayoritaria para conseguir un mayor balance en la clase. Entrena la red neuronal con un conjunto de proteínas y realiza predicciones sobre las proteínas de CASP7.

El método [Vullo et al., 2006] introduce un predictor basado en el consenso de dos redes neuronales recursivas de dos capas. El método clasifica las componentes del autovector principal y utiliza información predicha de estructura secundaria y escalas de interacción hidrofóbica.

En [Liu et al., 2006] se desarrolló un red neuronal transitoria caótica cuya topología está formada por tres capas de neuronas. La capa de salida representa una probabilidad de contacto. Una capa oculta contiene diez neuronas. Dispone de una capa de entrada con diferentes números de neuronas dependiendo de la cantidad de información codificada. Este método utiliza 1050 neuronas para 5 pares de residuos, 10 neuronas para la clasificación de residuos según su hidrofobicidad, polaridad, acidez-basicidad y 6 neuronas para información de estructura secundaria.

[Punta and Rost, 2005] es un método que combina diferentes fuentes de información de propiedades de proteínas, características biofísicas, perfiles evolutivos, estructura secundaria e información de alineamiento. Este método está centrado en la mejora de las predicciones a largo alcance.

[Zhang et al., 2005] es un método basado en un red neuronal de base radial. En particular se introducen algoritmos genéticos para la optimización

de la parametrización de entrada. Es un método para la predicción de mapas de contactos que incluye un esquema de codificación binario. Este modelo genera una correspondencia no lineal multivariada que resulta de utilidad para problemas de clasificación no lineal con múltiples parámetros y modelos, como es el caso de la predicción de mapas de contactos de proteínas.

[Zhang and Huang, 2004a] introduce un algoritmo genético para optimizar la función de base radial y los centros ocultos de una red neuronal de base radial. Su codificación binaria es empleada para entrenar la red para la predicción de patrones de contactos entre residuos.

El método [Zhang and Huang, 2004b] predice distancias entre aminoácidos utilizando una red neuronal, cuya parametrización es optimizada previamente por un algoritmo genético. Es el primer y único artículo encontrado que predice mapas de distancias de proteínas. Utiliza información de estructura secundaria predicha para la predicción. El algoritmo entrena con 39 proteínas y realiza predicción sobre los 41 primeros aminoácidos de una proteína independiente, mostrando únicamente el resultado de esta predicción. Para evaluar la calidad de la predicción, discretiza en tres intervalos los valores de distancias entre aminoácidos hasta 30 angstroms (de 0 a 10, de 10 a 20 y de 20 a 30).

En la tabla 4.9 se muestra el tipo de validación, los criterios de evaluación y el resultado obtenido en cada uno de los métodos de redes neuronales que se han explicado.

Método	Validación	Ranking	MinSep	Umbral	Fuente	NumProts	Medida	Valor
[Aydin et al., 2012]	HOUT	-	-	-	CULLPDB	5199	ACC	0.84
[Di Lena et al., 2012]	CV10F	LX5	24	8	CASP9	2356	ACC	0.31
[Eickholt and Cheng, 2012]	HOUT	LX5	24	8	CASP9	111	ACC	0.29
[Dorn and N. de Souza, 2010]	NOVAL	-	-	-	PDB	5	RMSD	1.92
[Wolff et al., 2010]	NOVAL	-	-	-	CASP8	29	RMSD	6.4
[Lippi and Frasconi, 2009]	CV10F	NOLX	24	8	PDB	916	ACC	0.47
[Tegge et al., 2009]	HOUT	LX5	24	8	CASP8	116	ACC	0.31
[Walsh et al., 2009]	HOUT	LX5	24	8,13,19	CASP7	93	F1	0.11
[Xue et al., 2009]	HOUT	LX5	24	8	CASP7	82	ACC	0.26
[Shackelford and Karplus, 2007]	HOUT	LX10	24	8	CASP7	93	ACC	0.58
[Liu et al., 2006]	CVLOU	NOLX	5	8	PDBSELECT	125	COV	0.08
[Vullo et al., 2006]	HOUT	LX5	24	8	CASP6	11	ACC	0.19
[Punta and Rost, 2005]	HOUT	LX2	24	8	EVA	1847	ACC	0.20
[Zhang et al., 2005]	HOUT	NOLX	0	8	PDB	53	COV	0.27
[Zhang and Huang, 2004a]	HOUT	NOLX	0	8	PDB	173	COV	0.29
[Zhang and Huang, 2004b]	HOUT	-	-	-	PDB	40	COV	0.80

Tabla 4.9: Métodos de redes neuronales artificiales.

4.6.3. Métodos de máquinas de soporte vectorial (SVM)

Las máquinas de soporte vectorial (SVM) están basadas en la transformación de un espacio de entrada en un espacio de características (denominado kernel) de mayor dimensionalidad. Es en este espacio donde se lleva a cabo el proceso de aprendizaje. Se construye un hiperplano que busca maximizar el margen entre las diferentes clases. En total se han analizado 7 métodos en esta categoría, los cuales se describen a continuación.

El método [Savojardo et al., 2013] predice contactos de tipo puente disulfuro entre los aminoácidos cisteína. Para ello, utiliza medidas que recogen las mutaciones correlacionadas presentes en las distribuciones de aminoácidos del MSA de la proteína. Para evitar los falsos positivos debidos a las conexiones indirectas entre aminoácidos correlados que no se encuentran en contacto, este método estima la covarianza inversa y dispersa. A esta covarianza se añade la información mutua encontrada en las mutaciones correlacionadas. Una vez obtenidas ambas medidas, se agregan al perfil junto a características globales, individuales y por entornos, incluido el PSSM de los aminoácidos. Finalmente se realizan las predicciones con estos perfiles mediante support vector regression (aplicación de SVM para clase continua) utilizando validación cruzada.

[Savojardo et al., 2011] predice contactos entre cisteínas en proteínas de células eucariotas. Para realizar las predicciones de contactos utiliza predicciones externas de localización subcelular de las proteínas. Ésta es una información poco utilizada como entrada, pero que parece aportar gran valor predictivo. El método utiliza una variante de modelo oculto de Markov y regresión de soporte vectorial sobre los perfiles evolutivos y de localización subcelular.

[Wu et al., 2011] está compuesto por un conjunto de nueve predictor de contactos basados en SVM, utilizados con simulación de I-TASSER en combinación con restricciones de contactos basadas en plantillas dispersas. En este método se utiliza la función de energía original de I-TASSER y las predicciones de contactos generadas por versiones extendidas de SVMSEQ [Wu and Zhang, 2008].

[Lo et al., 2009] es un esquema jerárquico de predicción de contactos con aplicación en proteínas de membrana. Este método está compuesto en dos niveles. En el primer nivel, los contactos entre residuos son predichos a partir de las secuencias. En el segundo, las relaciones entre los pares de aminoácidos son finalmente predichas. Se combina el análisis estadístico de propensiones de contactos con los perfiles evolutivos, de accesibilidad relativa al solvente y características helicoidales.

El método [Cheng and Baldi, 2007] es un esquema de predicción de contactos basado en máquinas de soporte vectorial que utiliza un amplio perfil de características acerca de los pares de aminoácidos, su estructura secundaria, accesibilidad relativa al solvente, potenciales de contactos.

Además emplea características del entorno de los aminoácidos.

[Sander et al., 2006] es un método basado en homologías para la predicción de contactos entre residuos. Genera una base de datos de clusters de secuencias con estructuras parecidas. Para ello utiliza una medida de similitud de estructura y el algoritmo k-means de clustering. Emplea propiedades físico-químicas e información evolutiva para realizar una validación cruzada con tres métodos de aprendizaje automático: árboles de decisión (C5.0), máquinas de soporte vectorial (SVM) y bosques aleatorios (RF).

En el método [Han et al., 2005] se ha desarrollado un esquema de reconocimiento del plegamiento basado en máquinas de soporte vectorial. El alineamiento entre la secuencia objetivo y la plantilla es transformado en un vector de características del mismo orden de longitud que la plantilla en cuestión, el cual es evaluado por el componente SVM. La salida del algoritmo es una probabilidad de la relación entre la secuencia objetivo y la plantilla.

En la tabla 4.10 se muestran el tipo de validación, los criterios de evaluación y el resultado obtenido en cada uno de los métodos de SVM que se han explicado.

Método	Validación	Ranking	MinSep	Umbral	Fuente	NumProts	Medida	Valor
[Savojardo et al., 2013]	CV20F	NOLX	24	8	PDB	1797	ACC	0.66
[Savojardo et al., 2011]	CV20F	NOLX	0	8	PDB	1797	ACC	0.51
[Wu et al., 2011]	HOUT	LX5	24	8	CULLPDB	164	ACC	0.14
[Lo et al., 2009]	CVLOU	LX5	-	6	PDBTM	66	ACC	0.44
[Cheng and Baldi, 2007]	HOUT	LX5	24	8	CASP7	498	ACC	0.13
[Sander et al., 2006]	CV5F	NOLX	0	8	PDBSELECT	27	RMSD	1.19
[Han et al., 2005]	HOUT	-	-	-	PDB	2854	COV	0.46

Tabla 4.10: Métodos de máquinas de soporte vectorial.

4.6.4. Métodos de computación evolutiva (EC)

Los algoritmos evolutivos son métodos de optimización y búsqueda de soluciones basados en los postulados de la evolución biológica. En ellos se mantiene un conjunto de individuos que representan posibles soluciones, las cuales se cruzan y compiten entre sí. De tal manera que las soluciones más aptas son capaces de prevalecer a lo largo de las generaciones, evolucionando el sistema hacia la producción de mejores soluciones.

Los algoritmos evolutivos son utilizados principalmente en problemas con espacios de búsqueda extensos y no lineales, en donde otros métodos no son capaces de encontrar soluciones en un tiempo razonable. En total se han analizado 13 métodos en esta categoría, los cuales se describen a continuación.

[Bacardit et al., 2012] genera reglas de decisión mediante un algoritmo evolutivo para la predicción de contactos. Balancea la clase en los ejemplos de entrenamiento utilizando un muestreo aleatorio estratificado, conservando una proporción de 2:1 entre no-contactos y contactos, respectivamente. Posteriormente realiza un análisis de las reglas generadas en términos de los atributos más frecuentes que aparecen.

El método [Calvo et al., 2011] está basado en un algoritmo genético multiobjetivo, denominado Pitagoras-PSP. Este algoritmo utiliza una aproximación *ab initio* que predice el esqueleto de la proteína y los ángulos de torsión, empleando para ello una función de energía en su función objetivo, en cuyo caso es necesario minimizar. Los operadores de mutación mantienen ángulos de torsión en rangos de valores adecuados según la estructura secundaria de la cadena de aminoácidos.

[Chen and Li, 2010] está compuesto por clasificadores basados en algoritmos genéticos para la predicción de contactos a largo alcance. Dichos contactos se encuentran a una separación en la secuencia de al menos 24 residuos. Este método incorpora centros de perfiles de secuencia. Éstos representan un vector que codifica un par de residuos que pertenecen al mismo rango de contactos o no-contactos.

[Hoque et al., 2010] utiliza una búsqueda primero en profundidad sobre los perfiles de atributos de proteínas, para evitar que los clásicos operadores de cruce produzcan individuos no factibles. El coste computacional de dicha búsqueda es lineal y, por tanto, muy rápida. La codificación de los individuos representa un clásico modelo de dos o tres dimensiones de hidrofobicidad-polaridad de los aminoácidos de las secuencias proteínicas. Tras la ejecución del algoritmo genético, los individuos finales, los cuales representan modelos hidrofobicidad-polaridad 2D o 3D, son extrapolados a modelos 3D de proteínas con todos sus átomos mediante herramientas externas. El artículo no proporciona ninguna medida de evaluación que permita conocer el grado de ajuste del modelo predicho con la estructura real, debido a ello no se ha incluido ningún valor de resultado de este método.

[Zhang et al., 2010b] es un algoritmo genético para la predicción de ángulos de torsión. Este método introduce una búsqueda tabú en el operador de mutación del algoritmo genético que mejora la capacidad de búsqueda local. Usa un modelo off-lattice AB (modelo matemático con función de energía propia de una secuencia de monómeros). Los individuos del algoritmo genético incluyen los ángulos de torsión de la cadena de aminoácidos representada por el modelo off-lattice AB. El tamaño de la población puede variar de una generación a otra y se ha adopta esta decisión para mantener la diversidad de la población. El operador de mutación es un operador de búsqueda tabu. En dicha búsqueda se descartan los individuos que empeoren (aumenten) la energía del sistema. El artículo expone los valores de energía finales obtenidos en la experimentación, pero éstos no miden por sí mismos el grado de ajuste de los modelos 3D predichos a las estructuras reales, por tanto no se ha incluido ningún valor experimental obtenido con este método.

El método [Islam and Chetty, 2009] está basado en un algoritmo memético. Un algoritmo memético es un algoritmo evolutivo que incorpora una fase de búsqueda local de diferentes tipos. En concreto, este método utiliza un modelo de hidrofobicidad-polaridad y calcula la función de fitness con dos nuevos parámetros denominados requisito H y requisito P. El requisito H mide como de compacto se encuentra un residuo localizado en el núcleo hidrofóbico del centro de la proteína. El requisito P mide cómo de cerca se encuentra el residuo al resto de lugares del modelo de rejilla.

[Judy et al., 2009] está basado en un algoritmo evolutivo multiobjetivo, el cual representa las estructuras de proteínas mediante sus ángulos de torsión. El método supone una modificación del algoritmo clásico multiobjetivo introduciendo algunas variaciones. El algoritmo utiliza probabilidades adaptativas de cruce y mutación.

El método [Zhang and Han, 2007] emplea en primer lugar un algoritmo evolutivo que está basado en una representación de 19 bits para una proteína, donde los bits 0-8 representan cada posible par de aminoácidos. Los bits 9-12 representan una clasificación de los residuos (polar, no-polar, ácido o base). Los bits 13-15 representan cual es la estructura secundaria del residuo, entre hélice, lámina o giro. Los bits 16-17 representan la longitud de la secuencia y los bits 18-19 representan la separación en la secuencia. Este algoritmo evolutivo se utiliza para mejorar la función de base radial de una red neuronal de predicción de contactos entre residuos.

En el método [Cutello et al., 2006] se ha utilizado una estrategia evolutiva basada en el frente de Pareto con el objetivo de explorar el espacio de búsuqeda conformacional de las interacciones de mínima energía entre átomos con y sin enlaces. El método utiliza una representación de modelos de ángulos de torsión y una función de energía CHARMM como función de fitness.

El método [Gupta et al., 2005a] comienza con una población inicial de mapas de contactos aleatorios para una secuencia de aminoácidos

determinada. En la función de fitness se utiliza una red neuronal y cuatro propiedades físicas: carga, distancia en la secuencia, vecindad hidrofóbica y grado de los vértices. El mapa de contactos más preciso es seleccionado tras la última generación. Posteriormente, este mapa de contactos es comparado con un mapa de contactos plantilla para cada plegamiento usando teoría de grafos.

[MacCallum, 2004] utiliza mapas de autoorganización (*self-organizing maps* [Kohonen and Mäkisara, 1989]) en un algoritmo genético para la predicción de contactos entre los aminócidos de las proteínas.

En el método [Shi et al., 2004] se propone un análisis de características multiobjetivo y algoritmo de selección para la resolución del reconocimiento de plegamiento de proteínas.

[Cotta, 2003] incluye un modelo de hidrofobicidad-polaridad mediante una implementación de modelo de rejilla tridimensional. Este método adopta una función de fitness basada en una función que tiene valor 1 si las variables son iguales y 0 en caso contrario. También está basada en la distancia entre los aminoácidos objetivo. Además incluye el solapamiento del entorno aminoacídico y las energías de contacto libres de los pares de aminoácidos.

En la tabla 4.11 se muestran el tipo de validación, los criterios de evaluación y el resultado obtenido en cada uno de los métodos evolutivos que se han explicado.

Método	Validación	Ranking	MinSep	Umbral	Fuente	NumProts	Medida	Valor
[Bacardit et al., 2012]	HOUT	LX5	24	8	CASP9	28	ACC	0.21
[Calvo et al., 2011]	NOVAL	-	-	-	CASP8	4	RMSD	9.16
[Chen and Li, 2010]	CV2F	LX5	24	8	REPRDB	480	ACC	0.21
[Hoque et al., 2010]	NOVAL	-	-	-	PDB	5	-	-
[Zhang et al., 2010b]	NOVAL	-	-	-	PDB	4	-	-
[Islam and Chetty, 2009]	NOVAL	-	-	-	PDB	6	-	-
[Judy et al., 2009]	NOVAL	-	-	-	PDB	1	RMSD	4.23
[Zhang and Han, 2007]	HOUT	NOLX	14	8	PDB	61	-	-
[Cutello et al., 2006]	NOVAL	-	-	-	PDB	4	RMSD	2.16
[Gupta et al., 2005a]	NOVAL	NOLX	3	6	PDB	24	SIML	0.73
[MacCallum, 2004]	HOUT	LX2	24	8	CASP5	15	ACC	0.14
[Shi et al., 2004]	CV8F	NOLX	0	8	PDBSELECT	698	ACC	0.87
[Cotta, 2003]	NOVAL	-	-	-	PDB	8	-	-

Tabla 4.11: Métodos de computación evolutiva.

4.6.5. Métodos de razonamiento basado en casos (CBR)

El razonamiento basado en casos es un paradigma de razonamiento automático donde las experiencias son representadas como casos en una base de casos ocurridos. Esta base de casos es utilizada para el razonamiento llevado a cabo en la fase de predicción. Un caso representa un conocimiento acerca de un problema en particular e incluye parte de la descripción del mismo.

Los métodos de razonamiento basado en casos son particularmente útiles en dominios poco conocidos o en evolución, donde el conocimiento es difícil de formalizar. El razonamiento basado en casos está basado en la premisa de que problemas similares tienen soluciones similares.

Dentro de este grupo de métodos se encuentra el paradigma de los vecinos más cercanos, el cual ha sido utilizado en el propuesta que se presenta en el capítulo 5. En total se han analizado 4 métodos en esta categoría, los cuales se describen a continuación.

[Abu-Doleh et al., 2012] predice mapas de contactos de proteínas y está basado en varias etapas de ejecución paralela con sistemas de inferencia neuro-fuzzy y vecinos más cercanos. Usa 5 ventanas de aminoácidos. Realiza selección de atributos mediante la factorización de matriz no negativa (NMF) al 25 %. Selecciona el mejor valor de K que maximiza la precisión sobre el conjunto de entrenamiento. Para asegurar un comportamiento natural en la conectividad entre aminoácidos se realiza un proceso de filtrado de la estructura. Este filtro está basado en ventanas de aminoácidos, cuyo tamaño es seleccionado por un sistema experto.

El método [Colubri et al., 2006] analiza la mejora predictiva debido a la simplificación de la representación de la estructura de una proteína. Esta simplificación está basada en la eliminación de átomos no significativos de la molécula y la orientación habitual de los ángulos de torsión. Es un método basado en homologías con estructuras conocidas. El criterio de selección de dichas estructuras está basado en una función que depende de propensiones estadísticas de los distintos tipos de aminoácidos y sus estructuras secundarias. La búsqueda de las estructuras más similares se basa en la minimización de dicha función mediante un algoritmo de *simulated annealing*.

[Davies et al., 2006] es un método CBR jerárquico, en el sentido de que considera los mapas de contactos de proteínas a diferentes niveles de complejidad estructural. En una organización *bottom-up*, el método construye motivos de estructura secundaria utilizando el mapa de contactos y conocimiento geométrico de contactos entre hélices α. Se adaptan las estructuras conocidas alineándolas, en una búsqueda de las estructuras más similares, a las secuencias con estructuras desconocidas.

El método [Glasgow et al., 2006] propone un predictor de mapas de contactos basado en casos. La base de casos está representada y organizada

por nombre de proteína, secuencia, asignación de estructura secundaria, clase estructural y tipo de mapa de contactos. La solución consiste en un modelo tridimensional del esqueleto de la proteína obtenido a partir del mapa de contactos. El método considera únicamente proteínas con hélices α y deriva medidas de similitud para la comparación de mapas de contactos de proteínas resueltas y mapas de contactos predichos.

En la tabla 4.12 se muestran el tipo de validación, los criterios de evaluación y el resultado obtenido en cada uno de los métodos CBR que se han explicado.

Método	Validación	Ranking	MinSep	Umbral	Fuente	NumProts	Medida	Valor
[Abu-Doleh et al., 2012]	HOUT	LX5	24	8	EVA	500	ACC	0.34
[Colubri et al., 2006]	HOUT	-	-	-	PDB	50	RMSD	6.73
[Davies et al., 2006]	HOUT	NOLX	0	10	PDB	422	RMSD	1.24
[Glasgow et al., 2006]	HOUT	NOLX	0	10	PDB	100	RMSD	1.86

Tabla 4.12: Métodos de razonamiento basado en casos.

4.6.6. Otras aproximaciones algorítmicas (OAP)

Además de los métodos que se han explicado en los epígrafes anteriores, existen otras propuestas que no se ajustan a las aproximaciones algorítmicas más utilizadas. En este apartado abordaremos dichas propuestas, las cuales están basadas principalmente en el esquema de random forests, en la optimización lineal entera y en la covarianza inversa y dispersa. En total se han analizado 17 métodos en esta categoría, los cuales se describen a continuación.

El método [Jones et al., 2012] introduce por primera vez una estimación de la covarianza inversa y dispersa al problema de la predicción de mapas de contactos de proteínas. Los datos utilizados únicamente proceden de las mutaciones correlacionadas detectadas a partir de los alineamientos múltiples de secuencias. De este modo resuelve el problema del DCA, comentado anteriormente en este capítulo.

[Nugent and Jones, 2012] realiza una predicción de modelos 3D de proteínas transmembrana basada en una predicción doble de estructura secundaria y en mutaciones correlacionadas. Corrige los falsos positivos de dichas correlaciones mediante la covarianza inversa dispersa, como en el método [Jones et al., 2012]. No utiliza ninguna información estadística ni estructural adicional. Su precisión depende del número de secuencias homólogas alineadas en el MSA del target. En primer lugar predice contactos, posteriormente se ensamblan fragmentos estructurales y finalmente el modelo 3D final es refinado mediante herramientas externas.

[Sułkowska et al., 2012] realiza un MSA y resuelve el problema del DCA. Una vez obtenidos los contactos predichos, los acompaña de información de predicción de estructura secundaria, para incluirlos en un modelo basado en la estructura. Este modelo distingue dos componentes: una para interacciones locales y otra para no locales. A partir de este modelo, se asignan ángulos de torsión a los aminoácidos en contacto. Finalmente se genera un modelo 3D y se refina para ser comparado con las estructuras determinadas experimentalmente.

El método [Ashkenazy et al., 2011] combina datos estructurales a partir de diferentes plantillas con el objetivo de mejorar la predicción de mapas de contactos de proteínas no homólogas. El uso de varias plantillas mejora la predicción de los contactos y puede ser de utilidad para revelar nuevos tipos de conformaciones estructurales.

[Li et al., 2011] desarrolla un conjunto de modelos de árboles de decisión mediante el esquema de *random forests*. De este modo combina las predicciones de cada uno de los árboles de decisión aprendidos y produce la probabilidad de contacto final, la cual es almacenada en el mapa de contactos de la proteína. El método también utiliza una matriz de propensiones estadísticas, entre otros datos de entrada.

[Marks et al., 2011] utiliza un modelo de máxima entropía para resolver

el problema del DCA procedente de las mutaciones correlacionadas del MSA sobre un conjuntos de secuencias de proteínas homólogas.

El método [Morcos et al., 2011] predice contactos entre los aminoácidos de una proteína. Para ello, analiza las mutaciones correlacionadas de aminoácidos procedentes del MSA de la proteína. Para evitar los falsos positivos debidos a las conexiones indirectas entre aminoácidos correlados que no se encuentran en contacto, este método resuelve el problema de maximización de la entropía, y lo hace de una forma más eficiente que otros métodos anteriores. No genera ningún modelo 3D, sólo los mapas de contactos.

[Wang et al., 2011] predice contactos globales y entre las hélices α de proteínas transmembrana. Construye largos perfiles basados en información evolutiva y realiza luego una selección de atributos con CFS y BestFirst. Finalmente, utiliza el algoritmo de Random Forests implementado en R. Como inconveniente, los índices de mutaciones correlacionadas que utiliza (usa tres índices) no resuelven el problema del DCA, con lo que se cometen mas falsos positivos.

El método [Wei and Floudas, 2011] incorpora un modelo de optimización matemático especializado en la predicción de contactos de proteínas α transmembrana. Incorpora restricciones físicas en el modelo matemático y realiza predicciones de contactos sobre dos conjuntos diferentes de proteínas.

[Yang and Chen, 2011] es un método de predicción de mapas de contactos basado en regresión logística que realiza un consenso entre varios sistemas predictores externos. De este modo, realiza predicciones de contactos basadas en la evaluación de la regresión logística sobre las probabilidades de contactos proporcionadas por los predictores.

[Rajgaria et al., 2010] es un método de optimización lineal entera de predicción de contactos en proteínas de las familias estructurales: β, $\alpha+\beta$ y α/β. Se centra en la predicción de contactos entre hebras beta utilizando las hidrofobicidades de los aminoácidos. Resuelve un problema de optimización sujeto a restricciones, en concreto la minimización de una función de energía basada en estructura secundaria e hidrofobicidad.

[Wang et al., 2010] es una aproximación multinivel de combinación para la mejora de diversas tareas de predicción, combinando y complementando plantillas estructurales, alineamientos y modelos. Este método incorpora cinco servidores de predicción de estructuras de proteínas y un predictor humano.

El método [Björkholm et al., 2009] presenta una aproximación basada en los modelos ocultos de Markov y utiliza datos de secuencias homólogas como entrenamiento. Además emplea estructura secundaria predicha y una librería de vecindarios local con descriptores estructurales.

[Gao et al., 2009] es un método de consenso para la predicción de contactos entre aminoácidos basado en el modelo de programación lineal entera. Evalúa la correlación utilizando una estimación de máxima

verosimilitud y extrae datos predictivos de servidores utilizando análisis de componentes principales.

El método [Rajgaria et al., 2009] está basado en una optimización lineal entera que utiliza distancias de alta resolución dependientes de campos de fuerza para calcular la energía de interacción entre los diferentes residuos de una proteína. Este método predice aminoácidos hidrofóbicos en proteínas de tipo α.

[Shell et al., 2009] es un método de predicción de estructuras 3D basado únicamente en modelos de energía. Estos modelos tienen en cuenta los campos de fuerza y leyes físicas aplicadas a todos los átomos de la molécula. Es un método altamente costoso computacionalmente, ya que además utiliza modelos de solvatación combinados con simulaciones de dinámica molecular. Finalmente los modelos son refinados mediante una técnica que mimetiza las rutas físicas del plegamiento proteínico.

[Wolff et al., 2008] predice estructuras 3D mediante una aproximación de threading. La función de puntutación del threading está basada en la característica estructural denominada perfil de conectividad efectiva (ECP). En la búsqueda de la mejor plantilla estructural utiliza un algoritmo estocástico, el algoritmo de Monte Carlo. En concreto, esta búsqueda persigue minimizar una función de energía basada en los perfiles anteriormente citados. Se relacionan el número de contactos entre pares de aminoácidos de la proteína objetivo y la plantilla, para determinar el grado de ajuste conseguido. Finalmente el modelo 3D es generado y se mide su RMSD.

En la tabla 4.13 se muestran el tipo de validación, los criterios de evaluación y el resultado obtenido en cada uno de los métodos explicados en este apartado.

Método	Validación	Ranking	MinSep	Umbral	Fuente	NumProts	Medida	Valor
[Jones et al., 2012]	NOVAL	LX5	24	8	PFAM	150	ACC	0.62
[Nugent and Jones, 2012]	NOVAL	NOLX	24	8	PDB	28	RMSD	8.48
[Sułkowska et al., 2012]	NOVAL	NOLX	6	8	PDB	15	RMSD	5.37
[Ashkenazy et al., 2011]	HOUT	LX1	24	8	CASP8	65	ACC	0.69
[Li et al., 2011]	HOUT	LX5	24	8	CASP9	28	ACC	0.21
[Marks et al., 2011]	NOVAL	NOLX	-	8	PDB	15	RMSD	3.62
[Morcos et al., 2011]	NOVAL	NOLX	24	8	PFAM	131	COV	0.55
[Wang et al., 2011]	CVLOU	LX5	0	8	PDBTM	62	ACC	0.49
[Wei and Floudas, 2011]	HOUT	NOLX	0	14	PDB	65	ACC	0.60
[Yang and Chen, 2011]	CV10F	LX5	24	8	CASP9	80	ACC	0.3
[Rajgaria et al., 2010]	HOUT	NOLX	24	8	SCOP	687	ACC	0.57
[Wang et al., 2010]	HOUT	-	-	-	CASP8	120	GDT_TS	63
[Björkholm et al., 2009]	HOUT	LX5	24	8	ASTRAL	4013	ACC	0.23
[Gao et al., 2009]	NOVAL	LX5	24	8	CASP7	86	ACC	0.23
[Rajgaria et al., 2009]	HOUT	NOLX	24	8	PDBSELECT	11	ACC	0.77
[Shell et al., 2009]	NOVAL	-	-	-	CASP7	6	RMSD	5.9
[Wolff et al., 2008]	NOVAL	NOLX	0	8.5	PDB	1507	RMSD	5.8

Tabla 4.13: Otras aproximaciones algorítmicas.

4.6.7. Resumen

En la figura 4.13 las aproximaciones algorítmicas de los métodos que se han analizado. Además, en las tablas 4.5, 4.6 y 4.7 se muestra el tipo de aproximación algorítmica (en las columnas STAT, ANN, SVM, EC, CBR y OAP) de cada uno de los 65 métodos analizados, en orden cronológico descendente.

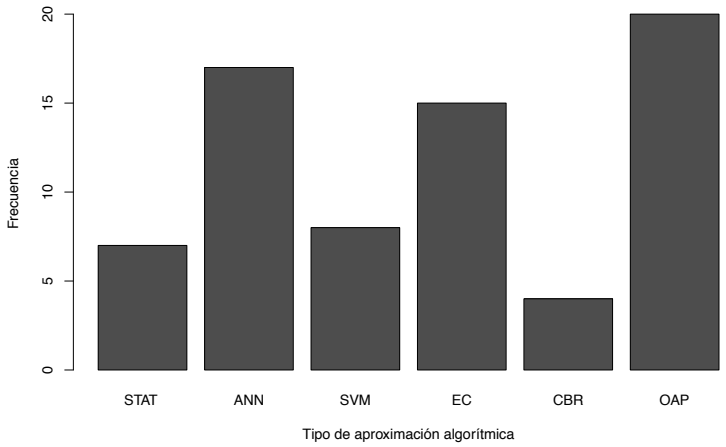


Figura 4.13: Aproximaciones algorítmicas de los métodos analizados.

4.7. Resumen

Para concluir el capítulo resumimos en la figura 4.14 las bases de datos que se han utilizado en los métodos analizados. Como se puede apreciar, la base de datos más utilizada es PDB, como se adelantó en el epígrafe 2.3.1. De todos modos, las proteínas del resto de bases de datos proceden en última instancia también de proteínas publicadas en PDB.

En la figura 4.15 se muestra la densidad de proteínas que se han utilizado en los métodos publicados en tres intervalos de tiempo (de 2003 a 2007, de 2008 a 2010 y de 2011 a 2013). Se aprecia cierta tendencia al uso de mayor número de proteínas en los métodos más modernos (línea roja), en buena parte debido a la disponibilidad de mayores recursos computacionales.

Con el objetivo de visualizar la evolución de los resultados obtenidos por los métodos de predicción de estructura terciaria, se ha procedido a comparar únicamente los métodos que realizan el mismo tipo de evaluación sobre la misma fuente de proteínas. De este modo, en la figura 4.16 se ha representado cronológicamente la precisión obtenida por los 11 métodos que

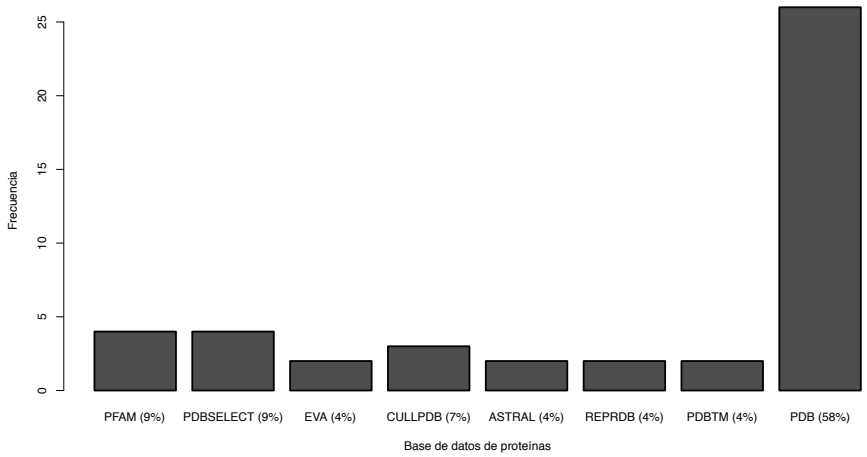


Figura 4.14: Bases de datos de proteínas utilizadas en los métodos analizados.

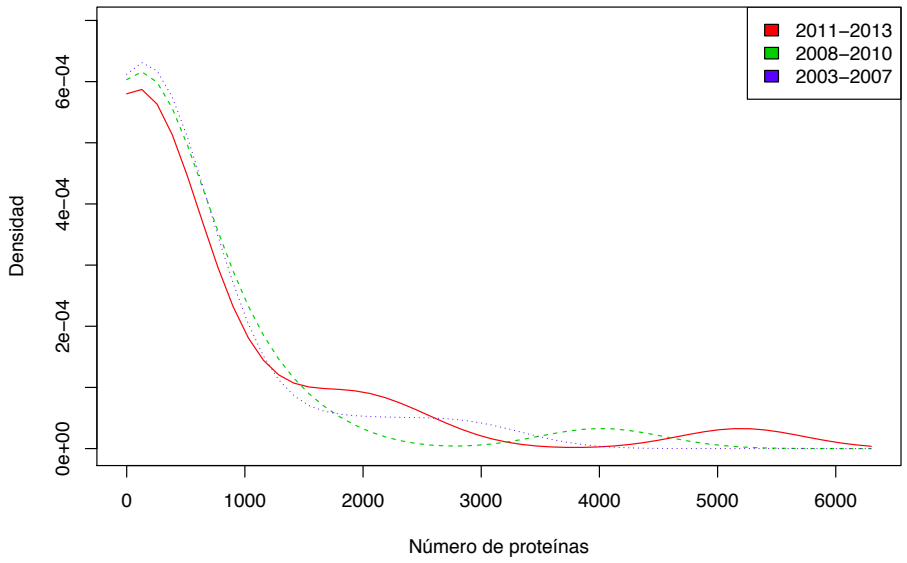


Figura 4.15: Número de proteínas utilizadas por intervalos anuales.

satisfacen la misma evaluación, consistente en LX5, MS24, 8 angstroms y proteínas de CASPx ($x \in \{7, 8, 9\}$).

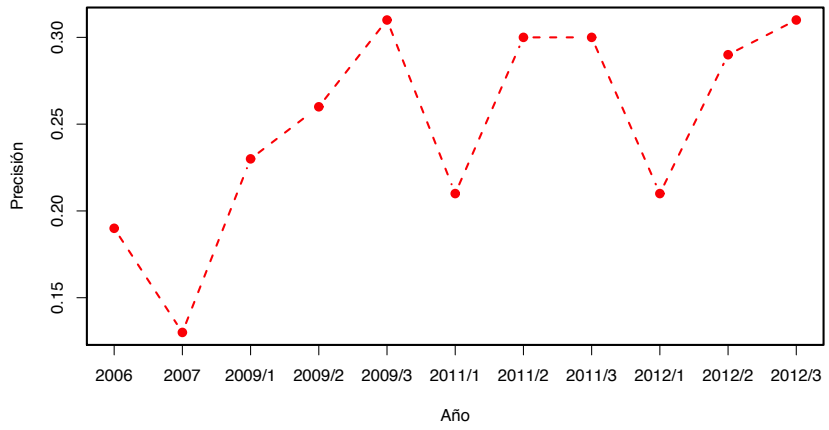


Figura 4.16: Precisión obtenida en proteínas de CASP7/8/9 evaluada con LX5, MS24 y umbral 8 angstroms.

La tendencia actual en la literatura de este campo pasa por el uso de mutaciones correlacionadas (CMU) realizando análisis de emparejamientos directos (DCA) [Marks et al., 2012]. Además, cada vez más los métodos se centran en predecir determinadas partes del problema en lugar de tratar de ser generales o universales, debido a la gran dificultad del problema. Ejemplo de ello son los métodos de predicción de contactos entre hebras β [Burkoff et al., 2013], entre hélices α [Wang et al., 2011, Wei and Floudas, 2011], predicción de contactos únicamente entre cisteínas [Savojardo et al., 2013, Savojardo et al., 2011] y predicción de tipos concretos de proteínas, como las transmembrana [Nugent and Jones, 2012].

Parte III

Propuestas

Capítulo 5

Predictor de mapas de distancias basado en similitud de propiedades de aminoácidos (ASPpred)

5.1. Introducción

En este capítulo se describe la propuesta de predicción de estructuras terciarias de proteínas realizada. Se trata de una aplicación denominada ASPpred, cuyo nombre hace alusión a un predictor de mapas de distancias basado en similitud de propiedades de aminoácidos (en inglés Aminoacid Subsequences Properties-based distance map PREDictor).

ASPpred está desarrollado en base a tres subsistemas. El primero de ellos se denomina ASPFgen, cuyo nombre significa generador de ficheros de propiedades de subsecuencias de aminoácidos. El segundo subsistema se denomina ASPnn, cuyo nombre significa vecinos más cercanos sobre propiedades de subsecuencias de aminoácidos. Por último, el tercer subsistema se denomina DMeval, cuyo nombre hace referencia a evaluador de mapas de distancias.

ASPFgen es una aplicación desarrollada en Java capaz de extraer propiedades físico-químicas de los aminoácidos de un conjunto de proteínas y producir una serie de ficheros en el formato ARFF de Weka [Hall et al., 2009] útiles para una posterior predicción de estructura de proteínas.

ASPnn es una aplicación también desarrollada en Java que utiliza los ficheros generados por ASPFgen y es capaz de predecir la estructura de una proteína mediante la generación de un mapa de distancias, estructura definida en el epígrafe 4.3.3. ASPnn emplea vecinos más cercanos como técnica de predicción (regresión).

DMeval es un aplicación escrita en Java que permite evaluar la calidad

de las predicciones calculando diversas medidas de evaluación a partir de un mapa de distancias como entrada.

Estos tres subsistemas han sido diseñados para funcionar de forma encadenada. En concreto, los resultados del primer subsistema pueden ser la entrada para el segundo, y los resultados de éste la entrada del tercero. No obstante, el primer subsistema, ASPFgen, también ha sido diseñado para servir de fuente de información para una predicción realizada por cualquier otro método externo de clasificación o regresión.

En el epígrafe 5.2 se describe la metodología de trabajo con la propuesta y sus subsistemas. En el epígrafe 5.3 se proporciona una visión global de las tareas llevadas a cabo por el sistema ASPpred. En último lugar, en los epígrafes 5.4, 5.5 y 5.6, se detallan todas las tareas realizadas por cada uno de los subsistemas.

5.2. Metodología de trabajo de ASPpred

Para realizar una predicción de estructuras de proteínas con ASPpred es necesario disponer de los archivos en el formato de la base de datos PDB, descargables de su página Web. En cada uno de estos archivos se encuentra, entre otros datos, la estructura de la proteína, la cual ha sido determinada experimentalmente en el laboratorio.

Por otra parte, el sistema ASPpred también necesita conocer cual es el conjunto de propiedades físico-químicas de aminoácidos que se desea utilizar para llevar a cabo la predicción. Existen varios conjuntos de propiedades que se han estudiado y que ofrecen las mejores tasas de sensibilidad y precisión. Estos conjuntos se brindan al usuario como predefinidos, aunque se pueden construir y utilizar otros de forma personalizada.

ASPpred utiliza una configuración de inicio que debe ser determinada y que sirve para especificar el comportamiento de ciertos aspectos del proceso de predicción. Entre los parámetros de configuración de inicio más relevantes se encuentra K, que determina el número de vecinos más cercanos a utilizar para las predicciones de proteínas.

El resultado que produce ASPpred consiste, por una parte, en las estructuras predichas para las proteínas incógnita (mediante mapas de distancias) y, por otra parte, en varias medidas de evaluación que permiten conocer la calidad de la predicción obtenida.

El procedimiento general empleado por ASPpred se muestra en la figura 5.1. Como se ha mencionado anteriormente, se parte de un subconjunto de proteínas descargadas del repositorio Protein Data Bank. También se parte de un subconjunto de propiedades físico-químicas que pueden proceder del repositorio AAindex.

El primer subsistema en actuar es ASPFgen, el cual toma sendos conjuntos de proteínas y propiedades y produce 400 ficheros ASPF. Los

ficheros ASPF son archivos de texto plano con el formato ARFF de Weka y contienen una linea de texto por cada subsecuencia analizada. El contenido de los archivos ASPF se describe con detalle en el epígrafe 5.4.4.

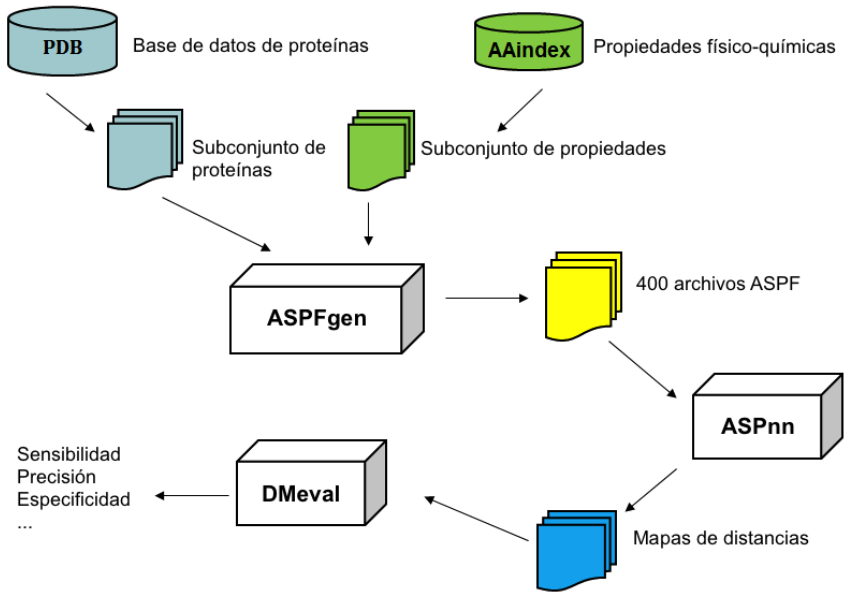


Figura 5.1: Procedimiento global del sistema ASPpred

A continuación el subsistema ASPnn tomará los 400 archivos ASPF y los utilizará para la predicción. Para cada proteína de entrada al sistema, ASPnn generará un mapa de distancias como resultado de las predicciones.

Por último, el subsistema DMeval calcula varias medidas de evaluación a partir de las predicciones obtenidas en los mapas de distancias. Las medidas que produce son: error relativo medio, sensibilidad, precisión, exactitud, especificidad, F-measure y MCC. Dicha evaluación se realiza mediante la comparación entre la estructura real de cada proteína y la predicha mediante el sistema.

En la tabla 5.1 se enumera la configuración de inicio utilizada por el sistema ASPpred, la cual es posible establecer a través de una línea de comandos. El sistema ASPpred posee también una configuración avanzada que se describe en las tablas 5.2 y 5.3, en la cual es posible indicar ciertos aspectos adicionales del funcionamiento global del sistema predictor.

5.2.1. Generación de datos con ASPFgen

Un primer escenario de uso del sistema ASPpred se tiene cuando se desea probar diversos métodos de predicción de carácter general, como algoritmos evolutivos, métodos de generación de árboles de decisión, generación de

Parámetro	Significado
proteins	Ruta de los archivos PDB de proteínas de entrada.
k	Número de vecinos más cercanos empleados en la predicción de distancias entre aminoácidos.
features	Elementos del vector de predicción que describe a un par de aminoácidos. Se explica en el epígrafe 5.4.3.
properties	Ruta y nombre del fichero de propiedades de aminoácidos en formato de AAindex.

Tabla 5.1: Parámetros básicos de configuración de ASPpred

Parámetro	Significado
num-threads	Número de hilos de ejecución que se desean utilizar.
alfa	Porcentaje de la longitud o número de aminoácidos de cada subsecuencia que se desea tomar desde cada extremo hacia el interior de la misma.
beta	Porcentaje de la longitud o número de aminoácidos de cada subsecuencia que se desea tomar desde cada extremo hacia el exterior de la misma.
validation	Método de validación que se realiza. Puede tomar estos valores: "none" para no realizar predicciones, "cv-Nf" para realizar una validación cruzada con N bolsas, "cv-lou" para realizar una validación cruzada dejando uno fuera (*leaving-one-out*) y "hout" para realizar predicciones sobre un conjunto de proteínas independiente.
minsep	Mínima separación en la secuencia (ver 4.4.2).
ranking-toplx	Tipo de ranking TopLx (ver 4.4.2). "NO" para no realizar ranking, 1 para realizar un ranking Top L, 2 para Top L/2 y 5 para Top L/5.
atom	Atomo tomado como referencia para calcular distancias entre aminoácidos.
which-pairs	Dado un par de aminoácidos de test, indica si se usan todos los pares de aminoácidos de training o solo aquellos del mismo tipo.

Tabla 5.2: Parámetros avanzados de configuración de ASPpred (1/2).

Parámetro	Significado
mean	Tipo de media utilizada al producir la predicción a partir de las clases de los vecinos más cercanos (están disponibles la aritmética, armónica, geométrica, cuadrática e híbrida). La híbrida permite indicar un valor umbral de separación en la secuencia, de manera que si el par de aminoácidos de test se encuentran a menor separación se aplica un tipo de media, y en caso contrario otro.
grouping	Indica si los vecinos más cercanos son agrupados por cercanía según cada atributo. Esto es, si se tienen N atributos, se encuentran los K vecinos más cercanos por cada atributo (en total KN vecinos).
dmap-post	Se realiza un posprocesado de las matrices de distancias que detecta y corrige elementos que incumplen la factibilidad geométrica. Para ello se evalúan las desigualdades triangulares en los valores de distancias.
k-study	Sirve para realizar un estudio del impacto del valor del parámetro K en los resultados. Para ello, ASPpred realiza predicciones con $K \in \{1, 2, 3, ..., K_{max}\}$ y devuelve una tabla que incluye la sensibilidad, precisión, etcétera. para cada valor de K.

Tabla 5.3: Parámetros avanzados de configuración de ASPpred (2/2).

reglas o redes neuronales, en el problema de la predicción de estructura de proteínas.

Para probar un algoritmo de predicción sobre el problema de la predicción de estructura de proteínas mediante ASPFgen, se necesita que el algoritmo sea capaz de hacer predicciones sobre clase continua; o bien sobre etiqueta discreta, previa discretización. Además, el algoritmo debe aceptar el formato ARFF de Weka para los datos de entrada. Cualquier algoritmo de clasificación o regresión incluido en la herramienta Weka que acepte clase continua puede utilizarse.

Una vez seleccionadas y descargadas las proteínas que formarán el conjunto de estudio sobre el cual aprender y realizar predicciones, se invoca a la aplicación ASPpred a través de una línea de comandos. En la tabla 5.4 se muestra la configuración necesaria para este escenario.

Nótese que se solicita la ejecución del componente ASPFgen para generar los ficheros para la predicción posterior, pero no se solicita la ejecución de

Parámetro	Valor
validation	"none"
proteins	Ruta de los archivos PDB de proteínas de entrada.
k	No se utiliza.
features	Elementos del vector de predicción que describe a un par de aminoácidos. Se explica en el epígrafe 5.4.3.
properties	Ruta y nombre del fichero de propiedades de aminoácidos en formato de AAindex.

Tabla 5.4: Configuración para el escenario ASPFgen.

ASPnn (validation = "none") puesto que se va a utilizar un algoritmo de predicción externo.

5.2.2. Predicción mediante vecinos más cercanos con ASPnn

Otro escenario de uso del sistema ASPpred se tiene cuando se desea predecir proteínas con estructura desconocida, una vez ya validado el método de predicción. En este contexto el objetivo es obtener una estructura de datos que defina una estructura de proteína representable en un computador: en nuestro caso, un mapa de distancias.

En este escenario se precisan dos conjuntos de proteínas: uno que contiene las proteínas que proporcionan conocimiento para la predicción, y otro formado por las proteínas incógnita para las que se desea conocer su estructura.

Para utilizar la aplicación ASPpred bajo este escenario se necesita indicar la configuración que se describe en la tabla 5.5. Nótese que en este caso intervienen tanto el subsistema ASPFgen como ASPnn y DMeval. Con ello se obtiene también una evaluación de la calidad de las predicciones efectuadas.

Parámetro	Valor
validation	"hout"
proteins	Ruta de PDB de proteínas de entrenamiento.
test_proteins	Ruta de PDB de proteínas de test.
k	Número de vecinos más cercanos.
features	Elementos del vector de predicción.
properties	Fichero de propiedades de aminoácidos AAindex.

Tabla 5.5: Configuración para el escenario ASPnn.

El resultado de la ejecución consiste en un conjunto de archivos de mapas de distancias situados en la misma localización del fichero ejecutable del programa. Estos archivos son de texto separado por comas. En este escenario se genera ficheros de mapas de distancias que contienen las predicciones de estructura de todas las proteínas incógnita.

A partir del mapa de distancias de una proteína se puede obtener mucha información sobre su estructura y la predicción efectuada. Además, es posible recrear una imagen tridimensional de la proteína, lo cual es de suma utilidad para muchos estudios biológicos.

5.2.3. Evaluación de mapas de distancias con DMeval

Otro escenario de uso del sistema ASPpred puede darse cuando se desea perfeccionar el propio sistema de predicción o ajustar de forma óptima los parámetros de los que depende. En este contexto, se realizan experimentos con el mismo conjunto de proteínas y diferentes configuraciones, para comprobar cuál es la que produce mejores resultados.

Para evaluar el sistema de predicción, ASPpred emplea la validación cruzada mediante bolsas o mediante la técnica leaving-one-out.

En este escenario se precisa un único conjunto de proteínas, el cual contiene las proteínas que proporcionan conocimiento a la predicción, y con las cuales se prueba a predecir; para finalmente comparar las predicciones con las estructuras reales conocidas.

Para utilizar la aplicación ASPpred bajo este escenario se necesita indicar la configuración que se describe en la tabla 5.6. Nótese que en este caso intervienen nuevamente todos los subsistemas.

Parámetro	Valor
validation	"cv-10f" o "cv-lou"
proteins	Ruta de PDB de proteínas de entrada.
k	Número de vecinos más cercanos.
features	Elementos del vector de predicción.
properties	Fichero de propiedades de aminoácidos AAindex.

Tabla 5.6: Configuración para el escenario DMeval.

El resultado de la ejecución, al igual que en el escenario anterior, consiste en un conjunto de archivos de mapas de distancias situados en la misma localización del fichero ejecutable del programa. En este escenario se generan tantos ficheros como bolsas en la validación cruzada (con cv-10f, 10 ficheros; con cv-lou, 1 fichero). Cada archivo de mapas de distancias contiene las predicciones de estructura de todas las proteínas incógnita de cada bolsa.

5.3. El sistema ASPpred

En esta sección se proporciona una visión global de las tareas que realiza cada subsistema para llevar a cabo el proceso completo de predicción de las estructuras de proteínas. Existen ocho tareas fundamentales que se llevan a cabo en dichos subsistemas, las cuales se enumeran en la tabla 5.7.

ASPFgen	Tarea 1. Lectura de archivos PDB de proteínas (training) Tarea 2. Extracción de subsecuencias Tarea 3. Creación de vectores de predicción Tarea 4. Clasificación y almacenamiento de vectores
ASPnn	Tarea 1. Lectura de archivos PDB de proteínas (test) Tarea 2. Extracción de subsecuencias Tarea 3. Creación de vectores de predicción Tarea 5. Búsqueda del vector de predicción más parecido Tarea 6. Asignación y almacenamiento de predicciones
DMeval	Tarea 7. Obtención del error de predicción Tarea 8. Obtención de sensibilidad y otras medidas

Tabla 5.7: Tareas del sistema ASPpred.

Nótese que tanto el subsistema ASPFgen como ASPnn realizan las tareas 1, 2 y 3. No obstante, la tarea 1 utiliza proteínas de training en el subsistema ASPFgen y proteínas de test en ASPnn.

La entrada al sistema consiste en dos carpetas de archivos en formato PDB de proteínas. La primera carpeta contiene las proteínas de entrenamiento (training protein set), de las cuales se aprende. La segunda carpeta contiene las proteínas de test (test protein set), con las cuales se predice. La salida del sistema consiste, por una parte, en mapas de distancias por cada proteína de test y, por otra parte, en valores de las distintas medidas de evaluación.

En las siguientes tres secciones (5.4, 5.5 y 5.6) se explican con detalle las tareas realizadas por cada uno de los tres subsistemas.

5.4. Subsistema ASPFgen

5.4.1. Tarea 1. Lectura de archivos PDB de proteínas

La lectura de archivos PDB es necesaria para tomar los datos de las proteínas, los cuales se utilizarán posteriormente en la predicción. En el caso del subsistema ASPFgen, se leen proteínas de training; en el caso de ASPnn, se leen proteínas de test. Los datos que se toman son las secuencias

de aminoácidos y las distancias relativas entre cada aminoácido y todos los demás dentro de cada secuencia.

Cada archivo PDB contiene una única proteína. Desde el punto de vista de las estructuras de datos del sistema, una proteína está formada por una o varias secuencias. Cada secuencia está formada por una serie ordenada de aminoácidos, en forma de cadenas de caracteres, y por una matriz cuadrada de números reales que contiene valores de distancia. Cada aminoácido está representado mediante una letra y existen 20 aminoácidos posibles.

Todo archivo con el formato PDB incluye un conjunto de líneas que comienzan por la palabra reservada "ATOM". Estas líneas son las utilizadas por el sistema ASPpred. Cada línea ATOM especifica la posición en el espacio de un átomo de un aminoácido de una secuencia concreta de una proteína. A partir del conjunto de líneas ATOM es posible construir la estructura de datos que el sistema contempla para una proteína.

La matriz o mapa de distancias de una secuencia es una matriz cuadrada de orden N, donde N es el número de aminoácidos que tiene dicha secuencia. El elemento (i, j) con $i < j$ de la matriz de distancias es la distancia medida en angstroms (Å) observada entre el aminoácido i-ésimo y el j-ésimo dentro de la secuencia. Para medir las distancias se utiliza un átomo de referencia, el cual es configurable mediante la opción "atom", como se explicó en la tabla 5.2. Los átomos de referencia más utilizados son el carbono alfa (CA) y el carbono beta (CB).

De este modo, se obtienen las tres coordenadas espaciales del átomo de referencia de todos los N aminoácidos de una secuencia. A continuación se calculan las $N(N-1)/2$ distancias euclídeas entre los N aminoácidos. Con estas distancias se rellena la triangular superior de la matriz de distancias contenida en la estructura de datos que el sistema guarda para la secuencia.

En la figura 5.2 se muestra un ejemplo artificial de fragmento de archivo PDB en el que se tiene una secuencia con cuatro aminoácidos. En la figura 5.3 se muestran las estructuras de datos que se obtienen tras leer el fragmento de archivo PDB.

Cada línea ATOM especifica las columnas que se explican a continuación. La primera columna indica que es un registro ATOM, la segunda representa un número de serie para cada átomo, la tercera es la representación del átomo utilizado (por ejemplo CA es carbono alfa, N es nitrogeno y C un carbono). La cuarta columna indica el símbolo de tres letras del aminoácido en cuestión. La quinta columna es el nombre o símbolo de la secuencia actual. La sexta columna indica el número de orden en la cadena del aminoácido al cual pertenece dicho átomo. Las columnas séptima, octava y novena indican las coordenadas ortogonales del átomo en los ejes, x, y z respectivamente medidos en angstroms (Å). Las columnas décima y undécima, reflejan los valores de ocupancia y el factor de temperatura. Por último, la duodécima columna representa el símbolo químico del átomo.

```
ATOM 1  N  LYS A 1   4.889  13.266  43.405  1.00 59.70   N
ATOM 2  CA LYS A 1   4.481  12.039  44.080  1.00 55.95   C
ATOM 3  C  LYS A 1   5.447  11.691  45.235  1.00 56.42   C
ATOM 4  N  ALA A 2   6.748  11.849  44.996  1.00 66.43   N
ATOM 5  CA ALA A 2   7.766  11.495  45.989  1.00 65.28   C
ATOM 6  C  ALA A 2   7.732  12.422  47.219  1.00 63.65   C
ATOM 7  N  VAL A 3   5.026  14.070  48.368  1.00 58.00   N
ATOM 8  CA VAL A 3   3.844  13.743  49.153  1.00 52.09   C
ATOM 9  C  VAL A 3   4.096  12.495  50.021  1.00 58.95   C
ATOM 10 N  CYS A 4   5.827  12.370  48.168  1.00 58.00   N
ATOM 11 CA CYS A 4   3.844  13.743  49.153  1.00 52.09   C
ATOM 12 C  CYS A 4   4.096  12.495  50.021  1.00 58.95   C
```

Figura 5.2: Un fragmento de archivo PDB.

```
Proteina {
   Secuencia 1 {
      Cadena de aminoácidos = "LKVC"
      Mapa de distancias =
            L     K         V         C
        L   0.0   3.838154  5.389313  7.121544
        K   0.0   0.0       3.917833  5.472211
        V   0.0   0.0       0.0       3.840412
        C   0.0   0.0       0.0       0.0
   }
}
```

Figura 5.3: Estructura de datos de proteína obtenida.

Cada aminoácido se representa por un símbolo de una o tres letras. En los archivos PDB se usan tres letras para representar los aminoácidos, mientras que en el sistema ASPpred se utiliza una letra. Los tres aminoácidos de la secuencia de la figura 5.2 son LYS (L), ALA (K), VAL (V) y CYS (C). De este modo, la secuencia obtenida es "LKVC". La lista completa de símbolos de una y tres letras de aminoácidos se mostraron en la tabla 2.1.

El mapa de distancias se obtiene al calcular las distancias entre todos los pares de átomos CA de aminoácidos posibles en la secuencia y almacenarlas en la triangular superior de la matriz. La diagonal principal tiene todos sus elementos a cero ya que de un átomo a él mismo hay distancia cero. La triangular inferior se encuentra también a cero y está reservada para almacenar las distancias predichas entre todos los pares de aminoácidos.

5.4.2. Tarea 2. Extracción de subsecuencias

La extracción de subsecuencias consiste en encontrar todas las posibles subcadenas, de una longitud mínima determinada, dentro de la cadena de aminoácidos obtenida en la tarea 1. Cada carácter de la cadena de aminoácidos, como se ha visto anteriormente, representa un aminoácido mediante su símbolo de una letra.

La longitud mínima utilizada en el sistema ASPpred viene determinada por el párametro *minsep*, que indica la separación mínima en la secuencia. En concreto, la longitud mínima de las subsecuencias de aminoácidos será $minsep + 2$. En la figura 5.4 se muestra la estructura de datos de la proteína tras la extracción de subsecuencias con longitud mínima 3 ($minsep = 1$) a partir de la secuencia obtenida en la figura 5.3.

```
Proteina {
   Secuencia 1 {
      Cadena de aminoácidos = "LKVC"
      Subsecuencias = "LKV", "LKVC", "KVC"
      Mapa de distancias =
            L     K         V          C
        L   0.0   3.838154  5.389313   7.121544
        K   0.0   0.0       3.917833   5.472211
        V   0.0   0.0       0.0        3.840412
        C   0.0   0.0       0.0        0.0
   }
}
```

Figura 5.4: Extracción de subsecuencias.

5.4.3. Tarea 3. Creación de vectores de predicción

A partir de las subsecuencias obtenidas en la tarea 2, se construye un vector para cada una de ellas, los cuales contienen información útil para la predicción de la estructura de la proteína. A estos vectores los denominamos vectores de predicción.

El subsistema ASPFgen soporta diferentes tipos de vectores de predicción, según lo especificado en el parámetro de configuración *features*. En esta Tesis se aportan dos tipos de vectores: tipo A y tipo B. Los resultados experimentales con el tipo A se abordan en el capítulo 6 y en el epígrafe 7.2. Los resultados con el tipo B se encuentran en el epígrafe 7.3. Ambos tipos de vectores de predicción se definen a continuación.

Para poder definir los tipos de vectores de predicción es necesario, en primer lugar, definir los siguientes elementos. En primer lugar, una secuencia

de aminoácidos de longitud L se define como $s_1 \dots s_L$. La subsecuencia $s_b \dots s_e$ se encuentra situada dentro de la secuencia $s_1 \dots s_b \dots s_e \dots s_L$, donde $s_b \dots s_e$ es la subsecuencia, s_b es el primer aminoácido de la subsecuencia, s_e es el último aminoácido de la misma y $1 \leq b < e \leq L$.

Además, las propiedades físico-químicas se definen con $P_1 \dots P_m$, donde m es el número de propiedades utilizadas. El conjunto de propiedades empleado es definido en el parámetro de configuración *properties*. El valor de la propiedad P_i de un aminoácido s_j se define como $P_i(s_j)$.

Las propiedades físico-químicas utilizadas proceden de una selección de atributos realizada sobre el repositorio AAindex. Según la experimentación realizada, se ha realizado una selección de atributos u otra. Las selecciones de atributos realizadas se abordarán en las experimentaciones concretas, las cuales se explican en los capítulos 6 y 7.

Una vez introducida esta nomenclatura, el vector de predicción tipo A se define formalmente en la figura 5.5. Como se puede apreciar, el vector contiene la longitud de la subsecuencia, los valores promedio de las propiedades físico-químicas de sus aminoácidos interiores y la distancia entre los extremos de la subsecuencia. En total, el vector tipo A tiene $m + 1$ atributos mas la clase (D).

L	$\overline{P}_1$	$\cdots$	$\overline{P}_m$	D
$(e-b+1)/L_{max}$	$\frac{1}{e-b+1}\sum_{i=b+1}^{e-1} P_1(s_i)$	$\cdots$	$\frac{1}{e-b+1}\sum_{i=b+1}^{e-1} P_m(s_i)$	$d(s_b, s_e)$

Figura 5.5: Vector de predicción tipo A para la subsecuencia $s_b \dots s_e$.

La longitud de cada subsecuencia (L) se almacena normalizada entre 0 y 1, para ello se divide la longitud de cada subsecuencia entre la longitud máxima de todas las proteínas del conjunto de training (L_{max}). La normalización es importante para que todos los atributos del vector de predicción estén en la misma escala y contribuyan por igual a la predicción.

Las propiedades $P_1 \dots P_m$ atribuibles a cada aminoácido interior a la subsecuencia se promedian y se almacenan en el vector de predicción ($\overline{P}_1 \dots \overline{P}_m$). Finalmente, se añade a cada vector la distancia real (D) entre los aminoácidos extremos (primero y último de la subsecuencia).

El vector de predicción tipo B se define formalmente en la figura 5.6. Como se puede apreciar, contiene dos atributos por cada propiedad (B_i y E_i, con $i \in \{1, \cdots, m\}$) y la distancia entre los aminoácidos extremos de la subsecuencia (D). B_i representa la distribución de la propiedad P_i en la secuencia completa con una ponderación que decrece según nos alejamos del primer aminoácido de la subsecuencia (s_b). E_i es análogo a B_i sólo que con respecto al último aminoácido de la subsecuencia (s_e). En total, el vector tipo B tiene $2m$ atributos mas la clase (D).

Para ilustrar con un ejemplo la construcción de los vectores de predicción,

B_1	$\cdots$	B_m	E_1	$\cdots$	E_m	D
$P_1(s_b) + \sum_{\substack{j=1 \\ j \neq b}}^{L} \frac{P_1(s_j)}{L\|b-j\|}$	$\cdots$	$P_m(s_b) + \sum_{\substack{j=1 \\ j \neq b}}^{L} \frac{P_m(s_j)}{L\|b-j\|}$	$P_1(s_e) + \sum_{\substack{j=1 \\ j \neq e}}^{L} \frac{P_1(s_j)}{L\|e-j\|}$	$\cdots$	$P_m(s_e) + \sum_{\substack{j=1 \\ j \neq e}}^{L} \frac{P_m(s_j)}{L\|e-j\|}$	$d(s_b, s_e)$

Figura 5.6: Vector de predicción tipo B para la subsecuencia $s_b \ldots s_e$.

trataremos a partir de ahora únicamente con el tipo A. Consideremos para este ejemplo tres propiedades físico-químicas reales de aminoácidos obtenidas de AAindex: P_1 es la propiedad *Amphiphilicity index*, P_2 es *Hydrostatic pressure asymmetry index* y P_3 es *Free energy of solution in water (kcal/mol)*.

En la tabla 5.8 se muestran los valores de estas propiedades para los cuatro aminoácidos que intervienen en la secuencia de ejemplo "LKVC". Tal como se aprecia en dicha tabla, los valores de las propiedades han sido normalizados, puesto que no lo están tal como se obtienen de AAindex.

Aminoácido	P_1	P_2	P_3
L	0,000	0,630	0,463
K	0,530	0,692	0,313
V	0,000	0,273	0,369
C	0,000	0,267	0,947

Tabla 5.8: Tres propiedades y sus valores en cuatro aminoácidos.

Una vez que se han tomado los valores de las propiedades de aminoácidos, se promedian para cada aminoácido interior a la subsecuencia. Por ejemplo, para la subsecuencia "LKVC"se promedian los valores de los aminoácidos K y V, tal como se ilustra en la figura 5.7.

	P_1	P_2	P_3
K	0,530	0,692	0,313
V	0,000	0,273	0,369
Promedios	0,265	0,482	0,341
	$\overline{P}_1$	$\overline{P}_2$	$\overline{P}_3$

Figura 5.7: Cálculo de valores medios de propiedades para aminoácidos interiores a la subsecuencia LKVC.

En la tabla 5.9 se muestran completos todos los vectores de predicción que se construyen a partir de la secuencia "LKVC"de ejemplo y de su mapa de distancias mostrado en la figura 5.4.

La longitud de secuencia máxima es 4, puesto que sólo hay una secuencia de training ("LKVC") y tiene cuatro aminoácidos. Por ello, las longitudes

	L	$\overline{P}_1$	$\overline{P}_2$	$\overline{P}_3$	D
LKV	0,75	0,530	0,692	0,313	5,389313
LKVC	1,00	0,265	0,482	0,341	7,121544
KVC	0,75	0,000	0,273	0,369	5,472211

Tabla 5.9: Los vectores de predicción de la secuencia LKVC de ejemplo.

normalizadas (L) de las subsecuencias son 0,75, 1,00 y 0,75, ya que son de tamaño 3, 4 y 3 respectivamente. El valor de distancia que se incluye en cada vector de predicción es el que se obtiene del mapa de distancias de la secuencia entre los dos aminoácidos extremos de cada subsecuencia.

5.4.4. Tarea 4. Organización y almacenamiento de vectores de predicción

Los vectores de predicción, una vez construidos, son organizados en función de sus dos aminoácidos extremos y almacenados en ficheros independientes. De este modo, al existir 20 aminoácidos posibles, los vectores de predicción se clasifican en 20 x 20 = 400 ficheros, según los aminoácidos extremos de cada vector. En la tabla 5.10 se muestra cómo se clasifican y en qué ficheros se almacenan los vectores de predicción obtenidos en la tarea anterior.

Subsecuencia	Inicio	Fin	Vector	Fichero
LKV	L	V	0,75 0,530 0,692 0,313 5,389313	L-V.aspf
LKVC	L	C	1,00 0,265 0,482 0,341 7,121544	L-C.aspf
KVC	K	C	0,75 0,000 0,273 0,369 5,472211	K-C.aspf

Tabla 5.10: Clasificación de subsecuencias y almacenamiento de vectores en ficheros

Los ficheros donde se almacenan los vectores de predicción se denominan ASPF (ficheros de propiedades de subsecuencias de aminoácidos) y tienen el formato ARFF de la herramienta Weka. En la figura 5.8 se muestra el contenido del archivo L-V.aspf que contiene el vector de predicción con extremos L y V anteriormente obtenido.

5.5. Subsistema ASPnn

Una vez obtenidos todos los archivos ASPF de las proteínas de training, y si el usuario desea realizar predicciones mediante ASPnn, se realizan nuevamente las tareas 1, 2 y 3 sobre todas las proteínas del conjunto de test.

```
% ASPF File
% Gualberto Asencio Cortes
%
% Protein set: prueba1 (1 files)
% Attribute set: tres_propiedades.txt
% Maximum length: 4

@relation L-V
@attribute length real
@attribute P1 real
@attribute P2 real
@attribute P3 real
@attribute class real

@data
0.75, 0.530, 0.692, 0.313, 5.389313
```

Figura 5.8: Fichero L-V.aspf que contiene un vector de predicción con extremos L y V.

Tras la realización de dichas tareas, se tienen en la memoria del computador todos los vectores de predicción de las proteínas de test organizados según lo descrito en la Tarea 4. No obstante, estos vectores de test no se almacenan en disco como ocurre con los de training, ya que no son necesarios tras realizar las predicciones.

5.5.1. Tarea 5. Búsqueda del vector de predicción más parecido

Una vez obtenidos todos los vectores de predicción de las proteínas de test, se realiza una búsqueda secuencial completa a partir de cada uno de ellos sobre los vectores de predicción de training. El objetivo es encontrar el vector de predicción de training que guarda mayor similitud con cada vector de predicción de test. Para el proceso de búsqueda sólo se consideran los vectores de training con los mismos extremos que el vector de test. En la figura 5.9 se ilustra el escenario de la búsqueda.

L^{ts} es la longitud de la subsecuencia de test. L^{tr} es la longitud de la subsecuencia de training con mayor similitud a la subsecuencia de test. $\overline{P}_1^{ts} \dots \overline{P}_k^{ts}$ son los valores medios de las propiedades de los aminoácidos interiores de la subsecuencia de test y $\overline{P}_1^{tr} \dots \overline{P}_k^{tr}$ los de la subsecuencia de training más próxima. La distancia a predecir se simboliza con ? y se asignará a la misma la distancia D^{tr} del vector de training más parecido.

El vector de training con mayor similitud al vector de test satisface la

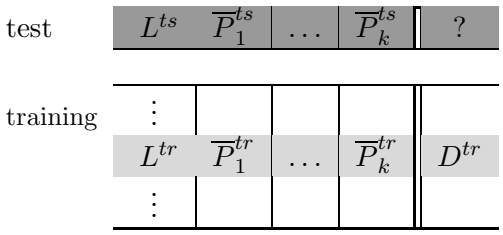


Figura 5.9: Búsqueda del vector de training más parecido a un vector de test.

condición 5.1. Se utiliza una distancia euclídea entre los vectores de test y de training, en la que intervienen las longitudes de las subsecuencias y los valores promedios de las propiedades de sus aminoácidos.

$$min\sqrt{(L^{ts}-L^{tr})^2+(\overline{P}_1^{ts}-\overline{P}_1^{tr})^2+\cdots+(\overline{P}_k^{ts}-\overline{P}_k^{tr})^2} \qquad (5.1)$$

Para ilustrar el proceso de búsqueda y posterior asignación y almacenamiento de predicciones, continuamos utilizando como ejemplo los vectores de training obtenidos en las tareas 1, 2 y 3 explicadas anteriormente.

Supongamos que se añade al conjunto de training la secuencia de proteína "MKPCC", de la cual se han extraído sus subsecuencias (MKP, MKPC, MKPCC, KPC, KPCC y PCC) y se han calculado sus vectores de predicción del modo explicado anteriormente.

Tras la incorporación de esta nueva secuencia de ejemplo y posterior organización y almacenamiento de sus vectores de predicción, se tiene la base de conocimiento que se muestra en la tabla 5.11. Nótese que al añadir la nueva secuencia, la longitud máxima de secuencia es ahora 5, en lugar de 4, lo cual afecta a la normalización del campo longitud (L) de cada vector.

Supongamos que tenemos una secuencia de test de ejemplo ("MKCC"), de la cual se conoce la estructura de proteína. Pese a que la estructura es conocida, se ignora para realizar la predicción y después se utiliza para evaluar la calidad de dicha predicción, comparando la estructura real con la predicha.

En la figura 5.10 se muestra el mapa de distancias con las distancias reales de la secuencia de test, y en la tabla 5.12 se muestran los vectores de predicción de la misma.

Como se ha indicado anteriormente, para cada vector de test se calcula la distancia euclídea con cada vector de training de iguales aminoácidos extremos y se obtiene el vector más "cercano". En la tabla 5.13 se ilustra el resultado del proceso de búsqueda. Se señalan en negrita las subsecuencias de test, de training más próximas y los valores mínimos de distancia euclídea o diferencia (Dif) entre ambas.

		L	$\overline{P}_1$	$\overline{P}_2$	$\overline{P}_3$	D
K-C.aspf	KVC	0,6	0,000	0,273	0,369	5,472211
	KPC	0,6	0,000	0,348	0,000	5,898690
	KPCC	0,8	0,000	0,307	0,473	7,905224
L-C.aspf	LKVC	0,8	0,265	0,482	0,341	7,121544
L-V.aspf	LKV	0,6	0,530	0,692	0,313	5,389313
M-C.aspf	MKPC	0,8	0,265	0,52	0,156	8,418224
	MKPCC	1,0	0,176	0,435	0,42	9,553188
M-P.aspf	MKP	0,6	0,530	0,692	0,313	5,395521
P-C.aspf	PCC	0,6	0,000	0,267	0,947	5,402254

Tabla 5.11: Conjunto de training formado por los vectores de predicción organizados tras incorporar la nueva secuencia MKPCC.

	M	K	C	C
M	0	3,841013	5,414033	7,911034
K		0	3,826711	5,452234
C			0	3,909833
C				0

Figura 5.10: Mapa de distancias con distancias reales de la secuencia MKCC de test.

	L	$\overline{P}_1$	$\overline{P}_2$	$\overline{P}_3$	D
MKC	0,6	0,530	0,692	0,313	?
MKCC	0,8	0,265	0,479	0,630	?
KCC	0,6	0,000	0,267	0,947	?

Tabla 5.12: Vectores de predicción de la secuencia MKCC de test.

5.5.2. Tarea 6. Asignación y almacenamiento de predicciones

Al campo distancia (D) de cada vector de test se le asigna el valor del campo distancia del vector de training más próximo. En la tabla 5.14 se muestran los vectores de test, usados como ejemplo en la tarea anterior, en los que se les ha asignado la distancia del vector de training más cercano. La distancia predicha asignada a cada vector de test representa la distancia entre los aminoácidos extremos de la subsecuencia a la que se refiere dicho vector.

Finalmente, las distancias predichas son almacenadas en la triangular inferior de la matriz de distancias de la secuencia de test, tal como se muestra

Test	L	$\overline{P}_1$	$\overline{P}_2$	$\overline{P}_3$	Train	L	$\overline{P}_1$	$\overline{P}_2$	$\overline{P}_3$	Dif
MKC	0,6	0,530	0,692	0,313	**MKPC**	0,8	0,265	0,52	0,156	**0,405**
MKC	0,6	0,530	0,692	0,313	MKPCC	1,0	0,176	0,435	0,42	0,602
MKCC	0,8	0,265	0,479	0,630	MKPC	0,8	0,265	0,52	0,156	0,475
MKCC	0,8	0,265	0,479	0,630	**MKPCC**	1,0	0,176	0,435	0,42	**0,306**
KCC	0,6	0,000	0,267	0,947	KVC	0,6	0,000	0,273	0,369	0,578
KCC	0,6	0,000	0,267	0,947	KPC	0,6	0,000	0,348	0,000	0,950
KCC	0,6	0,000	0,267	0,947	**KPCC**	0,8	0,000	0,307	0,473	**0,516**

Tabla 5.13: Resultados de la búsqueda en vectores de training.

	L	$\overline{P}_1$	$\overline{P}_2$	$\overline{P}_3$	D
MKC	0,6	0,530	0,692	0,313	**8,418224**
MKCC	0,8	0,265	0,479	0,630	**9,553188**
KCC	0,6	0,000	0,267	0,947	**7,905224**

Tabla 5.14: Vectores de predicción con distancias predichas para la secuencia MKCC de test.

en la figura 5.11.

	M	K	C	C
M	0	3,841013	5,414033	7,911034
K	-	0	3,826711	5,452234
C	8,418224	-	0	3,909833
C	9,553188	7,905224	-	0

Figura 5.11: Mapa de distancias con distancias reales y predichas de la secuencia MKCC de test.

Nótese que, al ser 3 la longitud mínima de subsecuencia, las predicciones entre aminoácidos consecutivos en la cadena no se realizan. Esto se aprecia en la ausencia de valores justo debajo de la diagonal principal en la matriz de distancias. Las predicciones de distancia entre aminoácidos consecutivos no tiene relevancia para la comunidad científica, puesto que siempre presentan aproximadamente las mismas distancias entre ellos (en torno a 3,8 Å).

5.6. Subsistema DMeval

Una vez se han realizado todas las predicciones, éstas residen en las matrices de distancias ubicadas en memoria y, si el usuario lo desea, también en ficheros. A partir de ellas, el subsistema DMeval genera una serie de

medidas para evaluar la calidad de las predicciones.

5.6.1. Tarea 7. Obtención del error de predicción a partir del mapa de distancias

La primera medida que se obtiene para evaluar la calidad de las predicciones es el error relativo medio $\overline{\varepsilon_r}$, el cual se define según la fórmula 5.2.

$$\overline{\varepsilon_r} = \frac{1}{n} \sum_{\substack{i=1..n \\ j=1..n \\ j-i>minsep}} \frac{|d_{ij} - d_{ji}|}{d_{ij}} \tag{5.2}$$

El valor d_{ij} es la distancia real entre el aminoácido i-ésimo y j-ésimo de la secuencia, obtenida de la triangular superior del mapa de distancias de la misma. El valor d_{ji} es la distancia predicha entre los aminoácidos i-ésimo y j-ésimo, obtenida de la triangular inferior.

Para el mapa de distancias de la figura 5.11, el error relativo medio es 0,70206. Este error tiene un valor alto, aunque se debe fundamentalmente a las predicciones no realizadas sobre aminoácidos consecutivos. Sin embargo, si asignamos el valor 3,8 a dichas predicciones, tal como se apuntó en la tarea 6, el error relativo medio sería 0,20968.

5.6.2. Tarea 8. Obtención de sensibilidad y otras medidas para distintos umbrales

El subsistema DMeval genera las medidas de exactitud, sensibilidad, especificidad, precisión y MCC, definidas en el epígrafe 3.4.2. Como se explicó anteriormente, estas medidas se utilizan para evaluar la calidad de las predicciones sobre una clase discreta. Dado que la clase, como hemos visto, es un valor real (una distancia), para utilizar estas medidas es necesario discretizar previamente la clase.

Tal como se pudo comprobar en el capítulo 4, lo habitual en la literatura es evaluar predicciones de contactos (binarios) entre aminoácidos y, por ello, se ha procedido aquí a discretizar la clase en dos intervalos.

Se ha utilizado una distancia umbral para fijar un punto de corte en los valores de distancia continuos. De este modo, los valores de distancia menores a este umbral se consideran con valor 1, y los valores mayores al umbral con valor 0. El significado de la nueva clase binarizada es 1 (contacto entre aminoácidos) y 0 (no contacto). Como se comprobó en el capítulo 4, en la literatura se utilizan diferentes umbrales para la definición de contactos, siendo el umbral de 8 Å el más utilizado.

En la tabla 5.15 se muestran los valores de estas medidas obtenidas a partir del mapa de distancias de la figura 5.11 utilizando dos umbrales distintos: 5 y 8 Å.

Umbral	TP	TN	FP	FN	Exact.	Sens.	Espec.	Prec.
5 Å	0	3	0	0	1,00	0,00	1,00	0,00
8 Å	1	0	0	2	0,33	0,33	0,00	1,00

Tabla 5.15: Exactitud, sensibilidad, especificidad y precisión para dos umbrales diferentes.

Como se ha comprobado en la tabla 5.15, para los datos del ejemplo se consigue mejor sensibilidad y precisión (las dos medidas más relevantes en la literatura) con umbral 8 Å que con 5 Å. El efecto en los resultados del umbral de contacto es analizado en detalle en el capítulo 6.

5.7. Resumen

En este capítulo se ha explicado el procedimiento llevado a cabo por ASPpred, el sistema propuesto para la predicción de estructuras de proteínas. Este sistema toma propiedades físico-químicas de aminoácidos como entrada, utiliza un esquema de vecinos más cercanos y produce mapas de distancias como salida. La evaluación de los resultados está basada en la binarización de las distancias predichas utilizando un umbral de contacto.

Parte IV

Resultados

Capítulo 6

Experimentación principal

6.1. Introducción

En el presente capítulo se exponen los resultados principales obtenidos con el sistema predictor ASPpred propuesto en el capítulo 5. Se han utilizado 5 conjuntos de proteínas, los cuales se describen en la sección 6.2.

Se ha realizado una selección de atributos sobre un superconjunto de propiedades de aminoácidos, la cual se explica en la sección 6.3. La configuración de la experimentación realizada se indica en la sección 6.4.

Previamente a la ejecución del sistema predictor se han realizado dos estudios para visualizar y comprender la distribución de los datos de entrada. Estos estudios se abordan en la sección 6.5. Los resultados de las predicciones sobre los conjuntos de proteínas se exponen en el epígrafe 6.6.

Tras la generación de las predicciones de mapas de distancias, se realizan múltiples análisis para descubrir determinados aspectos del comportamiento predictivo del sistema ASPpred. Estos análisis se abordan en la sección 6.7.

En último lugar, se realiza una comparativa del método propuesto con otro de la literatura, la cual puede encontrarse en la sección 6.8.

6.2. Conjuntos de proteínas

El objetivo que se ha seguido en la elección de los conjuntos de proteínas es el de utilizar proteínas no homólogas, es decir, proteínas con secuencias lo más diferentes posibles. De esta forma, se permite comprobar si el método de predicción es lo suficientemente general, y no que funcione únicamente para familias concretas de proteínas.

Se denomina porcentaje de identidad de un conjunto de proteínas, al porcentaje máximo de regiones comunes en las secuencias de todos los pares de proteínas de dicho conjunto. Cuanto menor es el porcentaje de identidad de un conjunto, más heterogéneas son sus secuencias. Para que las

secuencias no se consideren homólogas, su porcentaje de identidad debe ser como máximo del 30 %.

Además del porcentaje de identidad, existen otros dos parámetros para la obtención de proteínas a partir de las bases de datos: resolución y R-factor. Ambos parámetros miden la calidad de la obtención experimental de las proteínas. Cuanto más cercanos a cero sean los valores de ambos parámetros, más exactos son los modelos estructurales obtenidos en el laboratorio y, a la vez, menos modelos existen. Al utilizar modelos estructurales de mayor calidad, se introduce menor error en los datos de entrada y, por tanto, el predictor aprende modelos que se encuentran más ajustados a las estructuras reales de proteínas.

Se han considerado 5 conjuntos de proteínas para probar nuestra propuesta ASPpred. Las características de estos conjuntos se describen a continuación.

El primer conjunto de proteínas (CP1) contiene 20 proteínas, elegidas aleatoriamente de un superconjunto de 12830 proteínas. Este superconjunto fue descargado de la Web de PDB en 2010 utilizando una búsqueda avanzada en la que se obtuvieron todas las proteínas publicadas en PDB con porcentaje de identidad del 30 % o inferior. Con este reducido conjunto de proteínas se ha perseguido comprobar el comportamiento predictivo de ASPpred con escasos datos de entrenamiento.

Para el segundo conjunto de proteínas (CP2) se ha utilizado la aplicación CULLPDB [Wang and Dunbrack, 2003] de Dunbrack que permite obtener selecciones de proteínas de alta calidad. Se han seleccionado proteínas de más de 70 aminoácidos (para evitar proteínas de longitud corta, más fáciles de predecir), con resolución entre 0-1.0, R-factor entre 0-0.2 y con un máximo del 10 % de identidad. Además se han suprimido proteínas obtenidas mediante rayos X y proteínas de las que sólo se han recabado la posición de carbonos alfa (CA). El conjunto CP2 finalmente contiene 118 proteínas.

En el tercer conjunto (CP3) se ha utilizado la aplicación PDBselect [Griep and Hobohm, 2010] que permite obtener, al igual que CULLPDB, selecciones de proteínas de alta calidad. Para este experimento se han seleccionado proteínas de longitud mayor que 40 aminoácidos, resolución entre 0-1.4, R-factor entre 0-0.12 e identidad máxima del 25 %. En este conjunto se han relajado algo las condiciones de selección para producir mayor número de proteínas. La selección generada posee 170 proteínas.

En el cuarto conjunto de proteínas (CP4) se ha utilizado también la aplicación CULLPDB. Se han seleccionado proteínas de más de 70 aminoácidos, con resolución entre 0-1.1, R-factor entre 0-0.2 y con tan sólo el 5 % de identidad o menos. En este experimento se ha perseguido disponer de un conjunto mayor de proteínas que fueran incluso más heterogéneas que en los experimentos anteriores. El conjunto final tiene 221 proteínas.

El quinto conjunto de proteínas (CP5) contiene todas las proteínas disponibles en la base de datos PDBselect a fecha de febrero de 2011 con

un identidad máxima en la secuencia del 25 %. Este conjunto contiene 5130 proteínas.

6.3. Selección de propiedades

El conjunto de propiedades de aminoácidos que se ha utilizado en la experimentación que se describe en este capítulo ha sido obtenido mediante un proceso de selección de atributos sobre el repositorio completo de 544 propiedades de AAindex.

Tras una exhaustiva exploración de los algoritmos de evaluación de atributos y esquemas de búsqueda disponibles, se han encontrado los mejores resultados utilizando Relief [Kononenko, 1994] como algoritmo de evaluación y Ranker como esquema de búsqueda. Se han empleado 10 vecinos más cercanos en la configuración del algoritmo Reief en la herramienta Weka [Hall et al., 2009].

El conjunto de datos de entrada utilizado para el proceso de selección de atributos está formado por las proteínas del conjunto CP4. El conjunto inicial de atributos lo forman las 544 propiedades de AAindex y la clase es la distancia discretizada con umbral de 8 angstroms entre los pares de aminoácidos.

El esquema de búsqueda Ranker produce un ranking de atributos. Se han realizado predicciones con todos los subconjuntos que comienzan en el primer atributo del ranking (subconjuntos $\{\#1\}, \{\#1, \#2\}, \cdots, \{\#1, \cdots, \#N\}$). Se han encontrado los mejores resultados con los 30 primeros atributos (subconjunto $\{\#1, \cdots, \#30\}$), los cuales se han utilizado de aquí en adelante y se muestran en la tabla 6.1. Tanto los conjuntos de proteínas como la selección de propiedades empleada se encuentran disponibles en la dirección http://www.upo.es/eps/asencio/asppred.

6.4. Configuración de la experimentación

Se han realizado 5 experimentos para probar el funcionamiento del sistema ASPpred, uno para cada conjunto de proteínas. Se ha establecido una configuración inicial idéntica para todos los experimentos, salvo para el quinto experimento. El único elemento que varía de un experimento a otro es únicamente el conjunto de proteínas que utiliza.

En los experimentos 1, 2, 3 y 4 se ha empleado una validación cruzada con 10 bolsas. En el experimento 5 se ha usado *hold-out*, con entrenamiento el conjunto CP2 y test el conjunto CP5. A continuación se describe la configuración general empleada en todos los experimentos del capítulo.

En la configuración del sistema ASPpred se han establecido 400 hilos de ejecución. De este modo, se paraleliza la generación y organización de los vectores de predicción (según se explicó en los puntos 5.4.3 y 5.4.4) por

Nombre	Descripción
AURR980120	Normalized positional residue frequency at helix termini C4'
BUNA790101	alpha-NH chemical shifts
BUNA790103	Spin-spin coupling constants 3JHalpha-NH
CHAM820102	Free energy of solution in water, kcal/mole
DIGM050101	Hydrostatic pressure asymmetry index, PAI
FAUJ880111	Positive charge
FAUJ880112	Negative charge
GARJ730101	Partition coefficient
JOND750102	pK (-COOH)
KARP850103	Flexibility parameter for two rigid neighbors
KHAG800101	The Kerr-constant increments
MAXF760103	Normalized frequency of zeta R
MITS020101	Amphiphilicity index
MONM990201	Averaged turn propensities in a transmembrane helix
NADH010107	Hydropathy scale based on self-information values in the two-state model (50 % accessibility)
PRAM820101	Intercept in regression analysis
QIAN880139	Weights for coil at the window position of 6
RICJ880101	Relative preference value at N"
RICJ880104	Relative preference value at N1
RICJ880114	Relative preference value at C1
RICJ880117	Relative preference value at C"
SUEM840102	Zimm-Bragg parameter sigma x 1.0E4
TANS770102	Normalized frequency of isolated helix
TANS770108	Normalized frequency of zeta R
VASM830101	Relative population of conformational state A
VELV850101	Electron-ion interaction potential
WERD780102	Free energy change of epsilon(i) to epsilon(ex)
WERD780103	Free energy change of alpha(Ri) to alpha(Rh)
WILM950104	Hydrophobicity coefficient in RP-HPLC, C18 with 0.1 %TFA/2-PrOH/MeCN/H2O
YUTK870103	Activation Gibbs energy of unfolding, pH7.0

Tabla 6.1: La selección de 30 propiedades utilizada.

cada par de aminoácidos ($20 \times 20 = 400$ tipos de pares de aminoácidos posibles). De este modo se consigue el mayor rendimiento sobre la máquina de ejecución de experimentos. Las características de esta máquina y los tiempos de ejecución conseguidos se abordan en el epígrafe 6.7.5.

Se ha utilizado el vector de predicción tipo A, descrito en el apartado

5.4.3. Se ha utilizado una separación mínima en la secuencia de 1 aminoácido. Se ha utilizado el carbono beta (CB) como átomo de referencia para el cálculo de distancias entre aminoácidos. No obstante, para el aminoácido Glicina (GLY), el cual no tiene carbono beta, se ha utilizado el carbono alfa (CA), tal como se emplea en la literatura. No se ha utilizado ranking TopLx y se han empleado diferentes umbrales de contacto que se indicarán más adelante.

6.5. Estudio previo de los datos de entrada

6.5.1. Estudio de puntos $(\overline{P}_i, D)$

Las figuras 6.1 y 6.2 muestran la distribución de distancias entre aminoácidos en función del promedio de una propiedad en los aminoácidos que hay entre ellos, es decir, la distribución de los puntos $(\overline{P}_i, D)$ de los vectores de predicción tipo A.

Para este estudio se han utilizado las proteínas del conjunto CP4. Se ha incluido la distribución de los puntos $(\overline{P}_i, D)$ de dos propiedades (WILM9501040 en la figura 6.1 y GARJ730101 en la figura 6.2), aunque las distribuciones del resto de propiedades son similares. El eje de abcisas representa el valor promedio de la propiedad en los aminoácidos interiores entre dos dados y el eje de ordenadas la distancia entre estos dos.

Como se puede apreciar en las figuras 6.1 y 6.2, las distancias entre aminoácidos parecen seguir una distribución normal: con media 0.402 y desviación 0.31, en el caso de la propiedad WILM9501040; con media 0.047 y desviación 0.059, en el caso de la propiedad GARJ730101.

6.5.2. Estudio de puntos $(P_i(s_b), P_i(s_e), D)$

Se ha realizado otro estudio que proporciona más detalle de la distribución de distancias según las propiedades de los aminoácidos. Este estudio muestra las distancias medias entre pares de aminoácidos según sus propiedades. A diferencia del estudio anterior, los propiedades representadas son las de los dos aminoácidos cuya distancia es analizada. Es decir, se representan en este estudio los puntos $(P_i(s_b), P_i(s_e), D)$, utilizando la misma nomenclatura introducida en 5.4.3.

Debido a que son tres las variables que se analizan, se ha optado por una representación en superficies tridimensionales, con el objetivo de encontrar posibles patrones visualmente. Dados dos valores concretos $(P_i(s_b), P_i(s_e))$, se ha utilizado una media armónica para resumir las distancias de todos los pares de aminoácidos con dichos valores de propiedades. Se ha empleado una media armónica para mitigar el impacto de ciertos valores inusualmente altos de distancia. Se ha generado una gráfica 3D para cada propiedad del conjunto seleccionado previamente (30 gráficas).

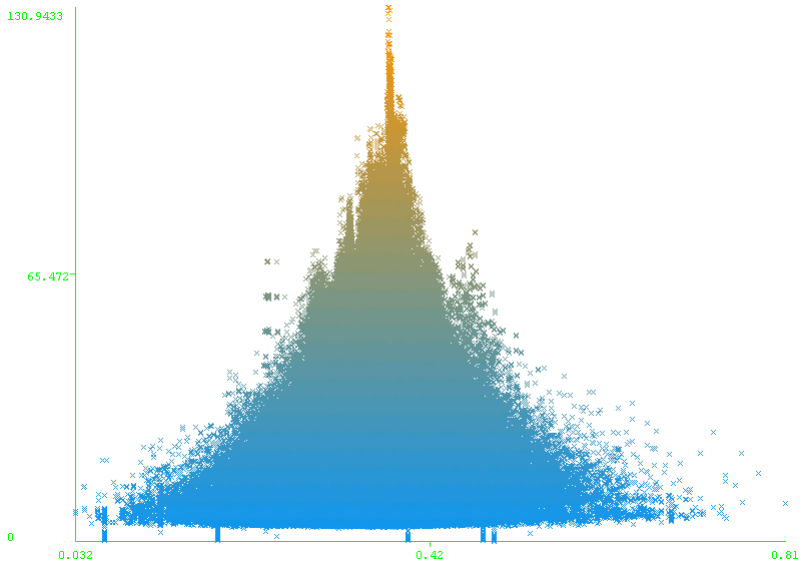


Figura 6.1: Distribución de la propiedad WILM950104. El eje X representa el valor promedio de la propiedad en los aminoácidos interiores entre dos dados y el eje Y la distancia entre los dos.

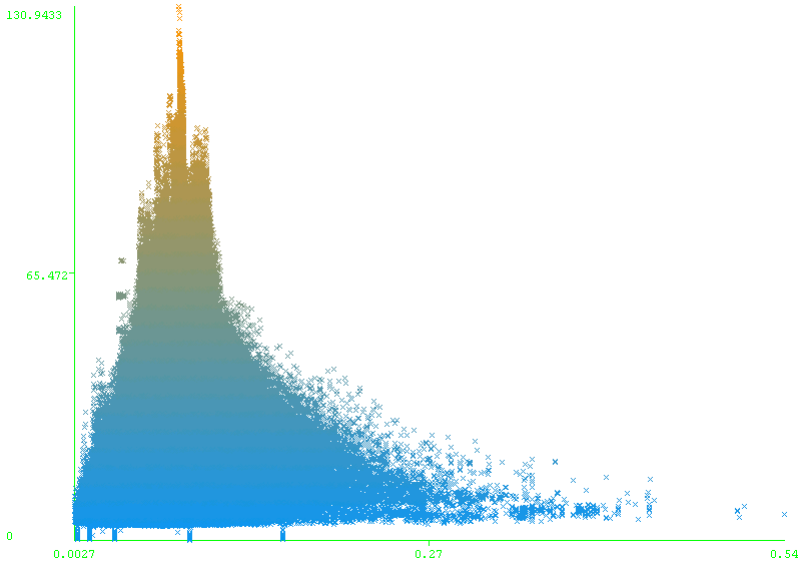
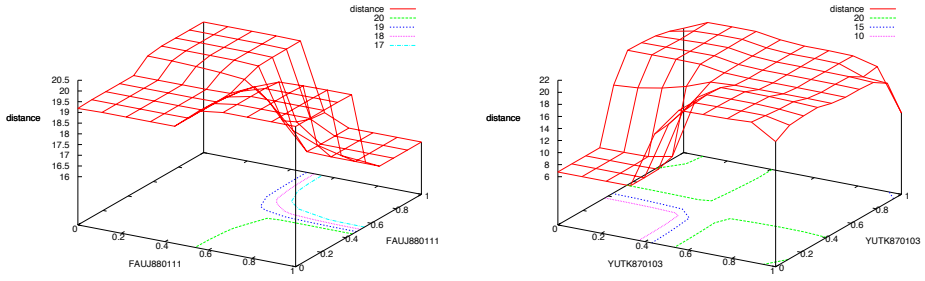


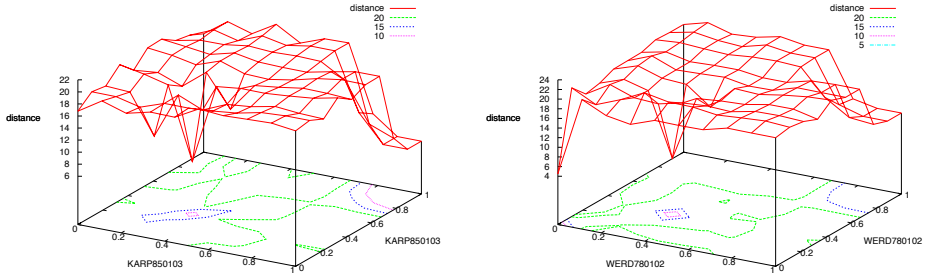
Figura 6.2: Distribución de la propiedad GARJ730101. El eje X representa el valor promedio de la propiedad en los aminoácidos interiores entre dos dados y el eje Y la distancia entre los dos.

Tras analizar las superficies 3D generadas, se han encontrado tres patrones que se repiten. El primer patrón se ha denominado “superficies



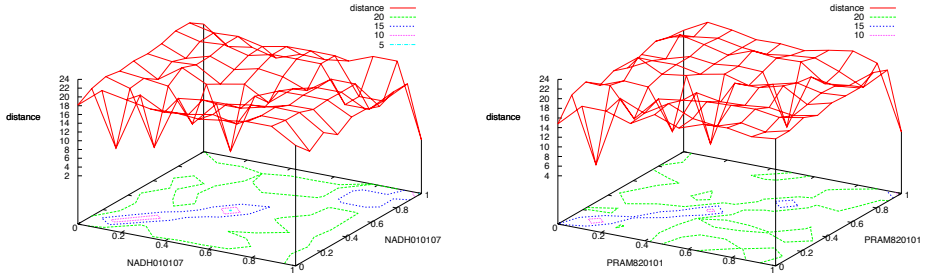
(a) FAUJ880111 (b) YUTK870103

Figura 6.3: Tipo de patrón “superficies escalonadas”.



(a) KARP850103 (b) WERD780102

Figura 6.4: Tipo de patrón “superficies cornisa”.



(a) NADH010107 (b) PRAM820101

Figura 6.5: Tipo de patrón “superficies valle”.

escalonadas”. Las superficies de este tipo presentan varios planos paralelos al formado por los ejes $(P_i(s_b), P_i(s_e))$ de diferente altura (D). Se han incluido dos superficies de este tipo en la figura 6.3 (para las propiedades (a) FAUJ880111 y (b) YUTK870103). Aparte de éstas dos, las propiedades

FAUJ880112 y MONM990201 presentan el mismo comportamiento.

El segundo patrón que se ha encontrado se ha denominado "superficies cornisa". Las superficies de este tipo presentan valores bajos de distancia para valores bajos o altos de las propiedades de los aminoácidos. Se han incluido dos superficies de este tipo en la figura 6.4 (para las propiedades (a) KARP850103 y (b) WERD780102). Existen, además, otras propiedades que presentan el mismo comportamiento (QIAN880139, RICJ880101, BUNA790101, BUNA790103, KHAG800101, MAXF760103, RICJ880117 y GARJ730101).

El tercer patrón se ha denominado "superficies valle". Las superficies de este tipo presentan valores bajos de distancia cuando la propiedad de los dos aminoácidos es similar. Se han incluido dos superficies de este tipo en la figura 6.5 (para las propiedades (a) NADH010107 y (b) PRAM820101). Existen, además, otras propiedades que presentan el mismo comportamiento (AURR980120, CHAM820102, JOND750102, TANS770108, RICJ880114, VASM830101 y VELV850101).

A partir del estudio realizado con las superficies de puntos $(P_i(s_b), P_i(s_e), D)$, podrían extraerse reglas para la predicción de distancias entre aminoácidos a partir de sus propiedades. Dada la naturaleza de la aproximación de predicción que se propone, basada en la similitud de las propiedades de aminoácidos, el comportamiento ofrecido por las propiedades del segundo y, especialmente, del tercer patrón es deseable, pues existe cierta correspondencia con la distancia a predecir.

6.6. Evaluación de las predicciones realizadas

6.6.1. Evaluación de cada experimento

En las tablas 6.2 y 6.3 se muestran los resultados de los 5 experimentos anteriormente explicados. Se consignan la media y desviación típica de la sensibilidad, precisión, exactitud y especificidad. En la tabla 6.2 se ha utilizado un umbral de contacto de 4 Å y en la tabla 6.3 de 8 Å.

Experimento	Sensibilidad	Precisión	Exactitud	Especificidad
1	0.10±0.05	0.08±0.10	0.99±0.00	0.99±0.00
2	0.31±0.10	0.39±0.11	0.99±0.00	0.99±0.00
3	0.48±0.04	0.43±0.05	0.99±0.01	0.99±0.01
4	0.40±0.05	0.41±0.05	0.99±0.01	0.99±0.01
5	0.14±0.08	0.14±0.08	0.99±0.05	0.99±0.05

Tabla 6.2: Eficacia de ASPpred usando 4 Å como umbral de contacto ($\mu \pm \sigma$).

Experimento	Sensibilidad	Precisión	Exactitud	Especificidad
1	0.39±0.06	0.41±0.08	0.97±0.03	0.98±0.01
2	0.39±0.07	0.40±0.07	0.95±0.01	0.97±0.02
3	0.38±0.02	0.38±0.02	0.95±0.02	0.97±0.01
4	0.40±0.03	0.41±0.03	0.95±0.01	0.97±0.01
5	0.51±0.11	0.51±0.11	0.92±0.06	0.95±0.07

Tabla 6.3: Eficacia de ASPpred usando 8 Å como umbral de contacto ($\mu \pm \sigma$).

Para 8 Å de umbral, la sensibilidad y la precisión tienen valores parecidos en los experimentos 1 a 4. Por tanto, según estas pruebas, el predictor ASPpred no necesita gran volumen de entrenamiento para ofrecer los mismos resultados.

En el experimento 5 se ha conseguido mejor sensibilidad y precisión que en los otros experimentos; no obstante, la desviación típica es mayor. Esto último puede ser debido al gran número de tipos diferentes de proteínas (múltiples clases estructurales y número de dominios, por ejemplo) presentes en todo PDBselect.

Los valores de error relativo medio ($\overline{\varepsilon_r}$) obtenidos son 0,7114, 0,5174, 0,5212, 0,5041 y 0,4305 en los experimentos 1, 2, 3, 4 y 5, respectivamente.

6.6.2. Evaluación según clase estructural

Para proporcionar más detalle acerca del comportamiento predictivo de ASPpred, y teniendo en cuenta el alto valor de la desviación típica de la sensibilidad y precisión de algunos experimentos, se han incluido en el anexo A de esta Tesis cinco tablas (A.1, A.2, A.3, A.4 y A.5) con los resultados obtenidos según la clase estructural CATH [Orengo et al., 2002] de las proteínas que se predicen. Se consignan, para cada experimento, las diferentes clases estructurales presentes en el conjunto de proteínas, el número de secuencias de cada clase y la media y desviación típica de la sensibilidad y precisión obtenida por ASPpred.

Además, se ha preparado información más detallada acerca de las proteínas de cada experimento, indicando su clase estructural CATH, su ID de la base de datos CATH, su descripción, número de dominios PFAM [Finn et al., 2008], resolución, longitud de la secuencia y, finalmente la sensibilidad y precisión obtenida por ASPpred para cada proteína. Toda esta información se encuentra disponible en la dirección http://www.upo.es/eps/asencio/asppred.

6.6.3. Mapas de distancias y contactos

La figura 6.6 muestra el mapa de distancias predicho para la proteína 3CCD (de 85 aminoácidos) del conjunto CP2. Se ha usado una escala de color para representar los valores de distancias, comenzando desde la mínima distancia (color rojo) hasta la máxima (color azul). Como se puede apreciar en la figura 6.6, la triangular inferior de la matriz de distancias (predicción) es bastante similar a la triangular superior (observación).

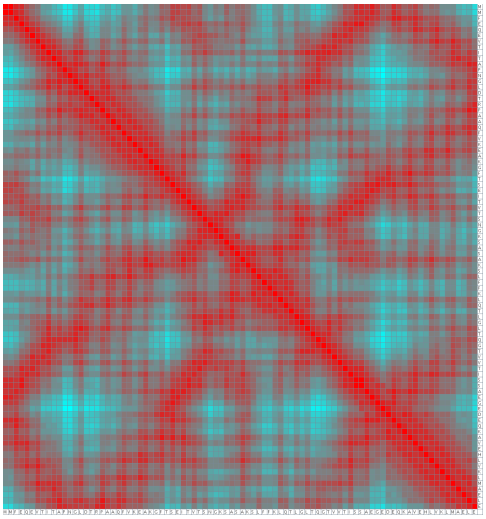


Figura 6.6: Mapa de distancias predicho por ASPpred para la proteína 3CCD.

La figura 6.7 muestra el mapa de contactos de la misma proteína 3CCD, obtenido a partir del mapa de distancias mostrado en la figura 6.6 aplicando un umbral de contacto de 8 angstroms. Al igual que el mapa de distancias, existe una gran similitud entre la parte real y la predicha en el mapa de contactos.

6.7. Análisis de las predicciones realizadas

6.7.1. Efecto del umbral de contacto

Se ha realizado un análisis visual de la sensibilidad, precisión, exactitud y especificidad para diferentes umbrales de contacto, utilizando los conjuntos de proteínas CP1, CP2, CP3 y CP4. Los resultados de este análisis se muestran en la figura 6.8. En cada gráfico de la figura se muestra el umbral de contacto en el eje de abcisas (medido en angstroms) y la sensibilidad, precisión, exactitud y especificidad en el eje de ordenadas.

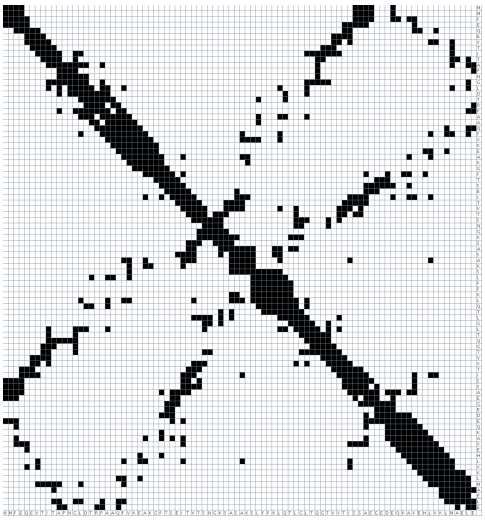


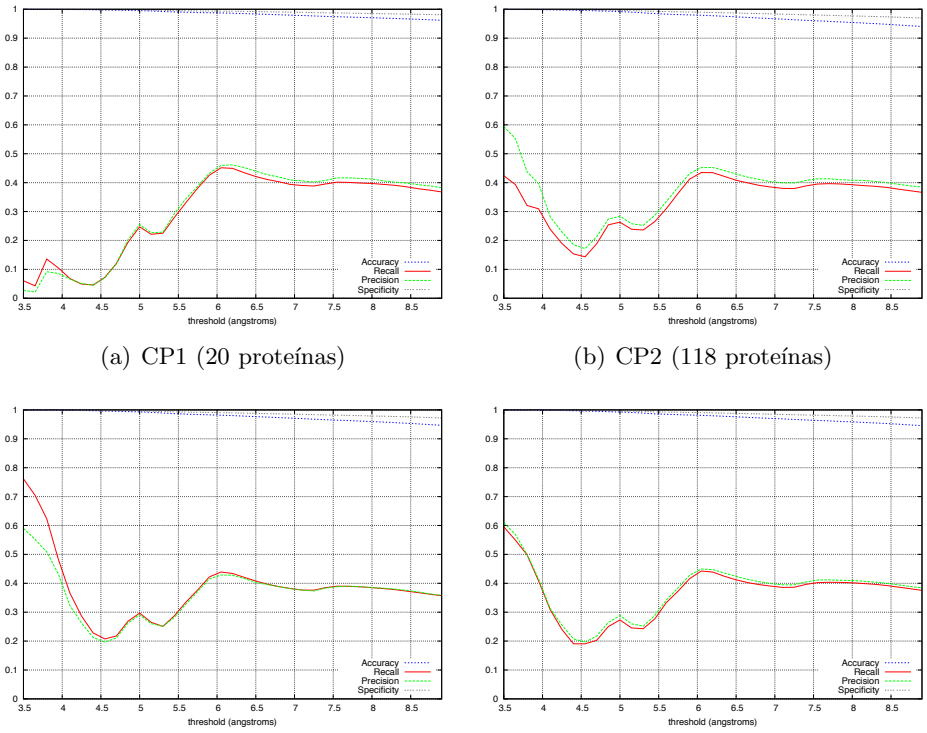
Figura 6.7: Mapa de contactos predicho por ASPpred para la proteína 3CCD usando 8Å como umbral de contacto.

A la vista de los resultados obtenidos en la figura 6.8 extraemos las siguientes conclusiones. En primer lugar, se aprecia un comportamiento parecido en las gráficas de los experimentos 2, 3 y 4 que las diferencia de la del experimento 1. En concreto, las diferencias se producen para valores de umbral bajos, de 3,5 a 4,8 Å aproximadamente. Esto puede deberse al volumen de proteínas del conjunto de entrenamiento: a mayor número de proteínas de entrenamiento, mayor es la cobertura del espacio de búsqueda (espacio de las propiedades de los aminoácidos) y por tanto mejores son, en este caso, los valores de sensibilidad y precisión.

Otra lectura posible, dada la similitud de las tendencias de sensibilidad y precisión en los experimentos 2, 3 y 4, es que la respuesta del método ante la diversidad de proteínas parece ser la misma con independencia del tipo de proteína a predecir. De hecho, los conjuntos de proteínas de dichos experimentos tienen una identidad muy baja (hasta del 5 % o inferior en el CP4). Este resultado es deseable, ya que se busca la generalidad del método, como se comentó anteriormente.

Por otra parte, es destacable que los valores de exactitud y especificidad son siempre muy elevados (próximos o iguales a 1). Estos valores altos de exactitud y especificidad tienen una explicación clara si tenemos en cuenta que los valores de TN son muy altos comparados con los de TP, FP y FN, debido a que la clase discretizada (contacto binario entre aminoácidos) está desbalanceada en este problema (en una proporción de 1/60 de la clase positiva con respecto a la negativa).

Por ello, tanto la exactitud como la especificidad, que son directamente



(a) CP1 (20 proteínas)

(b) CP2 (118 proteínas)

(c) CP3 (170 proteínas)

(d) CP4 (221 proteínas)

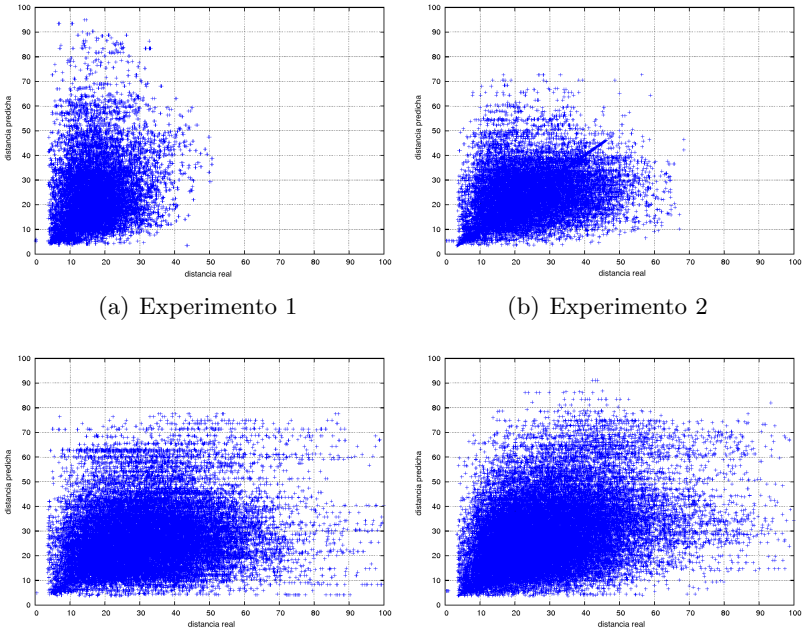
Figura 6.8: Sensibilidad, precisión, exactitud y especificidad según umbral de contacto, utilizando los conjuntos de proteínas CP1, CP2, CP3 y CP4.

proporcionales a TN, tienen unos valores altos. Por contra, las medidas de sensibilidad y precisión son directamente proporcionales a la tasa de TP, la cual se mantiene en valores más pequeños y con mayor variabilidad en función del umbral.

6.7.2. Distribución de distancias reales y predichas

Con el objetivo de analizar exactamente las predicciones de distancias producidas por ASPpred, se han incluido gráficos de puntos $\langle distanciaReal, distanciaPredicha \rangle$ para los experimentos 1, 2, 3 y 4, los cuales se muestran en la figura 6.9.

A partir de la figura 6.9 se puede apreciar el efecto del volumen de datos de entrenamiento (cada experimento, del primero al cuarto, contiene más proteínas que el anterior) y la tendencia de las predicciones. En el experimento 1 las distancias predichas son, en gran medida, mayores que las distancias reales. Sucesivamente, al incrementar el volumen de entrenamiento, se aprecia cierta tendencia a que las distancias predichas sean menores que las reales, en especial para distancias superiores a 40 Å.



(a) Experimento 1 (b) Experimento 2

(c) Experimento 3 (d) Experimento 4

Figura 6.9: Distancias reales y predichas en los experimentos 1 a 4.

6.7.3. Impacto de la hidrofobicidad en el error cometido

Una característica bien conocida de las proteínas es que en su núcleo interno se encuentran aminoácidos que son hidrofóbicos, los cuales, por una cuestión de espacio, tal como se explicó en el epígrafe 4.2.2, se encuentran cercanos unos de otros (las distancias entre ellos son parecidas).

Además, los aminoácidos de la superficie de la proteína son hidrofílicos (no hidrofóbicos) y las distancias entre ellos suelen ser mayores. Por tanto, en ambos casos puede ser relativamente sencillo predecir las distancias en función de la hidrofobicidad y el predictor ASPpred podría estar realizando sus mejores predicciones sólo en estos casos, mostrándose quizás incapaz de resolver distancias en otros casos.

Para esclarecer esta duda y comprobar si las mejores predicciones de ASPpred corresponden sólo a aminoácidos del núcleo o de la superficie proteínica, se ha incluido la tabla 6.4.

La idea es comprobar si el error es menor para aminoácidos con la misma hidrofobicidad; esto es, aminoácidos cuya diferencia de hidrofobicidad es cercana a cero. En la tabla 6.4 se muestra el error absoluto (media y desviación estándar) entre las distancias reales y predichas obtenido para cada rango de la diferencia de hidrofobicidad entre los pares de aminoácidos. La propiedad de hidrofobicidad, en la selección de 30 propiedades utilizadas, es la WILM950104. Se han utilizado las proteínas del conjunto CP4 para

Hidrofobicidad	Error absoluto ($\mu \pm \sigma$)
[0, 0,2)	13.327±9.472
[0,2, 0,4)	13.029±9.240
[0,4, 0,6)	12.873±9.295
[0,6, 0,8)	12.637±9.222
[0,8, 1]	12.379±9.131

Tabla 6.4: Análisis del error en función de la diferencia de hidrofobicidad (en el experimento 4).

este análisis.

Como se puede apreciar en la tabla 6.4, ASPpred obtiene aproximadamente el mismo error (en media y desviación estándar) con independencia del rango de hidrofobicidad. Además, el error medio y su desviación decrecen un poco cuando la diferencia de hidrofobicidad es más alta. Por tanto, se ha probado que ASPpred no produce sus mejores predicciones únicamente en aminoácidos del núcleo y de la superficie de la proteína.

6.7.4. Estudio del número de vecinos más cercanos

Se ha realizado un estudio de la eficacia de ASPpred en función del número de vecinos más cercanos (parámetro K del algoritmo). En el caso concreto de ASPpred, el número de vecinos más cercanos es el número de vectores de predicción de entrenamiento más similares utilizados. Cuando ASPpred utiliza varios vectores de predicción ($K > 1$), calcula la distancia predicha a partir de la media aritmética de las distancias (D) de dichos vectores.

En este estudio se ha probado con $K = \{1, 3, 5, 7, 9, 11, 13, 15\}$. En la tabla 6.5 se muestran los resultados de este estudio para las proteínas de los conjuntos CP1, CP2, CP3 y CP4 usando 8 Å como umbral de contacto.

Como se puede observar en la tabla 6.5, la precisión mejora considerablemente cuando el parámetro K se incrementa, hasta 0.88 en los conjuntos CP2, CP3 y CP4 y $K \geq 13$. Sin embargo, la sensibilidad decrece cuando K se incrementa, hasta 0.37 en los conjuntos CP2, CP3 y CP4 y $K \geq 13$. No obstante, el incremento de la precisión es notablemente más alto que el decremento de la sensibilidad. Por tanto, según este estudio, el mejor y menor valor de K es 13, ya que, aunque no se han expuesto los resultados con $K > 15$, la sensibilidad y precisión se estabiliza a partir de $K = 13$.

Para validar estadísticamente la significancia de las diferencias de precisión y sensibilidad según el valor de K, los resultados de la experimentación realizada en este estudio han sido sometidos a un test de

CP	K	Sensibilidad	Precisión	Exactitud	Especificidad
1	1	0.39±0.06	0.41±0.08	0.97±0.03	0.98±0.01
	3	0.40±0.05	0.63±0.07	0.97±0.02	0.98±0.00
	5	0.39±0.05	0.73±0.05	0.98±0.02	0.98±0.00
	7	0.38±0.05	0.78±0.05	0.98±0.01	0.98±0.00
	9	0.37±0.04	0.80±0.04	0.98±0.00	0.99±0.00
	11	0.36±0.04	0.83±0.04	0.99±0.00	0.99±0.00
	13	0.35±0.04	0.84±0.04	0.99±0.00	0.99±0.00
	15	0.35±0.04	0.86±0.04	0.97±0.00	0.98±0.00
2	1	0.39±0.07	0.40±0.07	0.95±0.01	0.97±0.02
	3	0.40±0.04	0.73±0.02	0.98±0.01	0.98±0.00
	5	0.39±0.03	0.81±0.01	0.98±0.00	0.99±0.00
	7	0.38±0.03	0.85±0.01	0.99±0.00	0.99±0.00
	9	0.38±0.03	0.86±0.01	0.99±0.00	0.99±0.00
	11	0.37±0.03	0.87±0.01	0.99±0.00	0.99±0.00
	13	0.37±0.03	0.88±0.01	0.99±0.00	0.99±0.00
	15	0.37±0.03	0.88±0.01	0.99±0.00	0.99±0.00
3	1	0.38±0.02	0.38±0.02	0.95±0.02	0.97±0.01
	3	0.40±0.02	0.72±0.01	0.97±0.01	0.98±0.00
	5	0.39±0.01	0.81±0.01	0.98±0.00	0.99±0.00
	7	0.38±0.01	0.84±0.01	0.99±0.00	0.99±0.00
	9	0.38±0.01	0.86±0.01	0.99±0.00	0.99±0.00
	11	0.37±0.01	0.87±0.01	0.99±0.00	0.99±0.00
	13	0.37±0.01	0.88±0.01	0.99±0.00	0.99±0.00
	15	0.37±0.01	0.88±0.01	0.99±0.00	0.99±0.00
4	1	0.40±0.03	0.41±0.03	0.95±0.01	0.97±0.01
	3	0.40±0.04	0.72±0.01	0.96±0.01	0.97±0.00
	5	0.39±0.04	0.81±0.01	0.97±0.00	0.98±0.00
	7	0.39±0.04	0.84±0.00	0.99±0.00	0.99±0.00
	9	0.38±0.04	0.86±0.00	0.99±0.00	0.99±0.00
	11	0.38±0.04	0.87±0.00	0.99±0.00	0.99±0.00
	13	0.37±0.04	0.88±0.00	0.99±0.00	0.99±0.00
	15	0.37±0.04	0.88±0.00	0.99±0.00	0.99±0.00

Tabla 6.5: Estudio de la eficacia de ASPpred según el número (K) de vecinos más cercanos para 8 Å de umbral de contacto ($\mu \pm \sigma$).

significancia estadística. Con el objetivo de utilizar suficientes ejemplos en el test estadístico, se han incluido los valores de sensibilidad y precisión obtenidos para cada proteína de cada conjunto: 20, 118, 170 y 221 valores para cada valor de K (de los conjuntos CP1, CP2, CP3 y CP4,

respectivamente).

Mediante un test previo de D'Agostino-Pearson [D'Agostino et al., 1990] se ha comprobado que los datos utilizados para el test estadístico no satisfacen el criterio de normalidad. Por esta razón, se ha seleccionado un test no paramétrico para validar estadísticamente las diferencias de precisión y sensibilidad. En concreto, el proceso completo que se ha seguido se encuentra descrito en [Garcia and Herrera, 2008], en el cual se hace uso del test de Friedman. De este modo, se han realizado 8 tests de Friedman, dos tests para cada conjunto de proteínas: uno para los valores de sensibilidad y otro para los de precisión.

Tras la ejecución de los tests estadísticos, todos los *p-values* de los tests de precisión han sido menores o iguales a $1{,}21 * 10^{-7}$ y, por tanto, la hipótesis nula queda rechazada. Por contra, los tests estadísticos de la sensibilidad han arrojado valores mayores o iguales a 0,059, por tanto, las diferencias de sensibilidad no son estadísticamente significativas. En resumen, concluimos que las mejoras de precisión, conseguidas al aumentar el valor de K, son significativas, mientras que el descenso de la sensibilidad no lo es.

6.7.5. Estudio del tiempo de ejecución

En la tabla 6.6 se muestran los tiempos de ejecución (en minutos) arrojados por ASPpred en la predicción de los conjuntos CP1, CP2, CP3 y CP4 para cada valor de K. Se indica entre paréntesis el número de proteínas de cada conjunto. El tiempo de preprocesado se produce una sola vez, es decir, no se consume dicho tiempo para cada valor de K, tan sólo en la primera ejecución. La máquina utilizada en todas las experimentaciones de esta Tesis es una Dell Precision T7400, la cual dispone de 2 Intel Xeon X5482 a 3.2GHz con 32GB RAM y HD SATA2 de 7200rpm.

K	CP1 (20)	CP2 (118)	CP3 (170)	CP4 (221)
Preprocesado	2.07	3.52	7.56	8.40
1	0.40	15.57	71.12	93.81
3	0.40	15.59	73.94	94.12
5	0.39	15.58	74.23	95.03
7	0.38	15.61	74.55	95.43
9	0.37	15.60	75.14	96.77
11	0.36	15.62	75.53	97.01
13	0.35	15.64	75.62	97.69
15	0.35	15.65	80.43	99.86

Tabla 6.6: Tiempos de ejecución (en minutos) para cada conjunto de proteínas (CP) y valor de K.

6.8. Comparación con otras propuestas

Con el objetivo de evaluar la calidad de las predicciones con respecto a otras propuestas en la literatura, se ha planteado, en primer lugar, la comparación con el único método encontrado de predicción de mapas de distancias. Sin embargo, se ha declinado la comparación con este método debido a que, como se comentó en el capítulo 4, el método evalúa la calidad de la predicción discretizando las distancias en tres intervalos hasta 30 angstroms (de 0 a 10, de 10 a 20 y de 20 a 30). Como se explicó en la evaluación de los mapas de distancias (epígrafe 4.4.1), se ha optado por evaluar los mapas de distancias tal como se evalúan los mapas de contactos, convirtiendo los primeros a estos últimos utilizando el umbral de contacto más empleado: 8 angstroms. Por esta razón, la comparación se ha realizado con un método de predicción de mapas de contactos. En concreto, se ha escogido la propuesta de Zhang et al. [Zhang et al., 2005], denominada RBFNN y explicada en el epígrafe 4.6.2.

En la comparación se han utilizado las mismas proteínas de entrenamiento y test que en el artículo de referencia en las mismas condiciones experimentales (con 8 angstroms de umbral de contacto). [Zhang et al., 2005] utiliza *hold-out* como esquema de validación. En concreto utiliza 5 proteínas de test y 5 conjuntos de entrenamiento, uno para cada proteína de test. Los conjuntos de entrenamiento tienen: 9 proteínas para la proteína de test 1TTF, 12 para 1E88, 1 para 1NAR, 12 para 1BTJ_B y 8 para 1J7E_A. [Zhang et al., 2005] emplea la sensibilidad como medida de evaluación (denominada *accuracy* (A_p) por los autores).

En la tabla 6.7 se muestran los resultados de la comparación de ASPpred con el método RBFNN. En esta tabla se consignan dos medidas adicionales que aparecen en el artículo de Zhang et al. [Zhang et al., 2005]. La primera de ellas la denominan los autores N_p y es el número de contactos correctamente predichos (TP, *true positives*). La segunda medida, N_d, es el número total de contactos real de la proteína (equivalente a $TP + FN$). Finalmente, A_p se calcula como N_p/N_d (sensibilidad, $\frac{TP}{TP+FN}$).

Como se puede apreciar en la tabla 6.7, la sensibilidad media de ASPpred es 50.82 % más alta que la de RBFNN. ASPpred ofrece peor sensibilidad que RBFNN sólo para la proteína 1NAR debido a que, como se ha comentado anteriormente, sólo se ha utilizado una proteína de entrenamiento, lo cual parece ser insuficiente para crear un modelo de conocimiento efectivo. En conclusión, los resultados arrojados por ASPpred suponen una mejora importante con respecto al método seleccionado.

Proteína (longitud)	RBFNN			ASPpred		
	N_p	N_d	A_p	N_p	N_d	A_p
1TTF (94)	376	1421	26.46	1307	1421	91.96
1E88 (160)	1006	3352	30.01	3075	3352	91.73
1NAR (290)	3346	10524	31.79	1797	10524	17.07
1BTJ_B (337)	3796	14283	26.58	14026	14283	98.20
1J7E (458)	6589	25026	26.33	23407	25026	93.53
Promedio			27.67			78.49

N_p: contactos bien predichos; N_d: contactos reales; A_p: sensibilidad (%).

Tabla 6.7: Comparación de ASPpred con RBFNN usando 8 Å de umbral de contacto.

6.9. Resumen

En este capítulo se han expuesto los resultados principales obtenidos con el sistema predictor ASPpred. Se han realizado 5 experimentos con 5 conjuntos de proteínas. Se ha llevado a cabo una selección de atributos, para reducir el amplio espacio de búsqueda formado por todas las propiedades de aminoácidos del repositorio AAindex. Se han realizado varios estudios previos acerca de la distribución de las distancias entre aminoácidos en función de sus propiedades. Una vez realizadas las predicciones de los experimentos y mostrados sus resultados, se han analizado diversos aspectos relacionados con las predicciones obtenidas por ASPpred. Finalmente, se ha comparado la propuesta de esta Tesis con otro método de la literatura, arrojando resultados que lo mejoran notablemente.

Capítulo 7

Aplicación a proteínas de interés biológico

7.1. Introducción

En el capítulo anterior se explicaron los experimentos y estudios principales realizados con la propuesta de PEP de esta Tesis, ASPpred. En este capítulo se ha aplicado esta propuesta a dos conjuntos de proteínas de interés biológico: las proteínas de virus y de mitocondrias. El objetivo ha sido el de comprobar la eficacia y calidad de las predicciones realizadas por ASPpred con proteínas localizadas en determinadas regiones subcelulares y con reciente interés en la comunidad científica.

Para cada uno de los conjuntos de proteínas que se abordan en este capítulo se aportan, en primer lugar, varias publicaciones en las que el tipo de proteínas en cuestión cobra especial relevancia. Posteriormente, se indican los detalles de la obtención de las proteínas. Para cada experimentación se ha realizado una selección de propiedades de aminoácidos, la cual también se encuentra explicada. Se indican también todos los detalles de las condiciones experimentales. Finalmente, se muestran y contrastan los resultados obtenidos.

7.2. Predicción de proteínas de virus

7.2.1. Motivación

Dentro de las proteínas presentes en los virus, uno de los tipos de proteínas más interesantes desde el punto de vista biológico es el de las proteínas de cápsides de virus. La cápside o envoltura de un virus es la capa o membrana que cubre y protege el material genético del virus. Dicha cápside está formada por una estructura regular proteínica.

Desde el punto de vista biológico, las proteínas de cápsides de virus han

sido estudiadas y sus estructuras experimentales han sido analizadas, en cuanto a su acoplamiento y organización tridimensional, tal como se demuestra en múltiples y relevantes trabajos recientes [Keef and Twarock, 2009, Azza et al., 2009, Twarock, 2006, Tama and Brooks III, 2005].

Además se han realizado distintas simulaciones tridimensionales de la formación de este tipo de proteínas con el objetivo de comprobar sus propiedades mecánicas mediante nanoindentación [Ahadi et al., 2009], o en la investigación de la cinética estocástica producida en su ensamblaje [Hemberg et al., 2006].

Las proteínas de cápsides de virus han sido también analizadas para predecir sus lugares de interacción con otras proteínas [Carrillo-Tripp et al., 2008], en el marco del clásico problema biomédico del *protein docking*.

7.2.2. Conjuntos de proteínas

Las proteínas que se han utilizado en esta experimentación proceden de la cápside o envoltura de los virus. En concreto, se han recopilado todas las proteínas de este tipo disponibles en la base de datos PDB, eliminando aquellas que son redundantes. De este modo, se ha obtenido un conjunto de proteínas no homólogas con un porcentaje de identidad en la secuencia máximo del 30 %. El conjunto contiene 63 proteínas de cápsides de virus (*viral capsid*, GO ID: 19028). La longitud de la secuencia proteínica más larga es de 1284 aminoácidos.

En la tabla 7.1 se muestran los identificadores PDB de las proteínas que se han utilizado en esta experimentación, organizadas en tres grupos según la longitud L de sus secuencias: $L < 150$, $L150 - 300$ y $L > 300$.

$L < 150$	1TD4	1CD3	2IZW	1C8D	1MUK	3IYH
1C5E	1VDO	1EI7	2VTU	1DZL	1OPO	3IYK
1GFF	1W8X	1F15	2VVF	1EJ6	1P2Z	3IYL
1HGZ	2C0W	1F2N	2WLP	1FN9	1QHD	3JYR
1IFK	2KX4	1JS9	2ZL7	1HX6	1SVA	3KIC
1IFL	2QUD	1STM	3FMG	1IHM	1YUE	3KZ4
1IFP	2VF9	1VPS	3KML	1KVP	2BBD	
1JMU	$L150-300$	1X36	$L > 300$	1LP3	2JHP	
1MSC	1AUY	1ZA7	1A6C	1M1C	2TBV	
1QBE	1C8N	2BUK	1BVP	1M3Y	2XVR	

Tabla 7.1: Las 63 proteínas de cápsides de virus utilizadas para entrenar y predecir con ASPpred (organizadas por la longitud (L) de sus secuencias).

7.2.3. Selección de propiedades

El conjunto de propiedades de aminoácidos que se ha utilizado en esta experimentación ha sido obtenido mediante un proceso de selección de atributos sobre el repositorio completo de 544 propiedades de AAindex.

Tras una exhaustiva exploración de los algoritmos de evaluación de atributos y esquemas de búsqueda disponibles, se han encontrado los mejores resultados utilizando CFS [Hall, 1999] como evaluación y BARS [Ruiz et al., 2008] como esquema de búsqueda.

BARS es un algoritmo aglomerativo debido a su forma de construir los subconjuntos de atributos. El método BARS comienza por la generación de un ranking de atributos. A continuación obtiene pares de atributos a partir de los primeros elementos del ranking en combinación con cada uno de los atributos restantes. Los pares de atributos son ordenados según el valor de la función o algoritmo de evaluación. Este proceso se repite del mismo modo, es decir, los subconjuntos constituidos por los primeros conjuntos de la nueva lista se comparan con el resto de los conjuntos. Al final, el algoritmo BARS devuelve el subconjunto mejor posicionado de todos los subconjuntos evaluados.

El conjunto de datos de entrada utilizado para el proceso de selección de atributos está formado por las proteínas publicadas en el trabajo de Fariselli et al. 2001 [Fariselli et al., 2001]. Este conjunto contiene 173 proteínas con una identidad máxima en la secuencia del 25 %. El conjunto inicial de atributos lo forman las 544 propiedades de AAindex y la clase es la distancia discretizada con umbral de 8 angstroms entre los pares de aminoácidos.

Tras el proceso de selección de atributos, se han obtenido 3 propiedades de aminoácidos, las cuales se muestran en la tabla 7.2.

Nombre	**Descripción**
CHOC760102	Residue accessible surface area in folded protein
RACS820112	Average relative fractional occurrence in ER(i-1)
WEBA780101	RF value in high salt chromatography

Tabla 7.2: La selección de 3 propiedades utilizada.

7.2.4. Configuración de la experimentación

Se ha empleado una validación basada en *leaving-one-out* por dos motivos. Por una parte, debido al reducido número de proteínas utilizadas (63 proteínas), un esquema *leaving-one-out* maximiza el volumen del conjunto de entrenamiento dado un único conjunto de datos para entrenar y predecir. En segundo lugar, se ha perseguido evitar el efecto de la

aleatoriedad introducida en la elección de las bolsas en una validación cruzada con bolsas.

El vector de predicción utilizado en esta experimentación es de tipo A (ver epígrafe 5.4.3). La separación mínima en la secuencia ha sido establecida en 7 aminoácidos. Este valor fue utilizado en [Fariselli et al., 2001] y evita la evaluación de las predicciones más sencillas, tal como se explicó en 4.4.2. Se ha utilizado el carbono beta (CB) como átomo de referencia para el cálculo de distancias entre aminoácidos. El umbral de contacto ha sido 8 angstroms y no se ha realizado ranking TopLx.

7.2.5. Resultados

En la tabla 7.3 se muestran los resultados obtenidos en la experimentación realizada sobre las 63 proteínas de cápsides de virus. Se ha indicado la sensibilidad, precisión, exactitud, especificidad y coeficiente de correlación de Mathews (ver epígrafe 3.4.2) para el total de las proteínas y para cada grupo según la longitud de sus secuencias. Para cada grupo de proteínas se indica entre paréntesis el número de proteínas que incluye.

Proteínas	Sens.	Prec.	Exact.	Especif.	MCC
Todas (63)	0.77	0.75	0.99	0.99	0.75
$L < 150$ (16)	0.85	0.83	0.99	0.99	0.84
$150 \leq L < 300$ (19)	0.80	0.75	0.99	0.99	0.77
$L \geq 300$ (28)	0.75	0.73	0.99	0.99	0.73

Tabla 7.3: Eficacia de ASPpred en la predicción de proteínas de cápsides de virus.

Como se puede apreciar en la tabla 7.3, ASPpred ha obtenido una precisión de 0.75 y sensibilidad 0.77 para el grupo completo de 63 proteínas. Aunque las condiciones experimentales no coinciden debido a que el conjunto de proteínas es distinto, la propuesta de Fariselli et al. 2001 [Fariselli et al., 2001] consiguió una precisión de 0.21 para 173 proteínas (con idéntico átomo de referencia, umbral de contacto y mínima separación en la secuencia).

Generalmente la precisión en la predicción de estructuras de proteínas con secuencias largas (más de 300 aminoácidos) es menor que la obtenida en proteínas con secuencias más cortas. Por ejemplo, en la propuesta de Fariselli et al. 2001 [Fariselli et al., 2001] se obtuvo una precisión de 0.11 para secuencias de 300 o más aminoácidos. ASPpred consigue una precisión de 0.73 para el subconjunto de proteínas de cápsides de virus del mismo orden de longitud (300 o más aminoácidos).

En la figura 7.1 se muestra como ejemplo el mapa de distancias predicho por ASPpred para la proteína de cápside de virus 1M3Y de 413 aminoácidos.

Se ha usado una escala de color para representar los valores de distancias, comenzando desde la mínima distancia (color rojo) hasta la máxima (color azul). Como se puede apreciar en la figura 7.1, la triangular inferior de la matriz de distancias (predicción) es bastante similar a la triangular superior (observación).

La figura 7.2 muestra el mapa de contactos de la misma proteína 1M3Y, obtenido a partir del mapa de distancias mostrado en la figura 7.1 aplicando un umbral de contacto de 8 angstroms. Al igual que el mapa de distancias, existe una gran similitud entre la parte real y la predicha en el mapa de contactos.

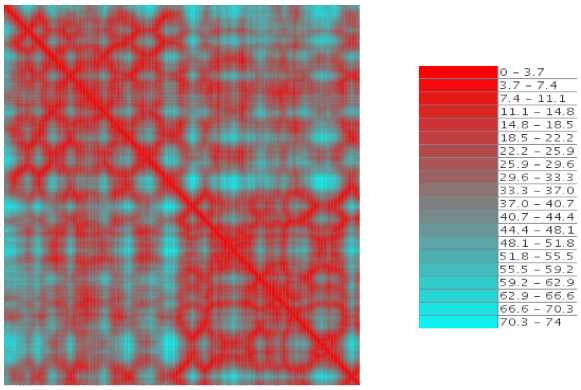


Figura 7.1: Mapa de distancias predicho por ASPpred para la proteína 1M3Y y su escala de color.

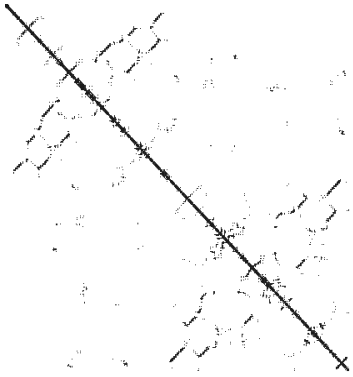


Figura 7.2: Mapa de contactos predicho por ASPpred para la proteína 1M3Y utilizando umbral de contacto de 8 Å.

7.3. Predicción de proteínas de mitocondrias

7.3.1. Motivación

Las mitocondrias son importantes orgánulos de las células eucariotas y se encargan de suministrar la mayor parte de la energía necesaria para la actividad celular. Además, las mitocondrias están asociadas con procesos de la muerte celular y con múltiples enfermedades humanas.

Las mitocondrias están rodeadas por dos membranas bien diferenciadas que separan tres espacios: el citosol, el espacio intermembrana y la matriz mitocondrial. En concreto, la matriz mitocondrial contiene iones, metabolitos a oxidar, ADN circular bicatenario, ARN mitocondrial, ribosomas y proteínas específicas. En la matriz mitocondrial tienen lugar diversas rutas metabólicas clave para la vida, como el ciclo de Krebs y la beta-oxidación de los ácidos grasos.

Las proteínas de la matriz mitocondrial han sido estudiadas y analizadas ampliamente en la literatura, especialmente desde el punto de vista de la clasificación proteínica. La clasificación proteínica consiste, en este contexto, en determinar si una proteína es de mitocondria o no lo es únicamente a partir de su secuencia de aminoácidos. En concreto, se han desarrollado numerosas aproximaciones que permiten predecir si una proteína pertenece al grupo de las mitocondrias [Afridi et al., 2012, Tan et al., 2007, Jiang et al., 2006, Kumar et al., 2006, Guda et al., 2004].

Otro problema también abordado con las proteínas de mitocondrias es el de la predicción del lugar concreto dentro de la mitocondria al que pertenece una proteína, ya que además de proteínas de la matriz también existen proteínas formando cada una de las membranas de la mitocondria. Por ejemplo, el método propuesto por [Du and Li, 2006] utiliza propiedades físico-químicas de aminoácidos (PCP) y su composición (AAC) para predecir el lugar dentro de la mitocondria al que pertenece una proteína a partir de su secuencia de aminoácidos.

7.3.2. Conjuntos de proteínas

Las proteínas que se han utilizado en esta experimentación proceden de la matriz de las mitocondrias. En concreto, se han recopilado todas las proteínas de este tipo disponibles en la base de datos PDB, eliminando aquellas que son redundantes. De este modo, se ha obtenido un conjunto de proteínas no homólogas con un porcentaje de identidad en la secuencia máximo del 30 %. El conjunto contiene 74 proteínas de cápsides de virus (*mitochondrial matrix*, GO ID: 5759). La longitud de la secuencia proteínica más larga es de 1094 aminoácidos.

En la tabla 7.1 se muestran los identificadores PDB de las proteínas que se han utilizado en esta experimentación, organizadas en tres grupos según la longitud L de sus secuencias: $L \leq 300$, $L300 - 450$ y $L > 450$.

$L \leq 300$	2CW6	1CSH	2DFD	3CMQ	1PJ3	3C5E
1BWY	2GRA	1D2E	2E0A	3EXE	1WDK	3DLX
1EFV	2HDH	1F0Y	2IB8	3GH0	1WLE	3E04
1KKC	2O23	1GKZ	2IZZ	3KGW	1ZMD	3IHJ
1MJ3	2WYT	1HW4	2OAT	7AAT	2FGE	3IKL
1QQ2	3ED7	1I4W	2QB7	$L > 450$	2J6L	3IKM
1R4W	3EMN	1OTH	2QFY	1A4E	2JDI	3MW9
1RHS	3QUW	1RX0	2R2N	1CJC	2UXW	3N9Y
1TG6	3ULL	1W6U	3AF0	1G5H	2WYA	3OEE
1XX4	5CYT	2A1H	3BLX	1HR6	2XIJ	3OU5
1ZD8	$L300-450$	2BFD	3BPT	1OHV	2ZT5	3SPA

Tabla 7.4: Las 74 proteínas de matriz de mitocondrias utilizadas para entrenar y predecir con ASPpred (organizadas por la longitud (L) de sus secuencias).

7.3.3. Selección de propiedades

El conjunto de propiedades de aminoácidos que se ha utilizado en esta experimentación ha sido obtenido mediante un proceso de selección de atributos sobre el repositorio completo de 544 propiedades de AAindex.

Tras una exhaustiva exploración de los algoritmos de evaluación de atributos y esquemas de búsqueda disponibles, se han encontrado los mejores resultados utilizando regresión lineal como evaluación y BARS como esquema de búsqueda.

Se ha utilizado regresión lineal como función de evaluación debido a que se encuentra en consonancia con la naturaleza continua de la clase del conjunto de datos de entrada (distancias entre aminoácidos).

El conjunto de datos de entrada utilizado para el proceso de selección de atributos está formado por las proteínas publicadas en el trabajo de Fariselli et al. 2001 [Fariselli et al., 2001]. El conjunto inicial de atributos lo forman las 544 propiedades de AAindex y la clase es la distancia entre los pares de aminoácidos.

Tras el proceso de selección de atributos, se han obtenido 16 propiedades de aminoácidos, las cuales se muestran en la tabla 7.5.

7.3.4. Configuración de la experimentación

Se ha empleado una validación basada en *leaving-one-out* por dos motivos. Por una parte, debido al reducido número de proteínas utilizadas (74 proteínas), un esquema *leaving-one-out* maximiza el volumen del conjunto de entrenamiento dado un único conjunto de datos para entrenar y predecir. En segundo lugar, se ha perseguido evitar el efecto de la

Nombre	Descripción
CHOC760104	Proportion of residues 100 % buried
LEVM760104	Side chain torsion angle phi(AAAR)
MEIH800103	Average side chain orientation angle
PALJ810107	Normalized frequency of alpha-helix in all-alpha class
QIAN880112	Weights for alpha-helix at the window position of 5
WOLS870101	Principal property value z1
ONEK900101	Delta G values for the peptides extrapolated to 0 M urea
BLAM930101	Alpha helix propensity of position 44 in T4 lysozyme
PARS000101	p-Values of mesophilic proteins based on the distributions of B values
NADH010102	Hydropathy scale based on self-information values in the two-state model (9 % accessibility)
SUYM030101	Linker propensity index
WOLR790101	Hydrophobicity index
JACR890101	Weights from the IFH scale
MIYS990103	Optimized relative partition energies - method B
MIYS990104	Optimized relative partition energies - method C
MIYS990105	Optimized relative partition energies - method D

Tabla 7.5: La selección de 16 propiedades utilizada.

aleatoriedad introducida en la elección de las bolsas en una validación cruzada con bolsas.

El vector de predicción utilizado en esta experimentación es de tipo B (ver epígrafe 5.4.3). La separación mínima en la secuencia ha sido establecida en 7 aminoácidos. Este valor fue utilizado en [Fariselli et al., 2001] y evita la evaluación de las predicciones más sencillas, tal como se explicó en 4.4.2. Se ha utilizado el carbono beta (CB) como átomo de referencia para el cálculo de distancias entre aminoácidos. El umbral de contacto ha sido 8 angstroms y no se ha realizado ranking TopLx.

7.3.5. Resultados

En la tabla 7.6 se muestran los resultados obtenidos en la experimentación realizada sobre las 74 proteínas de matriz de mitocondrias. Se ha indicado la sensibilidad, precisión, exactitud, especificidad y coeficiente de correlación de Mathews (ver epígrafe 3.4.2) para el total de las proteínas y para cada grupo según la longitud de sus secuencias. Para cada grupo de proteínas se indica entre paréntesis el número de proteínas que incluye.

Como se puede apreciar en la tabla 7.6, ASPpred ha obtenido una precisión de 0.79 y sensibilidad 0.8 para el grupo completo de 74 proteínas.

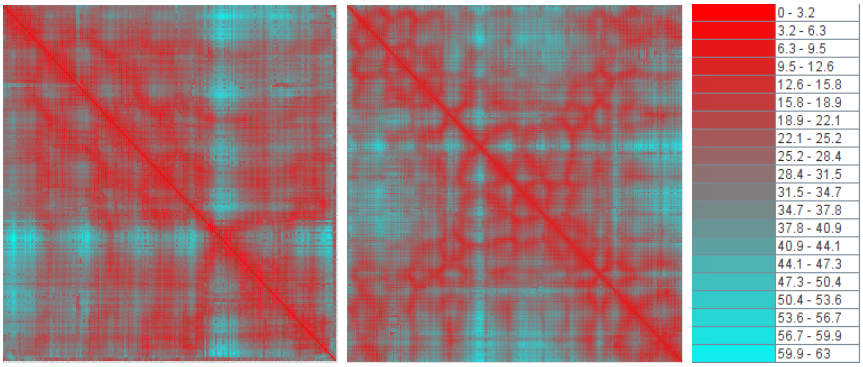
Proteínas	Sens.	Prec.	Exact.	Especif.	MCC
Todas (74)	0.80	0.79	0.97	0.97	0.82
$L \leq 300$ (20)	0.77	0.76	0.98	0.98	0.75
$300 < L \leq 450$ (27)	0.84	0.83	0.99	0.99	0.83
$L > 450$ (27)	0.77	0.76	0.95	0.95	0.82

Tabla 7.6: Eficacia de ASPpred en la predicción de proteínas de matriz de mitocondrias.

En esta experimentación las proteínas de longitud mayor o igual a 300 aminoácidos han sido divididas en dos grupos ($300 < L \leq 450$ y $L > 450$), a diferencia de la experimentación realizada con las proteínas de cápsides de virus.

Aunque las condiciones experimentales no coinciden debido a que el conjunto de proteínas es distinto, el vector de predicción tipo B ha arrojado mejores resultados que el tipo A. En concreto, mientras que ASPpred con vector tipo A obtiene una precisión de 0.73 para proteínas $L \geq 300$ de virus, ASPpred con vector tipo B obtiene 0.83 y 0.76 para proteínas $300 < L \leq 450$ y $L > 450$, respectivamente, de mitocondrias.

En la figura 7.3 se muestran como ejemplo los mapas de distancias predichos por ASPpred para las proteínas de matriz de mitocondrias 1TG6 (277 aminoácidos) y 3BLX (349 aminoácidos). Se ha usado una escala de color para representar los valores de distancias, comenzando desde la mínima distancia (color rojo) hasta la máxima (color azul). Como se puede apreciar, la triangular inferior de las matrices de distancias (predicción) es bastante similar a la triangular superior (observación).



(a) 1TG6 (277 aminoácidos) (b) 3BLX (349 aminoácidos) (c) Escala color

Figura 7.3: Mapas de distancias para las proteínas 1TG6 (a) y 3BLX (b) con su escala de color (c).

7.4. Resumen

En este capítulo se ha aplicado la propuesta ASPpred a dos conjuntos de proteínas de interés biológico: las proteínas de virus y de mitocondrias. La experimentación ha sido realizada sobre proteínas no homólogas, de larga longitud (hasta 1284 aminoácidos) y con separación mínima en la secuencia de 7 aminoácidos. Estas tres condiciones hacen más difíciles las predicciones, como ya se ha explicado anteriormente en esta Tesis.

Se han empleado los dos tipos de vectores de predicción explicados en el capítulo 5, arrojando mejor resultado el tipo B en el caso de las proteínas de mitocondrias. El mejor resultado, en concreto, ha sido de 0.84 de sensibilidad y 0.83 de precisión para las proteínas de mitocondrias de longitud $300 < L \leq 450$. Además, los mapas de distancias dejan ver una gran similitud entre las distancias reales y predichas.

Parte V

Conclusiones

Capítulo 8

Conclusiones y trabajos futuros

8.1. Conclusiones

Las proteínas son las biomoléculas que tienen mayor diversidad estructural y desempeñan multitud de importantes funciones en todos los organismos vivos. Sin embargo, en la formación de las proteínas se producen anomalías que provocan o facilitan el desarrollo de importantes enfermedades como el cáncer o el Alzheimer, siendo de vital importancia el diseño de fármacos que permitan evitar sus desastrosas consecuencias. En dicho diseño de fármacos se precisa disponer de modelos estructurales de proteínas que, pese a que su secuencia es conocida, en la mayoría de los casos su estructura aún se ignora. Es por ello que la predicción de la estructura de una proteína a partir de su secuencia de aminoácidos resulta clave para la cura de este tipo de enfermedades.

La tendencia actual en la literatura de este campo pasa por emplear cada vez un mayor número de proteínas de entrenamiento, necesidad sugerida en gran medida por el uso cada vez más habitual de perfiles de características más extensos y, de este modo, poder cubrir lo máximo posible el gigantesco espacio de búsqueda de este problema. Sin embargo, esta tendencia requiere mayores recursos computacionales y, al ser más extensos los datos de entrada, los modelos de conocimiento basados en éstos son cada vez más complejos e ininteligibles. Además, cada vez son más los métodos que se centran en predecir determinadas partes del problema en lugar de tratar de ser generales o universales, debido a la gran dificultad del problema.

En esta Tesis se ha propuesto un nuevo método para la predicción de mapas de distancias mediante un esquema de vecinos más cercanos empleando propiedades físico-químicas de aminoácidos como entrada.

Se ha probado que proteínas con similares regiones estructurales comparten también, en la mayoría de los casos, ciertas características en sus

secuencias y este comportamiento está en consonancia con el procedimiento de predicción efectuado por el algoritmo de vecinos más cercanos, ya que los atributos de entrada son características de las secuencias y la clase una región de su estructura. Además, el algoritmo de vecinos más cercanos se caracteriza por carecer de un modelo de conocimiento propio, estando éste formado por todos los ejemplos de entrenamiento, eliminando el coste de generación del mismo. A diferencia de otros métodos como las redes neuronales artificiales, las máquinas de soporte vectorial o los algoritmos evolutivos, el algoritmo de vecinos más cercanos apenas posee parametrización, únicamente el número K de vecinos.

Aunque la estructura de datos más utilizada en la literatura para predecir estructuras de proteínas es el mapa de contactos, el mapa de distancias aporta mayor información acerca de la estructura de una proteína, lo cual es beneficioso para la reconstrucción de modelos tridimensionales de mayor calidad. Además, un mapa de distancias puede convertirse en un mapa de contactos con tan sólo aplicar el valor de umbral deseado. De este modo, cualquier aplicación que utilice mapas de contactos puede ser empleada a partir de un mapa de distancias. Y, especialmente, ¿por qué utilizar 8 angstroms como umbral estándar? Se ha observado que las interacciones entre aminoácidos cercanos en la proteína se producen a diferentes distancias, según sea el tipo de interacción y el lugar donde se produzca. Por tanto, fijar cualquier umbral de contacto puede despreciar un buen número de interacciones que se producen a distancias superiores al umbral, al mismo tiempo que puede reconocer interacciones inferiores al umbral que en realidad no lo son. Los mapas de distancias no adolecen de este problema, pues carecen de umbral.

Los resultados obtenidos por la propuesta metodológica de esta Tesis (ASPpred) suponen una mejora del 50.82 % en sensibilidad con respecto a una de las propuestas de predicción de contactos analizada de la literatura.

Se ha comprobado además que la precisión mejora considerablemente cuando el parámetro K se incrementa, desde una precision de 0.38 (para $K = 1$) hasta 0.88 (para $K = 13$). Además estas mejoras de precisión son estadísticamente significativas.

La propuesta ASPpred se ha aplicado a dos conjuntos de proteínas de interés biológico: las proteínas de virus y de mitocondrias. La experimentación ha sido realizada en condiciones que hacen más difíciles las predicciones (proteínas no homólogas, de larga longitud y con separación mínima en la secuencia de 7 aminoácidos). No obstante, se ha alcanzado un 0.84 de sensibilidad y 0.83 de precisión para las proteínas de mitocondrias de longitud $300 < L \leq 450$. Además, los mapas de distancias dejan ver una gran similitud entre las distancias reales y predichas.

8.2. Trabajos futuros

Como trabajos futuros se han identificado varias ideas que podrían mejorar los resultados producidos por ASPpred y hacerlos más competitivos. Por consiguiente, se proponen llevar a la práctica los siguientes puntos:

- Se utilizarán las mutaciones correlacionadas detectadas en los alineamientos múltiples de secuencias como dato de entrada junto a las propiedades físico-químicas. Para ello, se realizará un análisis del emparejamiento directo con el objetivo de evitar los falsos positivos debido a las correlaciones transitivas.

- Se realizarán y documentarán estudios sistemáticos de selección de atributos y su impacto en la sensibilidad y precisión obtenida por ASPpred. En este sentido, se realizará una selección de atributos para cada conjunto de entrenamiento en un esquema de validación cruzada (una selección de atributos para cada bolsa, en una validación cruzada con bolsas), en lugar de hacer una selección de atributos sobre conjuntos de proteínas independientes.

- Se utilizarán algoritmos de editado para reducir el tiempo en la búsqueda de los vecinos más cercanos y eliminar los ejemplos que no aportan información. Además, la reducción de ejemplos deberá tener en cuenta el gran desbalanceo de clases inherente a este problema. Se pretende realizar además estudios sistemáticos de editado y selección de atributos para distintos conjuntos de proteínas.

- Se empleará un ranking Top L/5 de predicciones de contactos y se utilizará una mínima separación en la secuencia de 24 aminoácidos, con el objetivo de hacer más comparables los resultados de ASPpred.

Aparte de los puntos anteriores, se pretende llevar a cabo una propuesta diferente para la predicción de contactos entre aminoácidos basada en la generación de reglas representadas por puntos de la clase positiva situados en diferentes espacios de atributos. Una vez generado el modelo de conocimiento basado en estos puntos-regla, las predicciones se realizarán atendiendo a criterios de vecindad. Esta propuesta promete salvar el problema del desbalanceo de la clase, a la vez que reduce considerablemente el espacio de búsqueda del vecino más cercano en la tarea de predicción, ya que utiliza tan sólo un subconjunto no redundante de los ejemplos de la clase minoritaria.

Parte VI

Apéndices

Apéndice A

Tablas de resultados según clase estructural

Clase estructural CATH	Secuencias	Sens.(μ)	Sens.(σ)	Prec.(μ)	Prec.(σ)
3-Layer(aba) Sandwich	4	0,45	0,00	0,38	0,00
Mainly Beta	4	0,47	0,01	0,62	0,01
Sin clase estructural	49	0,46	0,09	0,53	0,11

Tabla A.1: Evaluación experimento 1 según clase estructural.

Clase estructural CATH	Secuencias	Sens.(μ)	Sens.(σ)	Prec.(μ)	Prec.(σ)
3-Layer(aba) Sandwich	22	0,49	0,19	0,51	0,19
Mainly Beta	27	0,43	0,11	0,47	0,12
2-Layer Sandwich	12	0,53	0,21	0,58	0,18
Few Secondary Structures	1	0,45	0	0,59	0
Roll	10	0,45	0,03	0,48	0,06
Mainly Alpha	19	0,49	0,03	0,53	0,07
Alpha-Beta Complex	7	0,48	0,10	0,46	0,10
Alpha-Beta Barrel	11	0,47	0,16	0,50	0,16
4-Layer Sandwich	1	0,39	0	0,51	0
3-Layer(bba) Sandwich	1	0,37	0	0,38	0
Sin clase estructural	47	0,53	0,19	0,55	0,18

Tabla A.2: Evaluación experimento 2 según clase estructural.

Clase estructural CATH	Secuencias	Sens.(μ)	Sens.(σ)	Prec.(μ)	Prec.(σ)
3-Layer(aba) Sandwich	44	0,42	0,03	0,46	0,06
Mainly Beta	45	0,40	0,03	0,46	0,08
2-Layer Sandwich	18	0,45	0,03	0,51	0,06
Few Secondary Structures	2	0,45	0,02	0,53	0,02
Roll	11	0,45	0,05	0,50	0,07
Mainly Alpha	29	0,50	0,05	0,55	0,08
Alpha-Beta Complex	11	0,46	0,08	0,46	0,08
Alpha-Beta Barrel	17	0,42	0,02	0,42	0,06
4-Layer Sandwich	4	0,42	0,02	0,45	0,06
3-Layer(bba) Sandwich	1	0,39	0	0,36	0
Sin clase estructural	49	0,45	0,09	0,48	0,12

Tabla A.3: Evaluación experimento 3 según clase estructural.

Clase estructural CATH	Secuencias	Sens.(μ)	Sens.(σ)	Prec.(μ)	Prec.(σ)
3-Layer(aba) Sandwich	51	0,43	0,03	0,45	0,06
Mainly Beta	48	0,41	0,03	0,45	0,07
2-Layer Sandwich	23	0,45	0,03	0,51	0,06
Few Secondary Structures	1	0,46	0	0,56	0
Roll	13	0,46	0,03	0,50	0,06
Mainly Alpha	31	0,50	0,05	0,55	0,07
Alpha-Beta Complex	12	0,45	0,10	0,43	0,09
Alpha-Beta Barrel	19	0,45	0,13	0,47	0,13
4-Layer Sandwich	2	0,42	0,01	0,43	0,07
3-Layer(bba) Sandwich	2	0,39	0,02	0,35	0,00
Sin clase estructural	122	0,55	0,19	0,56	0,16

Tabla A.4: Evaluación experimento 4 según clase estructural.

Clase estructural CATH	Secuencias	Sens.(μ)	Sens.(σ)	Prec.(μ)	Prec.(σ)
3-Layer(aba) Sandwich	975	0,46	0,08	0,47	0,09
Mainly Beta	1536	0,45	0,08	0,46	0,09
2-Layer Sandwich	759	0,48	0,07	0,50	0,09
Few Secondary Structures	129	0,60	0,11	0,58	0,10
Roll	350	0,47	0,10	0,50	0,10
Mainly Alpha	1298	0,57	0,10	0,57	0,11
Alpha-Beta Complex	190	0,48	0,11	0,47	0,12
Alpha-Beta Barrel	112	0,48	0,16	0,50	0,17
Alpha-beta prism	2	0,39	0,00	0,36	0,00
5-stranded Propeller	2	0,42	0,00	0,43	0,00
4-Layer Sandwich	86	0,45	0,09	0,44	0,08
3-Layer(bba) Sandwich	28	0,44	0,11	0,45	0,13
Alpha-Beta Horseshoe	7	0,41	0,03	0,34	0,08
Box	17	0,46	0,11	0,43	0,05
Ribosomal Protein L15, K, 2	1	0,65	0	0,66	0
3-Layer(bab) Sandwich	2	0,44	0,02	0,53	0,02
Sin clase estructural	6306	0,53	0,13	0,53	0,13

Tabla A.5: Evaluación experimento 5 según clase estructural.

Apéndice B

Glosario

3D: Indica que un método predice un modelo tridimensional.

AAC: Indica que un método utiliza la composición de aminoácidos como dato de entrada para la predicción.

ABI: Indica que un método está basado en la aproximación biológica *ab-initio*.

ANN: Indica que un método está basado en redes neuronales artificiales (Artificial Neural Networks).

CBR: Indica que un método pertenece al grupo de los de razonamiento basado en casos (Case-based Reasoning).

CMAP: Indica que un método predice mapas de contactos como representación estructural de la proteína.

CMU: Indica que un método utiliza mutaciones correlacionadas a partir de datos evolutivos.

DCA: Indica que un método realiza un análisis del emparejamiento directo para evitar falsos positivos en las mutaciones correlacionadas.

DMAP: Indica que un método predice mapas de distancias como representación estructural de la proteína.

EC: Indica que un método está basado en computación evolutiva (Evolutionary Computation).

ECP: Indica que un método utiliza el perfil de conectividad efectivo como rasgo estructural característico derivado de los mapas de contactos.

ENER: Indica que un método utiliza alguna función de energía.

ENV: Indica que un método utiliza características de aminoácidos en una ventana o entorno a uno dado dentro de la cadena de aminoácidos.

FSEL: Indica que un método realiza selecciones de atributos propias para reducir el espacio de búsqueda.

GLOB: Indica que un método utiliza características derivadas de todos los aminoácidos de la cadena de aminoácidos.

HOM: Indica que un método está basado en la aproximación biológica de las homologías.

INDV: Indica que un método utiliza características de aminoácidos individuales de la cadena de aminoácidos.

OAP: Indica que un método está basado en otro tipo de aproximaciones algorítmicas.

PCP: Indica que un método utiliza propiedades físico-químicas de aminoácidos.

PSSM: Indica que un método utiliza información evolutiva en forma de la matriz PSSM (Position-specific scoring matrices).

Ramachandran plot: Es un gráfico que representa los ángulos de torsión del esqueleto de la proteína. Las regiones del gráfico dejan ver claramente la conformación de la estructura secundaria. Los gráficos de Ramachandran son muy usados para identificar áreas en una estructura con problemas geométricos.

SA: Indica que un método utiliza predicciones de accesibilidad al solvente.

SS: Indica que un método utiliza predicciones de estructura secundaria.

STAT: Indica que un método utiliza técnicas estadísticas en la predicción.

STPR: Indica que un método hace uso de estadísiticas de propensión aplicadas a aminoácidos de la cadena proteínica.

SVM: Indica que un método está basado en máquinas de soporte vectorial (Support Vector Machines).

TANG: Indica que un método predice ángulos de torsión como representación estructural de la proteína.

THR: Indica que un método está basado en la aproximación biológica de *threading*.

Apéndice C

Acrónimos

AA: Aminoácido.

ARFF: Formato de fichero de datos de Weka.

CASP: Critical Assessment of Techniques for Protein Structure Prediction.

CATH: Class, Architecture, Topology and Homologous superfamily.

CV: Validación cruzada.

CVFOLD: Validación cruzada con bolsas.

CVLOU: Técnica de validación *leave-one-out*.

FN: False negatives.

FP: False positives.

GDT-TS: Global Distance Test Total Score.

HOUT: Técnica de validación *hold-out*.

LX: Ranking Top L/x de CASP.

MCC: Mathews correlation coeficient.

MS: Mínima separación en la secuencia.

NN: Nearest neighbors.

NOVAL: Sin validación.

PDB: Protein Data Bank.

PSIBLAST: Position-Specific Iterative Basic Local Alignment Search Tool.

PEP: Predicción de estructuras de proteínas.

RBFNN: Radial basis function neural network.

RMSD: Root mean squared deviation.

RMSE: Root Mean Squared Error.

RRSE: Root Relative Squared Error.

SCOP: Structural Classification of Proteins database.

TN: True negatives.

TP: True positives.

Parte VII

Bibliografía

Bibliografía

[Abu-Doleh et al., 2012] Abu-Doleh, A. A., Al-Jarrah, O. M. and Alkhateeb, A. (2012). Protein contact map prediction using multi-stage hybrid intelligence inference systems. Journal of Biomedical Informatics *45*, 173–183.

[Afridi et al., 2012] Afridi, T. H., Khan, A. and Lee, Y. S. (2012). Mito-GSAAC: mitochondria prediction using genetic ensemble classifier and split amino acid composition. Amino acids *42*, 1443–1454.

[Ahadi et al., 2009] Ahadi, A., Colomo, J. and Evilevitch, A. (2009). Three-dimensional simulation of nanoindentation response of viral capsids. Shape and size effects. The Journal of Physical Chemistry B *113*, 3370–3378.

[Altschul et al., 1997] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research *25*, 3389–3402.

[Anfinsen, 1972] Anfinsen, C. (1972). The formation and stabilization of protein structure. The Biochemical journal *128*, 737–749.

[Asencio-Cortés and Aguilar-Ruiz, 2009] Asencio-Cortés, G. and Aguilar-Ruiz, J. S. (2009). Prediccion de estructuras de proteinas mediante vecinos mas cercanos usando caracteristicas inherentes a los aminoacidos. In Actas de la XIII Conferencia de la Asociacion Espanola para la Inteligencia Artificial (CAEPIA2009).

[Asencio-Cortés and Aguilar-Ruiz, 2010] Asencio-Cortés, G. and Aguilar-Ruiz, J. S. (2010). Importancia de las propiedades fisico-quimicas de los aminoacidos en la prediccion de estructuras de proteinas usando vecinos mas cercanos. In Actas del XV Congreso Espanol sobre Tecnologias y Logica Fuzzy (ESTYLF 2010). ISBN: 978-84-92944-02-6 pp. 459–464,.

[Asencio-Cortés and Aguilar-Ruiz, 2011] Asencio-Cortés, G. and Aguilar-Ruiz, J. S. (2011). Predicting protein distance maps according

to physicochemical properties. Journal of Integrative Bioinformatics *8(3):181.*

[Asencio-Cortés et al., 2011a] Asencio-Cortés, G., Aguilar-Ruiz, J. S. and Chamorro, A. E. M. (2011a). Prediccion de mapas de distancias de proteinas basada en vecinos mas cercanos. In XIV Conferencia de la Asociacion Espanola para la Inteligencia Artificial (CAEPIA2011).

[Asencio-Cortés et al., 2011b] Asencio-Cortés, G., Aguilar-Ruiz, J. S. and Chamorro, A. E. M. (2011b). A Nearest Neighbour-Based Approach for Viral Protein Structure Prediction. In European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBio 2011) Lecture Notes in Computer Science pp. 69–76,.

[Asencio-Cortés et al., 2011c] Asencio-Cortés, G., Aguilar-Ruiz, J. S. and Chamorro, A. E. M. (2011c). Prediction of Protein Distance Maps by Assembling Fragments According to Physicochemical Similarities. In 5th International Conference on Practical Applications of Computational Biology and Bioinformatics (PACBB 2011) number 93 in Advances in Intelligent and Soft Computing pp. 271–278,.

[Asencio-Cortés et al., 2012] Asencio-Cortés, G., Aguilar-Ruiz, J. S., Chamorro, A. E. M., Ruiz, R. and Toca, C. E. S. (2012). Prediction of Mitochondrial Matrix Protein Structures Based on Feature Selection and Fragment Assembly. In European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBio 2012) Lecture Notes in Computer Science pp. 156–167,.

[Ashkenazy et al., 2011] Ashkenazy, H., Unger, R. and Kliger, Y. (2011). Hidden conformations in protein structures. Bioinformatics *27*, 1941–1947.

[Atsushi, 1980] Atsushi, I. (1980). Thermostability and aliphatic index of globular proteins. Journal of Biochemistry *88*, 1895–1898.

[Aydin et al., 2012] Aydin, Z., Thompson, J., Bilmes, J., Baker, D. and Noble, W. S. (2012). Protein Torsion Angle Class Prediction by a Hybrid Architecture of Bayesian and Neural Networks. In 13th International Conference on Bioinformatics and Computational Biology.

[Azza et al., 2009] Azza, S., Cambillau, C., Raoult, D. and Suzan-Monti, M. (2009). Revised Mimivirus major capsid protein sequence reveals intron-containing gene structure and extra domain. BMC molecular biology *10*, 39.

[Bacardit et al., 2012] Bacardit, J., Widera, P., Márquez-Chamorro, A., Divina, F., Aguilar-Ruiz, J. S. and Krasnogor, N. (2012). Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features. Bioinformatics *28*, 2441–2448.

[Bastolla et al., 2008] Bastolla, U., Ortíz, A. R., Porto, M. and Teichert, F. (2008). Effective connectivity profile: a structural representation that evidences the relationship between protein structures and sequences. Proteins: Structure, Function, and Bioinformatics *73*, 872–888.

[Bastolla et al., 2005] Bastolla, U., Porto, M., Roman, H. E. and Vendruscolo, M. (2005). Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. Proteins: Structure, Function, and Bioinformatics *58*, 22–30.

[Berman et al., 2000] Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. and Bourne, P. (2000). The Protein Data Bank. Nucl. Acids Res. *28*, 235–242.

[Björkholm et al., 2009] Björkholm, P., Daniluk, P., Kryshtafovych, A., Fidelis, K., Andersson, R. and Hvidsten, T. R. (2009). Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue–residue contacts. Bioinformatics *25*, 1264–1270.

[Blundell et al., 1987] Blundell, T., Sibanda, B., Sternberg, M. and Thornton, J. (1987). Knowledge-based prediction of protein structures. Nature *326*, 26.

[Burkoff et al., 2013] Burkoff, N. S., Várnai, C. and Wild, D. L. (2013). Predicting protein β-sheet contacts using a maximum entropy-based correlated mutation measure. Bioinformatics *29*, 580–587.

[Calvo et al., 2011] Calvo, J., Ortega, J. and Anguita, M. (2011). Pitagoras-psp: including domain knowledge in a multi-objective approach for protein structure prediction. Neurocomputing *74*, 2675–2682.

[Cao et al., 2004] Cao, H., Ihm, Y., Wang, C.-Z., Morris, J. R., Su, M., Dobbs, D. and Ho, K.-M. (2004). Three-dimensional threading approach to protein structure recognition. Polymer *45*, 687–697.

[Carrillo-Tripp et al., 2008] Carrillo-Tripp, M., Brooks, C. L. and Reddy, V. S. (2008). A novel method to map and compare protein–protein interactions in spherical viral capsids. Proteins: Structure, Function, and Bioinformatics *73*, 644–655.

[Chandonia et al., 2004] Chandonia, J.-M., Hon, G., Walker, N. S., Conte, L. L., Koehl, P., Levitt, M. and Brenner, S. E. (2004). The ASTRAL compendium in 2004. Nucleic acids research *32*, D189–D192.

[Chen and Li, 2010] Chen, P. and Li, J. (2010). Prediction of protein long-range contacts using an ensemble of genetic algorithm classifiers with sequence profile centers. BMC structural biology *10*, S2.

[Cheng and Baldi, 2007] Cheng, J. and Baldi, P. (2007). Improved residue contact prediction using support vector machines and a large feature set. BMC bioinformatics *8*, 113.

[Colubri et al., 2006] Colubri, A., Jha, A. K., Shen, M.-y., Sali, A., Berry, R. S., Sosnick, T. R. and Freed, K. F. (2006). Minimalist representations and the importance of nearest neighbor effects in protein folding simulations. Journal of molecular biology *363*, 835–857.

[Consortium, 2012] Consortium, T. U. (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Research *40*, D71–D75.

[Conte et al., 2000] Conte, L. L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G. and Chothia, C. (2000). SCOP: a structural classification of proteins database. Nucleic acids research *28*, 257–259.

[Cornell et al., 1995] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. and Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. Journal of the American Chemical Society *117*, 5179–5197.

[Cotta, 2003] Cotta, C. (2003). Protein structure prediction using evolutionary algorithms hybridized with backtracking. In Artificial Neural Nets Problem Solving Methods pp. 321–328. Springer.

[Cutello et al., 2006] Cutello, V., Narzisi, G. and Nicosia, G. (2006). A multi-objective evolutionary approach to the protein structure prediction problem. Journal of The Royal Society Interface *3*, 139–151.

[D'Agostino et al., 1990] D'Agostino, R. B., Belanger, A. and D'Agostino Jr, R. B. (1990). A Suggestion for Using Powerful and Informative Tests of Normality. The American Statistician *44*, 316–321.

[Davies et al., 2006] Davies, J., Glasgow, J. and Kuo, T. (2006). VISIO-SPATIAL CASE-BASED REASONING: A CASE STUDY IN PREDICTION OF PROTEIN STRUCTURE. Computational Intelligence *22*, 194–207.

[Di Lena et al., 2012] Di Lena, P., Nagata, K. and Baldi, P. (2012). Deep architectures for protein contact map prediction. Bioinformatics *28*, 2449–2457.

[Dorn and N. de Souza, 2010] Dorn, M. and N. de Souza, O. (2010). A3N: An artificial neural network n-gram-based method to approximate 3-D polypeptides structure prediction. Expert Systems with Applications *37*, 7497–7508.

[Du and Li, 2006] Du, P. and Li, Y. (2006). Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC bioinformatics *7*, 518.

[Eickholt and Cheng, 2012] Eickholt, J. and Cheng, J. (2012). Predicting protein residue-residue contacts using deep networks and boosting. Bioinformatics *28*, 3066–3072.

[Eickholt et al., 2011] Eickholt, J., Wang, Z. and Cheng, J. (2011). A conformation ensemble approach to protein residue-residue contact. BMC structural biology *11*, 38.

[Ekeberg et al., 2013] Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. and Aurell, E. (2013). Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. Physical Review E *87*, 012707.

[Fariselli et al., 2001] Fariselli, P., Olmea, O., Valencia, A. and Casadio, R. (2001). Prediction of contact maps with neural networks and correlated mutations. Protein Engineering *14*, 835–843.

[Finn et al., 2008] Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L. L. and Bateman, A. (2008). The Pfam protein families database. Nucleic Acids Research *36*, D281–D288.

[Gao et al., 2009] Gao, X., Bu, D., Xu, J. and Li, M. (2009). Improving consensus contact prediction via server correlation reduction. BMC structural biology *9*, 28.

[Garcia and Herrera, 2008] Garcia, S. and Herrera, F. (2008). An Extension on Statistical Comparisons of Classifiers over Multiple Data Sets for all Pairwise Comparisons. Journal of Machine Learning Research *9*, 2677–2694.

[Ginalski, 2006] Ginalski, K. (2006). Comparative modeling for protein structure prediction. Current opinion in structural biology *16*, 172–177.

[Glasgow et al., 2006] Glasgow, J., Kuo, T. and Davies, J. (2006). Protein structure from contact maps: A case-based reasoning approach. Information Systems Frontiers *8*, 29–36.

[Griep and Hobohm, 2010] Griep, S. and Hobohm, U. (2010). PDBselect 1992-2009 and PDBfilter-select. Nucl. Acids Res. *38*, D318–319.

[Gu and Bourne, 2003] Gu, J. and Bourne, P. (2003). Structural Bioinformatics (Methods of Biochemical Analysis). Wiley-Blackwell.

[Guda et al., 2004] Guda, C., Fahy, E. and Subramaniam, S. (2004). MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. Bioinformatics *20*, 1785–1794.

[Gupta et al., 2005a] Gupta, N., Mangal, N. and Biswas, S. (2005a). Evolution and similarity evaluation of protein structures in contact map space. Proteins: Structure, Function, and Bioinformatics *59*, 196–204.

[Gupta et al., 2005b] Gupta, N., Mangal, N. and Biswas, S. (2005b). Evolution and Similarity Evaluation of Protein Structures in Contact Map Space. Proteins: Structure, Function, and Bioinformatics *59*, 196–204.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. (2009). The WEKA data mining software: an update. SIGKDD Explorations *11*, 10–18.

[Hall, 1999] Hall, M. A. (1999). Correlation-based feature selection for machine learning. PhD thesis, The University of Waikato.

[Han et al., 2005] Han, S., Lee, B.-C., Yu, S. T., Jeong, C.-S., Lee, S. and Kim, D. (2005). Fold recognition by combining profile-profile alignment and support vector machine. Bioinformatics *21*, 2667–2673.

[Hemberg et al., 2006] Hemberg, M., Yaliraki, S. N. and Barahona, M. (2006). Stochastic kinetics of viral capsid assembly based on detailed protein structures. Biophysical journal *90*, 3029.

[Hoque et al., 2010] Hoque, M. T., Chetty, M., Lewis, A., Sattar, A. and Avery, V. M. (2010). DFS-generated pathways in GA crossover for protein structure prediction. Neurocomputing *73*, 2308–2316.

[Islam and Chetty, 2009] Islam, M. K. and Chetty, M. (2009). Novel Memetic Algorithm for Protein Structure Prediction. In AI 2009: Advances in Artificial Intelligence pp. 412–421. Springer.

[Jiang et al., 2006] Jiang, L., Li, M., Wen, Z., Wang, K. and Diao, Y. (2006). Prediction of mitochondrial proteins using discrete wavelet transform. The protein journal *25*, 241–249.

[Jones et al., 2012] Jones, D. T., Buchan, D. W., Cozzetto, D. and Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics *28*, 184–190.

[Judy et al., 2009] Judy, M., Ravichandran, K. and Murugesan, K. (2009). A multi-objective evolutionary algorithm for protein structure prediction with immune operators. Computer Methods in Biomechanics and Biomedical Engineering *12*, 407–413.

[Karplus, 2009] Karplus, K. (2009). SAM-T08, HMM-based protein structure prediction. Nucleic acids research *37*, W492–W497.

[Kawashima et al., 2008] Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T. and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. Nucleic Acids Res *36*, D202–D205.

[Keef and Twarock, 2009] Keef, T. and Twarock, R. (2009). Affine extensions of the icosahedral group with applications to the three-dimensional organisation of simple viruses. Journal of mathematical biology *59*, 287–313.

[Kitchen et al., 2004] Kitchen, D. B., Decornez, H., Furr, J. R. and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. Nature reviews Drug discovery *3*, 935–949.

[Klein et al., 1984] Klein, P., Kanehisa, M. and DeLisi, C. (1984). Prediction of protein function from sequence properties: Discriminant analysis of a data base. Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology *787*, 221–226.

[Kohonen and Mäkisara, 1989] Kohonen, T. and Mäkisara, K. (1989). The self-organizing feature maps. Physica Scripta *39*, 168.

[Kononenko, 1994] Kononenko, I. (1994). Estimating attributes: analysis and extensions of RELIEF. In Machine Learning: ECML-94 pp. 171–182, Springer.

[Kozma et al., 2013] Kozma, D., Simon, I. and Tusnády, G. E. (2013). PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic acids research *41*, D524–D529.

[Kuhn et al., 1995] Kuhn, L. A., Swanson, C. A., Pique, M. E., Tainer, J. A. and Getzoff, E. D. (1995). Atomic and residue hydrophilicity in the context of folded protein structures. Proteins: Structure, Function, and Bioinformatics *23*, 536–547.

[Kumar et al., 2006] Kumar, M., Verma, R. and Raghava, G. P. (2006). Prediction of mitochondrial proteins using support vector machine and hidden Markov model. Journal of Biological Chemistry *281*, 5357–5363.

[Lapedes et al., 1999] Lapedes, A. S., Giraud, B., Liu, L. and Stormo, G. D. (1999). Correlated mutations in models of protein sequences: phylogenetic and structural effects. Statistics in molecular biology and genetics *33*, 236–256.

[Lavor et al., 2012] Lavor, C., Liberti, L., Maculan, N. and Mucherino, A. (2012). Recent advances on the discretizable molecular distance geometry problem. European Journal of Operational Research *219*, 698–706.

[Lee, 1992] Lee, R. H. (1992). Protein model building using structural homology. Nature *356*, 543–544.

[Li et al., 2011] Li, Y., Fang, Y. and Fang, J. (2011). Predicting residue-residue contacts using random forest models. Bioinformatics *27*, 3379–3384.

[Lippi and Frasconi, 2009] Lippi, M. and Frasconi, P. (2009). Prediction of protein beta-residue contacts by Markov logic networks with grounding-specific weights. Bioinformatics *25*, 2326–2333.

[Liu et al., 2006] Liu, G., Zhu, Y., Zhou, W., Zhou, C. and Wang, R. (2006). Prediction of contact maps using modified transiently chaotic neural network. In Advances in Neural Networks-ISNN 2006 pp. 696–701. Springer.

[Lo et al., 2009] Lo, A., Chiu, Y.-Y., Rødland, E. A., Lyu, P.-C., Sung, T.-Y. and Hsu, W.-L. (2009). Predicting helix-helix interactions from residue contacts in membrane proteins. Bioinformatics *25*, 996–1003.

[MacCallum, 2004] MacCallum, R. M. (2004). Striped sheets and protein contact prediction. Bioinformatics *20*, i224–i231.

[Marks et al., 2011] Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R. and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS One *6*, e28766.

[Marks et al., 2012] Marks, D. S., Hopf, T. A. and Sander, C. (2012). Protein structure prediction from sequence variation. Nature biotechnology *30*, 1072–1080.

[Márquez-Chamorro et al., 2011] Márquez-Chamorro, A., Divina, F., Aguilar-Ruiz, J. and Asencio-Cortes, G. (2011). Un Algoritmo Genético para la Predicción de Mapas de Contacto Basado en Propiedades de Aminoácidos. In Actas de la XIV Conferencia de la Asociación Espanola para la Inteligencia Artificial (CAEPIA 2011).

[Márquez-Chamorro et al., 2012] Márquez-Chamorro, A. E., Asencio-Cortes, G., Divina, F. and Aguilar-Ruiz, J. S. (2012). Evolutionary

decision rules for predicting protein contact maps. Pattern Analysis and Applications -, 1–13.

[Márquez-Chamorro et al., 2011] Márquez-Chamorro, A. E., Divina, F., Aguilar-Ruiz, J. S. and Asencio-Cortés, G. (2011). A multi-objective genetic algorithm for the Protein Structure Prediction. In 11th International Conference on Intelligent Systems Design and Applications (ISDA) pp. 1086–1090, IEEE.

[Márquez-Chamorro et al., 2012] Márquez-Chamorro, A. E., Divina, F., Aguilar-Ruiz, J. S., Bacardit, J., Asencio-Cortés, G. and Santiesteban-Toca, C. E. (2012). A NSGA-II algorithm for the residue-residue contact prediction. In Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBio 2012) number 7246 in Lecture Notes in Computer Science pp. 234–244. Springer.

[Márquez-Chamorro et al., 2011a] Márquez-Chamorro, A. E., Divina, F., Aguilar-Ruiz, J. S. and Cortés, G. A. (2011a). An evolutionary approach for protein contact map prediction. In Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBio 2011) number 6623 in Lecture Notes in Computer Science pp. 101–110. Springer.

[Márquez-Chamorro et al., 2011b] Márquez-Chamorro, A. E., Divina, F., Aguilar-Ruiz, J. S. and Cortés, G. A. (2011b). Residue-Residue Contact Prediction Based on Evolutionary Computation. In 5th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2011) number 93 in Advances in Intelligent and Soft Computing pp. 279–283, Springer.

[Márquez-Chamorro et al., 2010] Márquez-Chamorro, A. E., Divina, F., Ruiz, J. S. A. and Cortés, G. A. (2010). Alpha helix prediction based on evolutionary computation. In Pattern Recognition in Bioinformatics number 6282 in Lecture Notes in Computer Science pp. 358–367. Springer.

[McGuffin et al., 2000] McGuffin, L. J., Bryson, K. and Jones, D. T. (2000). The PSIPRED protein structure prediction server. Bioinformatics *16*, 404–405.

[Miyazawa, 2013] Miyazawa, S. (2013). Prediction of Contact Residue Pairs Based on Co-Substitution between Sites in Protein Structures. PloS one *8*, e54252.

[Morcos et al., 2011] Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T. and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences *108*, E1293–E1301.

[Moult et al., 2011] Moult, J., Fidelis, K., Kryshtafovych, A. and Tramontano, A. (2011). Critical assessment of methods of protein structure prediction (CASP)—round IX. Proteins: Structure, Function, and Bioinformatics *79*, 1–5.

[Mucherino et al., 2012] Mucherino, A., Lavor, C. and Liberti, L. (2012). The discretizable distance geometry problem. Optimization Letters *6*, 1671–1686.

[Noguchi and Akiyama, 2003] Noguchi, T. and Akiyama, Y. (2003). PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. Nucleic acids research *31*, 492–493.

[Nugent and Jones, 2012] Nugent, T. and Jones, D. T. (2012). Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. Proceedings of the National Academy of Sciences *109*, E1540–E1547.

[Orengo et al., 2002] Orengo, C. A., Bray, J. E., Buchan, D. W. A., Harrison, A., Lee, D., Pearl, F. M. G., Sillitoe, I., Todd, A. E. and Thornton, J. M. (2002). The CATH protein family database: a resource for structural and functional annotation of genomes. Proteomics *2*, 11–21.

[Pollastri et al., 2002] Pollastri, G., Baldi, P., Fariselli, P. and Casadio, R. (2002). Prediction of coordination number and relative solvent accessibility in proteins. Proteins: Structure, Function, and Bioinformatics *47*, 142–153.

[Pontius et al., 1996] Pontius, J., Richelle, J., Wodak, S. J. et al. (1996). Deviations from standard atomic volumes as a quality measure for protein crystal structures. Journal of molecular biology *264*, 121–136.

[Punta et al., 2012] Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J. et al. (2012). The Pfam protein families database. Nucleic acids research *40*, D290–D301.

[Punta and Rost, 2005] Punta, M. and Rost, B. (2005). PROFcon: novel prediction of long-range contacts. Bioinformatics *21*, 2960–2968.

[Radzicka and Wolfenden, 1988] Radzicka, A. and Wolfenden, R. (1988). Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. Biochemistry *27*, 1664–1670.

[Rajgaria et al., 2009] Rajgaria, R., McAllister, S. and Floudas, C. (2009). Towards accurate residue-residue hydrophobic contact prediction for α

helical proteins via integer linear optimization. Proteins: Structure, Function, and Bioinformatics *74*, 929–947.

[Rajgaria et al., 2010] Rajgaria, R., Wei, Y. and Floudas, C. (2010). Contact prediction for beta and alpha-beta proteins using integer linear optimization and its impact on the first principles 3D structure prediction method ASTRO-FOLD. Proteins: Structure, Function, and Bioinformatics *78*, 1825–1846.

[Raman et al., 2006] Raman, P., Cherezov, V. and Caffrey, M. (2006). The membrane protein data bank. Cellular and molecular life sciences *63*, 36–51.

[Rost and Eyrich, 2001] Rost, B. and Eyrich, V. A. (2001). EVA: large-scale analysis of secondary structure prediction. Proteins: Structure, Function, and Bioinformatics *45*, 192–199.

[Rost et al., 1997] Rost, B., Schneider, R., Sander, C. et al. (1997). Protein fold recognition by prediction-based threading. Journal of molecular biology *270*, 471–480.

[Ruiz et al., 2008] Ruiz, R., Riquelme, J. C. and Aguilar-Ruiz, J. S. (2008). Best Agglomerative Ranked Subset for Feature Selection. Journal of Machine Learning Research - Proceedings Track *4*, 148–162.

[Sánchez-Linares et al., 2012] Sánchez-Linares, I., Pérez-Sánchez, H., Cecilia, J. M. and García, J. M. (2012). High-Throughput parallel blind Virtual Screening using BINDSURF. BMC bioinformatics *13*, S13.

[Sander and Schneider, 1991] Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins: Structure, Function, and Bioinformatics *9*, 56–68.

[Sander et al., 2006] Sander, O., Sommer, I. and Lengauer, T. (2006). Local protein structure prediction using discriminative models. BMC bioinformatics *7*, 14.

[Santiesteban-Toca et al., 2012] Santiesteban-Toca, C. E., Asencio-Cortés, G., Márquez-Chamorro, A. E. and Aguilar-Ruiz, J. S. (2012). Short-Range interactions and decision tree-based protein contact map predictor. In Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBio 2012) number 7246 in Lecture Notes in Computer Science pp. 224–233. Springer.

[Santiesteban-Toca et al., 2011] Santiesteban-Toca, C. E., Chamorro, A. E. M., Cortés, G. A. and Aguilar-Ruiz, J. S. (2011). A decision tree-based

method for protein contact map prediction. In Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBio 2011) number 6623 in Lecture Notes in Computer Science pp. 153–158. Springer.

[Savojardo et al., 2011] Savojardo, C., Fariselli, P., Alhamdoosh, M., Martelli, P. L., Pierleoni, A. and Casadio, R. (2011). Improving the prediction of disulfide bonds in Eukaryotes with machine learning methods and protein subcellular localization. Bioinformatics *27*, 2224–2230.

[Savojardo et al., 2013] Savojardo, C., Fariselli, P., Martelli, P. L. and Casadio, R. (2013). Prediction of disulfide connectivity in proteins with machine-learning methods and correlated mutations. BMC bioinformatics *14*, S10.

[Shackelford and Karplus, 2007] Shackelford, G. and Karplus, K. (2007). Contact prediction using mutual information and neural nets. Proteins: Structure, Function, and Bioinformatics *69*, 159–164.

[Shell et al., 2009] Shell, M. S., Ozkan, S. B., Voelz, V., Wu, G. A. and Dill, K. A. (2009). Blind test of physics-based prediction of protein structures. Biophysical journal *96*, 917–924.

[Shi et al., 2004] Shi, S. Y., Suganthan, P. and Deb, K. (2004). Multiclass protein fold recognition using multiobjective evolutionary algorithms. In Computational Intelligence in Bioinformatics and Computational Biology, 2004. CIBCB'04. Proceedings of the 2004 IEEE Symposium on pp. 61–66, IEEE.

[Shi et al., 2008] Shi, Y., Zhou, J., Arndt, D., Wishart, D. and Lin, G. (2008). Protein contact order prediction from primary sequences. BMC bioinformatics *9*, 255.

[Sułkowska et al., 2012] Sułkowska, J. I., Morcos, F., Weigt, M., Hwa, T. and Onuchic, J. N. (2012). Genomics-aided structure prediction. Proceedings of the National Academy of Sciences *109*, 10340–10345.

[Tama and Brooks III, 2005] Tama, F. and Brooks III, C. L. (2005). Diversity and identity of mechanical properties of icosahedral viral capsids studied with elastic network normal mode analysis. Journal of molecular biology *345*, 299–314.

[Tan et al., 2007] Tan, F., Feng, X., Fang, Z., Li, M., Guo, Y. and Jiang, L. (2007). Prediction of mitochondrial proteins based on genetic algorithm–partial least squares and support vector machine. Amino Acids *33*, 669–675.

[Tegge et al., 2009] Tegge, A. N., Wang, Z., Eickholt, J. and Cheng, J. (2009). NNcon: improved protein contact map prediction using 2D-recursive neural networks. Nucleic Acids Research *37*, W515–W518.

[Twarock, 2006] Twarock, R. (2006). Mathematical virology: a novel approach to the structure and assembly of viruses. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences *364*, 3357–3373.

[Vassura et al., 2011] Vassura, M., Di Lena, P., Margara, L., Mirto, M., Aloisio, G., Fariselli, P., Casadio, R. et al. (2011). Blurring contact maps of thousands of proteins: what we can learn by reconstructing 3 D structure. BioData mining *4*, 1.

[Vullo et al., 2006] Vullo, A., Walsh, I. and Pollastri, G. (2006). A two-stage approach for improved prediction of residue contact maps. BMC bioinformatics *7*, 180.

[Walsh et al., 2009] Walsh, I., Baù, D., Martin, A., Mooney, C., Vullo, A. and Pollastri, G. (2009). Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. BMC structural biology *9*, 5.

[Wang and Dunbrack, 2003] Wang, G. and Dunbrack, R. (2003). PISCES: a protein sequence culling server. Bioinformatics (Oxford, England) *19*, 1589–1591.

[Wang et al., 2011] Wang, X.-F., Chen, Z., Wang, C., Yan, R.-X., Zhang, Z. and Song, J. (2011). Predicting residue-residue contacts and helix-helix interactions in transmembrane proteins using an integrative feature-based random forest approach. PloS one *6*, e26767.

[Wang et al., 2010] Wang, Z., Eickholt, J. and Cheng, J. (2010). MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. Bioinformatics *26*, 882–888.

[Wei and Floudas, 2011] Wei, Y. and Floudas, C. (2011). Enhanced inter-helical residue contact prediction in transmembrane proteins. Chemical engineering science *66*, 4356–4369.

[Widera, 2010] Widera, P. (2010). Automated design of energy functions for protein structure prediction by means of genetic programming and improved structure similarity assessment. PhD thesis, University of Nottingham.

[Wolff et al., 2008] Wolff, K., Vendruscolo, M. and Porto, M. (2008). Stochastic reconstruction of protein structures from effective connectivity profiles. BMC Biophysics *1*, 5.

[Wolff et al., 2010] Wolff, K., Vendruscolo, M. and Porto, M. (2010). Efficient identification of near-native conformations in ab initio protein structure prediction using structural profiles. Proteins: Structure, Function, and Bioinformatics *78*, 249–258.

[Wu et al., 2011] Wu, S., Szilagyi, A. and Zhang, Y. (2011). Improving protein structure prediction using multiple sequence-based contact predictions. Structure *19*, 1182–1191.

[Wu and Zhang, 2008] Wu, S. and Zhang, Y. (2008). A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics *24*, 924–931.

[Xue et al., 2009] Xue, B., Faraggi, E. and Zhou, Y. (2009). Predicting residue-residue contact maps by a two-layer, integrated neural-network method. Proteins: Structure, Function, and Bioinformatics *76*, 176–183.

[Yang and Chen, 2011] Yang, J.-Y. and Chen, X. (2011). A consensus approach to predicting protein contact map via logistic regression. In Bioinformatics Research and Applications pp. 136–147. Springer.

[Zemla, 2003] Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res *31*, 3370–3374.

[Zhang and Han, 2007] Zhang, G.-Z. and Han, K. (2007). Hepatitis C virus contact map prediction based on binary encoding strategy. Computational Biology and Chemistry *31*, 233–238.

[Zhang et al., 2005] Zhang, G.-Z., Huang, D. and Quan, Z. (2005). Combining a binary input encoding scheme with RBFNN for globulin protein inter-residue contact map prediction. Pattern Recognition Letters *26*, 1543–1553.

[Zhang and Huang, 2004a] Zhang, G.-Z. and Huang, D.-S. (2004a). Prediction of inter-residue contacts map based on genetic algorithm optimized radial basis function neural network and binary input encoding scheme. Journal of computer-aided molecular design *18*, 797–810.

[Zhang and Huang, 2004b] Zhang, G.-Z. and Huang, D.-S. (2004b). Combing genetic algorithm with neural network technique for protein inter-residue spatial distance prediction. In Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on vol. 3, pp. 1687–1691, IEEE.

[Zhang et al., 2010a] Zhang, J., Wang, Q., Barz, B., He, Z., Kosztin, I., Shang, Y. and Xu, D. (2010a). MUFOLD: A new solution for protein 3D structure prediction. Proteins: Structure, Function, and Bioinformatics *78*, 1137–1152.

[Zhang et al., 2010b] Zhang, X., Wang, T., Luo, H., Yang, J., Deng, Y., Tang, J. and Yang, M. (2010b). 3D Protein structure prediction with genetic tabu search algorithm. BMC Systems Biology *4*, S6.

[Zhang, 2008] Zhang, Y. (2008). Progress and challenges in protein structure prediction. Current opinion in structural biology *18*, 342–348.

[Zhang, 2009] Zhang, Y. (2009). I-TASSER: Fully automated protein structure prediction in CASP8. Proteins: Structure, Function, and Bioinformatics *77*, 100–113.

[Zhang and Skolnick, 2005] Zhang, Y. and Skolnick, J. (2005). The protein structure prediction problem could be solved using the current PDB library. Proceedings of the National Academy of Sciences of the United States of America *102*, 1029–1034.

[Zhou et al., 2011] Zhou, Y., Duan, Y., Yang, Y., Faraggi, E. and Lei, H. (2011). Trends in template/fragment-free protein structure prediction. Theoretical chemistry accounts *128*, 3–16.