Construcción de un Clasificador de Polipéptidos Multi-criterio

Tesis de Licenciatura en Ciencias de la Computación

María Marta Ponzoni

martaponzoni@gmail.com

LU 127/06

Directores

Pablo E. Martínez López

Pablo Daniel Ghiringhelli

12 de noviembre de 2010

Departamento de Computación

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

De niña una vez le pregunté a mi padre si quería más a mi madre o a mí y el respondió: "Las quiero a ambas, de modos diferentes".

Quiero dedicar esta tesis a mis padres, a mi abuela, y a la persona que quiero de un modo diferente.

Agradecimientos

A mi familia, por apoyarme siempre, a lo largo de toda mi vida.

A Pablo, por ser mi soporte y mi equilibrio, y por haber sido mi compañero implícito en esta tesis.

A Fidel, por ser amigo y ser maestro, y por enseñarme que, como dijo alguien alguna vez: "lo importante no es llegar, lo importante es el camino".

A Daniel, por guiarme en el oscuro pero maravilloso mundo de las proteínas y los ácidos nucleicos, y por estar siempre dispuesto a contarme un poco más.

Y a ambos, por aceptar el desafío de una tesis juntos.

A Pablo Rodríguez, a Devdatt Dubhashi, a Vinay Jethava y a Lutz Krause, por los chats y mails tan enriquecedores.

A todos mis compañeros, con los que compartí mi carrera. En especial a Eddy, a Fede, a Ger, a Vicky y a Diego.

Y a todos los grandes docentes e investigadores que tiene esta hermosa facultad.

Resumen

Los baculovirus son virus eucarióticos que atacan específicamente a ciertas especies de artrópodos. Por este motivo, resultan ser excelentes candidatos para ser usados como bioinsecticidas específicos, ya que se ha demostrado que no afectan a organismos distintos del blanco, siendo inocuos para plantas, vertebrados e invertebrados no blanco. A pesar de esta ventaja, en comparación con los insecticidas de origen químico, cuentan con una limitación importante: su lento modo de acción. Por ello, los baculovirus se han mejorado genéticamente para aumentar su poder infectivo.

Existen proteínas muy importantes en el proceso de infección de las células del huésped, como la IE-1, la GP64 o la P74, cuyas funciones han sido ampliamente estudiadas. Pero más importante aún es la inmensa gama de proteínas que todavía se desconocen. Y es éste uno de nuestros motores. Dado que la certificación de la existencia de una proteína mediante experimentación en mesada húmeda es sumamente costosa en términos de tiempo y dinero, en este trabajo buscaremos ayudar en el proceso de selección de polipéptidos candidatos mediante métodos computacionales.

Como herramienta para conseguir el objetivo propuesto usaremos una implementación de máquina de vectores de soporte (*support vector machine, SVM*). Basándonos en criterios composicionales, como la frecuencia o la periodicidad de los aminoácidos dentro del polipéptido, construiremos vectores con los cuales entrenaremos y testearemos el desempeño de la SVM. Probaremos diversos criterios, kernels y conjuntos de datos de entrenamiento y seleccionaremos las combinaciones que logren una mejor clasificación de los polipéptidos, para así finalmente conformar un clasificador multi-criterio que pueda ser utilizado en la selección de los polipéptidos más prometedores.

El clasificador tendrá amplios fines prácticos. En el caso de proteomas ya anotados*, se empleará para detectar polipéptidos mal anotados (por ejemplo, para encontrar posibles proteínas funcionales que fueron equivocadamente excluidas en la anotación) y para seleccionar aquellos polipéptidos anotados que tengan mayor probabilidad de ser proteínas funcionales para darles prioridad al momento de realizar los experimentos de mesada húmeda. En el caso de genomas novedosos, secuenciados posteriormente a la finalización de este trabajo, el clasificador será una herramienta fundamental para poder definir rápidamente la fracción anotable dentro del proteoma teórico completo. De esta manera, aún cuando no descartamos el uso de otras metodologías confirmatorias, se agilizará significativamente el proceso de anotación.

* Anotación en las bases de datos de secuencias es el proceso de registro de la información biológica vinculada a la secuencia que se deposita (por ejemplo: regiones promotoras, marcos de lectura abiertos, proteínas codificadas, asignación de funciones, etc.) en un formato compatible para ser almacenada en una base de datos.

Abstract

Baculoviruses are eukaryotic viruses that attack specifically certain species of arthropods. For this reason, they are excellent candidates for being used as specific bioinsecticides, because it has been proved that they don't affect organisms different from target, being harmless for plants, vertebrates and non-target invertebrates. In spite of this advantage, compared to insecticides of chemical origin, they have an important limitation: they act slower. For that reason, baculoviruses have been genetically modified to increment their infectious power.

Some proteins are very important in the infection process of the host cells, like IE-1, GP64 and P74, whose functions have been deeply studied. But more important is the huge amount of proteins which haven't been discovered yet. And this is one of our motivations. Given that certification of the existence of a protein by means of laboratory experimentation is highly expensive in terms of time and money, in this work we will try to help in the selection of candidate polypeptides by means of computational methods.

In order to achieve our goal we will use an implementation of support vector machine (SVM). Using compositional criteria, such as the frequency or periodicity of the aminoacids in the polypeptide, we will build vectors to train and test the SVM performance. We will try different criteria, kernels and training data sets and choose the combinations that achieve the best classification of polypeptides, to finally make a multi-criterion classifier which could be used in the selection of the more promising polypeptides.

The classifier will have plenty of practical uses. In the case of proteomes which are already annotated*, it will be used to detect incorrectly annotated polypeptides (for example, to find out possible functional proteins that have been omitted in the annotation) and to choose those annotated polypeptides with higher probability of being a functional protein to give them priority at the moment of doing the laboratory experiments. In the case of new genomes, secuenciated after the finalization of this work, the classifier will be a fundamental tool for quickly defining the annotable fraction of the theoretical complete proteome. In this way, although we don't discard the use of other confirmatory methodologies, annotation process will be significantly speeded up.

* Annotation on sequence databases is the registration process of biological information associated with the sequence being deposited (for example: promoter regions, open reading frames, codified proteins, functional assignment, etc.) in a compatible format for being saved in a database.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. El problema

Los virus son entes biológicos cuya replicación sólo puede realizarse en el interior de células de seres vivos. Infectan todo tipo de organismos: desde plantas o animales hasta bacterias. La estructura elemental es un ácido nucleico (ADN o ARN) asociado con proteínas para constituir la nucleocápside y pueden estar cubiertos por una cápside de naturaleza proteica y/o envolturas de tipo membranoso.

Los baculovirus son una clase de virus que ataca específicamente a ciertas especies de artrópodos. Por este motivo, resultan ser excelentes candidatos para ser usados como bioinsecticidas específicos. Pero en comparación con otros métodos insecticidas de origen químico, estos organismos cuentan con una limitación importante: su lento modo de acción. Por ello, hace ya varias décadas los científicos han buscado mejorarlos genéticamente para aumentar su poder infectivo.

Existen proteínas muy importantes en el proceso de infección de las células del huésped, como la IE-1, la GP64 o la P74, cuyas funciones han sido ampliamente estudiadas. De la misma forma, existen muchas proteínas que aún no han sido descubiertas.

Día a día nuevos genomas de baculovirus son descriptos por científicos de todo el mundo. Cada uno de ellos contiene un conjunto de marcos de lectura abiertos (*open reading frames, ORFs*[*]). En base a cada genoma se define un orfeoma, es decir un conjunto de ORFs seleccionados por considerarse que poseen una mayor

[*]Los ORFs son porciones de ADN cuya secuencia nucleotídica teóricamente codifica para una proteína.

probabilidad de codificar para proteínas. A partir del orfeoma construido se obtiene un proteoma teórico que contiene la traducción de cada uno de los ORFs elegidos a su respectivo polipéptido, es decir a su proteína teórica.

Del conjunto de polipéptidos definidos en un proteoma teórico de baculovirus, no todos llegan a expresarse. Por medio de experimentación en mesada húmeda puede comprobarse la existencia de algunas proteínas en los organismos, pero este mecanismo es muy costoso en términos de tiempo y dinero. Estos resultados, además, son parciales, ya que la detección de la proteína indica que el ORF verdaderamente la codificaba, pero la no detección de la misma no aporta mayor información. Esto se debe a que, en distintas etapas del ciclo biológico, no todas las proteínas del organismo se encuentran presentes en el mismo.

Dadas las circunstancias, sería muy beneficioso contar con un mecanismo de preselección de polipéptidos candidatos, para así orientar los experimentos para la comprobación de la existencia de los mismos en baculovirus.

1.2. Solución propuesta

En este trabajo intentaremos resolver mediante métodos computacionales el problema de preseleccionar un conjunto de polipéptidos candidatos para que luego se verifique su existencia por medio de métodos de mesada húmeda. Para ello partiremos de proteomas ya definidos, no ocupándonos de las fases previas a su creación, antes descriptas. El objetivo será construir, entrenar y realizar pruebas de la efectividad de un clasificador capaz de determinar, con el menor error posible, cuáles de los polipéptidos definidos en los proteomas son factibles de expresarse en baculovirus. En particular, el algoritmo de clasificación que utilizaremos es el que nos proveen las SVMs [Vapnik (2000)].

La clasificación mediante SVMs (que explicamos en mayor profundidad en el capítulo 2) consta de dos fases. En la primera se entrena la SVM con elementos que fueron previamente etiquetados como positivos o negativos. En una segunda fase se testea el desempeño de la SVM con elementos sin etiquetar para que la misma clasifique, en base a su conocimiento, en positivos y negativos. Dada la imposibilidad de contar con casos negativos reales, es decir, con polipéptidos definidos en los proteomas que se haya comprobado no se expresan en baculovirus (ver sección 1.1), uno de los desafíos de este trabajo es decidir qué conjuntos de datos se utilizarán como negativos para el entrenamiento de la SVM.

En el capítulo 3 evaluaremos la utilización de dos conjuntos de datos que estimamos contienen un porcentaje significativo de casos negativos. Uno de ellos contendrá los polipéptidos que no fueron anotados en su correspondiente proteoma porque los autores consideraron muy poco probable su existencia, y el segundo aquellos que si bien fueron anotados por sus autores, su existencia no fue demostrada aún mediante procedimentos de mesada húmeda.

Otro desafío interesante que abordaremos es la elección de un conjunto de criterios que puedan complementarse para realizar una clasificación adecuada de los polipéptidos. Los criterios que evaluaremos serán de tipo composicional, como la frecuencia o la periodicidad de los aminoácidos dentro del polipéptido. Además, para cada uno de ellos, evaluaremos varios kernels y escogeremos el que logre un mejor desempeño en la clasificación.

Una vez seleccionadas las combinaciones que logren una mejor clasificación de los polipéptidos deberemos decidir de qué manera combinar los resultados obtenidos para construir un clasificador multi-criterio que pueda ser utilizado en la selección de los polipéptidos más prometedores. Por último, propondremos un protocolo para la anotación de polipéptidos utilizando nuestro clasificador y evaluaremos los resultados de experimentos ciegos en los cuales lo desafiaremos simulando situaciones reales.

1.3. Trabajo relacionado

El algoritmo de clasificación provisto por las SVMs [Vapnik (2000)] ha sido ampliamente utilizado en diversas aplicaciones como la clasificación de textos [Venkatesh & Sureshkumar (2009), Zhu *et al.* (2009)], el reconocimiento de rostros [Huang *et al.* (2002), Heisele *et al.* (2001)], el análisis de la escritura [Wu *et al.* (2005)], entre otras.

En los últimos años, las SVMs han empezado a ganar importancia en el ámbito de la bioinformática. Entre las aplicaciones relacionadas con el estudio de proteínas se destacan la búsqueda de similitudes en cadenas [Jaakkola *et al.* (1999), Liao & Noble (2002)] y la predicción de estructura y función biológica [Wang *et al.* (2008), Vert (2002)].

En diciembre de 2006 se publicó un trabajo orientado a la clasificación de marcos de lectura abiertos en organismos procariotas [Krause *et al.* (2006)]. En él se presenta GISMO, una herramienta que combina la utilización de un modelo oculto de Markov (*Hidden Markov Model, HMM*) para la búsqueda de dominios

proteicos y una SVM capaz de identificar regiones codificantes basándose en criterios composicionales.

Uno de los objetivos de esta tesis es realizar un aporte similar al de GISMO, pero orientado a un grupo particular de organismos cuya complejidad es comparable a la de organismos eucariotas: los baculovirus. Además, nos planteamos otro objetivo, no abordado por los autores de GISMO: desarrollar un mecanismo de priorización de polipéptidos según su probabilidad de ser proteínas funcionales, para orientar los experimentos de mesada húmeda.

1.4. Estructura de la tesis

En el capítulo 2 presentamos un resumen de los conceptos fundamentales detrás de las SVMs, que servirá para comprender mejor el resto de la tesis. Además describimos brevemente PyML, la implementación de SVM que utilizamos en este trabajo.

A continuación, en el capítulo 3 detallamos los experimentos realizados en el transcurso de la tesis, que dieron lugar a las decisiones tomadas a la hora de elegir cómo construir, entrenar y testear las distintas instancias de SVM utilizadas para la construcción del clasificador de polipéptidos multi-criterio.

Luego, en el capítulo 4 proponemos un algoritmo de anotación de genomas baculovirales y realizamos un experimento en el que simulamos uno de los casos de uso futuro para medir el desempeño del clasificador construido.

Finalmente, en el capítulo 5 abrimos nuevas posibilidades de investigación que ampliarían y mejorarían este trabajo y damos a conocer las conclusiones que nos dejó el mismo.

Por último, proveemos el apéndice A sobre baculovirus y el apéndice B con información útil a la hora de la instalación y el uso del clasificador.

Capítulo 2

Máquinas de Vectores de Soporte

En este capítulo resumimos los conceptos básicos detrás de las máquinas de vectores de soporte, por considerarlo clave para el entendimiento del resto de la tesis.

2.1. Definición

Las máquinas de vectores de soporte (*Support Vector Machines*, que abreviaremos *SVMs*) [Vapnik (1998), Vapnik (2000)] son una familia de algoritmos que permiten clasificar elementos en dos clases distintas (aunque pueden ser extendidas para clasificar en un número arbitrario de clases). La especificación del algoritmo de construcción de una SVM es la siguiente:

Entrada: un conjunto de elementos $x_i \in R^n$ y sus clases conocidas $y_i \in$ {-1,+1}

$$S = \{ (x_1, y_1), (x_2, y_2) \ldots (x_k, y_k) \}$$

Salida: un clasificador capaz de predecir la clase de cualquier $x \in R^n$

$$f\colon R^n \to \{\text{-1, +1}\}$$

2.2. Funcionamiento

Las SVMs mapean el conjunto de vectores $S_x = \{ x_1, x_2, \ldots x_k \}$ recibido como entrada en un espacio de mayor dimensión (*feature space*), a través de una función no-lineal $\mu\colon R^n \to R^p$ $(n \leq p)$. En este espacio se construye un hiperplano que separa los x_i según su clase, es decir según el y_i al cual estén asociados. En caso de que una separación total de los elementos no sea posible, se busca aquella que minimice el error, es decir la cantidad de x_i que queden del lado equivocado.

De esta manera, el hiperespacio queda dividido en dos, uno que representa a la clase $+1$ (los positivos), y otro a la clase -1 (los negativos). Así, la clasificación de un elemento desconocido x, se realiza determinando a qué subespacio pertenece $\mu(x)$, es decir el vector resultante de mapear x en el *feature space*.

2.2.1. Feature space

¿Por qué es necesario realizar un mapeo a un *feature space* antes de calcular el hiperplano separador? ¿No se puede realizar el cálculo directamente en el espacio de entrada? Estas son algunas de las preguntas que nos surgen al intentar comprender el funcionamiento de las SVMs.

Muchas veces separar a los vectores de entrada por medio de un hiperplano no resulta conveniente. Es decir, no existe un hiperplano que pueda separar los x_i según su clase y_i con un error aceptable (figura 2.1).

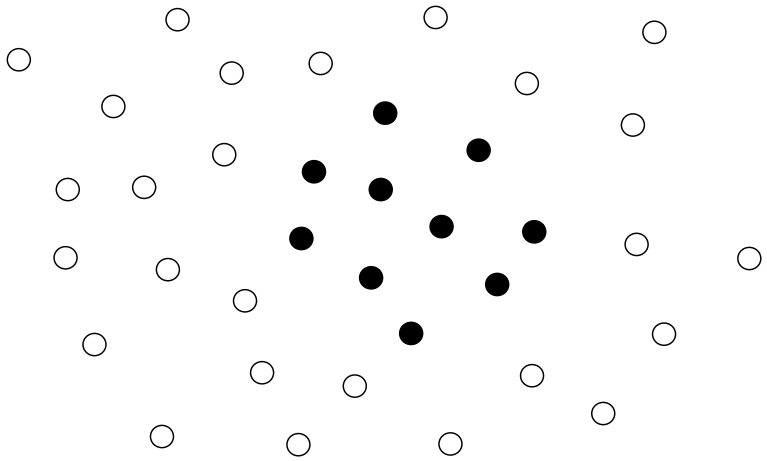

Figura 2.1: Ejemplo de vectores cuya separación por medio de un hiperplano no resulta conveniente.

En estos casos, es posible encontrar un mapeo no-lineal $\mu\colon R^n \to R^p$ $(n \leq p)$ para los vectores x_i en un espacio de mayor dimensión (*feature space*) en el que exista un hiperplano capaz de separarlos según su clase y_i con un error aceptable (figura 2.2).

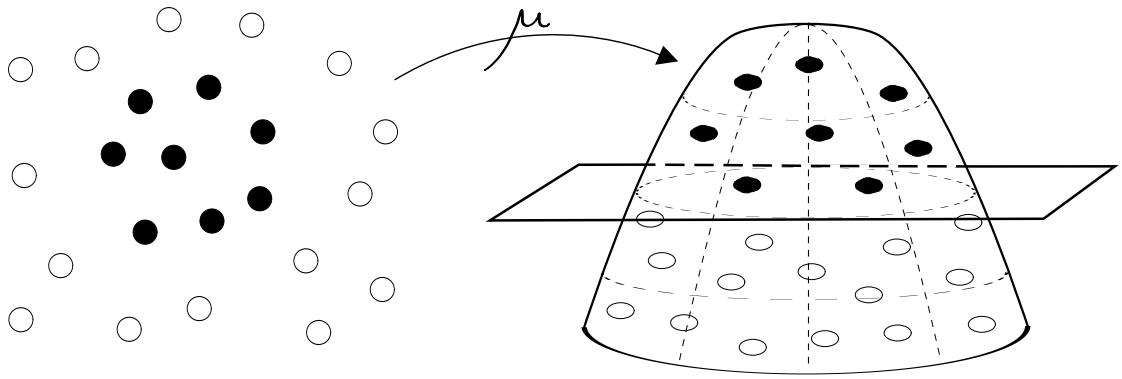

Figura 2.2: Ejemplo de mapeo no-lineal en un *feature space*.

2.2.2. Hiperplano separador óptimo *

En ocasiones existen muchos hiperplanos capaces de separar los x_i según su clase y_i (figura 2.3).

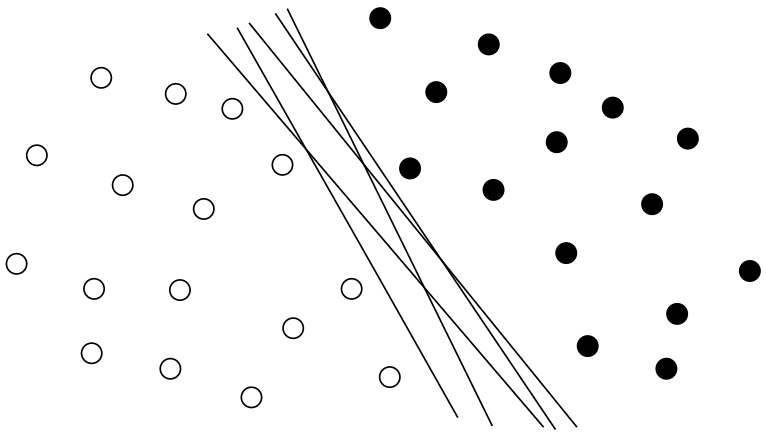

Figura 2.3: Posibles hiperplanos separadores.

El mejor hiperplano separador (figura 2.4) es aquel que separa los datos x_i (con respecto a su clase y_i) sin error y maximiza la distancia (*margen*) entre los vectores más cercanos al hiperplano. Estos vectores reciben el nombre de vectores de soporte (*support vectors*).

*En esta sección, por simplicidad, explicaremos cómo encontrar el mejor hiperplano separador para el caso en que los vectores pueden separarse completamente (es decir, sin error). Estos conceptos pueden extenderse fácilmente para aceptar un error Δ.

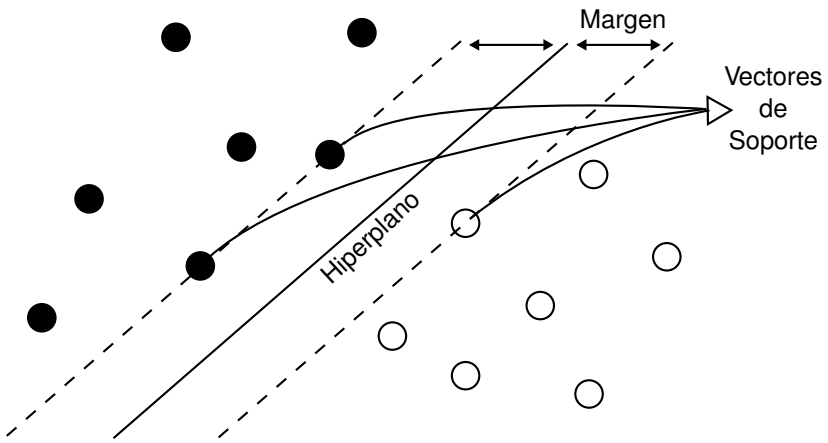

Figura 2.4: Mejor hiperplano separador.

Definimos el hiperplano por medio de la fórmula: $w \cdot x - b = 0$, con $w \in R^p$ y $b \in R$. El mismo debe cumplir las inecuaciones:

$$\begin{aligned} w \cdot x_i - b \geq 1 \text{ si } y_i = +1 \\ w \cdot x_i - b \leq -1 \text{ si } y_i = -1 \end{aligned} \tag{2.1}$$

donde $i \in \{1, 2, \ldots n\}$. Esto significa que los vectores positivos quedarán de un lado del hiperplano y los negativos del otro, separados por un doble margen (figura 2.5).

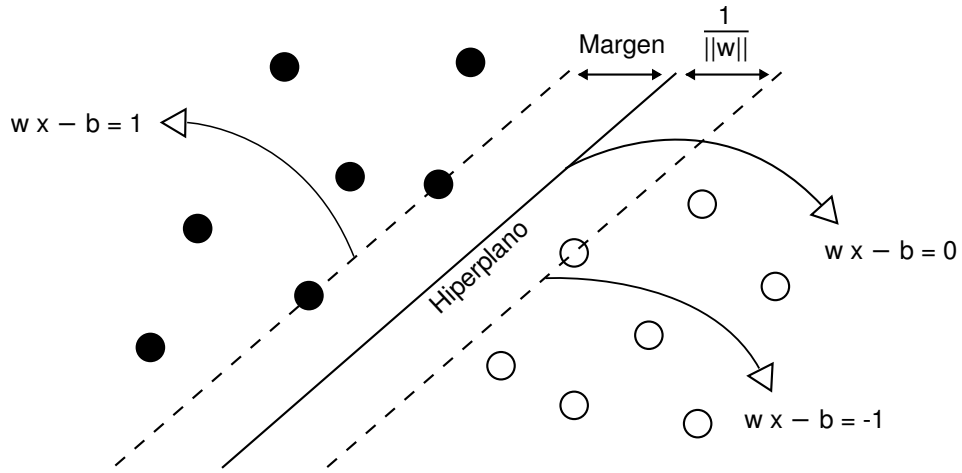

Figura 2.5: Ecuaciones del mejor hiperplano separador y de los dos hiperplanos paralelos que definen los márgenes de separación.

Además el hiperplano separador debe maximizar el margen, es decir la distancia a los vectores de soporte*: $d(w) = \frac{1}{\|w\|}$. Esto es equivalente a minimizar la función:

$$\Phi(w) = \frac{1}{2}\|w\|^2 = \frac{1}{2}(w \cdot w)$$

sujeta a las restricciones de (2.1). La solución de este problema de optimización es el punto silla de la función de Lagrange:

$$L(w, b, \alpha) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{n} \alpha_i (y_i(w \cdot x_i - b) - 1) \tag{2.2}$$

donde los α_i son los multiplicadores de Lagrange [Vapnik (2000)].

Una observación importante es que los únicos α_i que pueden ser distintos de cero en (2.2) son aquellos que acompañan a los x_i que cumplen: $y_i(w \cdot x_i - b) = 1$, es decir, los vectores de soporte. Esta propiedad nos garantiza que los únicos x_i necesarios para el cálculo del hiperplano separador son los correspondientes a vectores de soporte, que en la mayoría de los casos van a ser un pequeño porcentaje del total.

2.2.3. Funciones kernel

Cuando la dimensión del *feature space* Z es suficientemente grande (incluso puede ser infinita), realizar el mapeo de los vectores x_i en Z para luego calcular su producto interno se convierte en un problema intratable computacionalmente.

Una función kernel K permite calcular el producto interno entre dos vectores de Z implícitamente, es decir, sin tener que realizar el cálculo de las imágenes de los x_i en Z, o sea los $\mu(x_i)$.

Así, una función kernel cumple:

$$\forall x_1, x_2 \in X,\ K(x_1, x_2) = \mu(x_1) \cdot \mu(x_2)$$

* En particular los vectores de soporte x_i son los que cumplen: $y_i(w \cdot x_i - b) = 1$. La distancia con signo de un vector de soporte x_i al plano definido por la ecuación $w \cdot x - b = 0$ es $d = \frac{w \cdot x_i - b}{\|w\|}$. Entonces $y_i \cdot d = \frac{y_i(w \cdot x_i - b)}{\|w\|} = \frac{1}{\|w\|}$. Esto significa que todos los vectores de soporte se encuentran a distancia $\frac{\pm 1}{\|w\|}$ del hiperplano separador.

2.2.3.1. Ejemplos

Entre los kernels más ampliamente usados se encuentran los polinomiales y los gausianos. El kernel polinomial de grado uno (lineal) es el más sencillo de todos, los conjuntos de datos que son separados adecuadamente por un kernel de este tipo son los que tienen una distribución similar a los de la figura 2.6. Los kernels polinomiales de grado mayor o igual a dos son un poco más versátiles, pueden separar conjuntos de datos como los de la figura 2.7. Los kernels gausianos son mucho más complejos y permiten separar conjuntos de datos más sofisticados, como los de la figura 2.8. En la figura 2.9 mostramos un ejemplo de clasificación de un mismo conjunto de datos utilizando cuatro de los kernels mencionados anteriormente. En este ejemplo los que logran una mejor clasificación son los kernels gausianos, pero esto no es algo que vaya a ocurrir siempre, todo dependerá de la distribución de los datos que estemos clasificando.

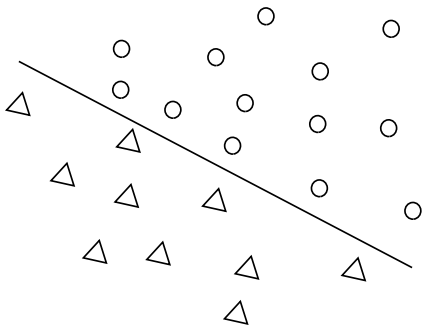

Figura 2.6: Conjunto de datos separables por un kernel lineal.

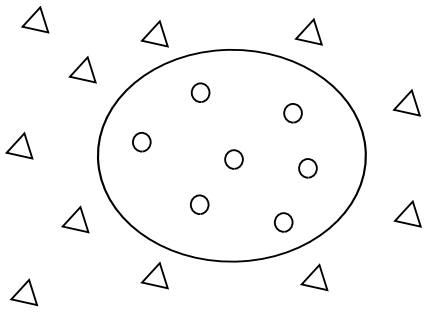

Figura 2.7: Conjunto de datos separables por un kernel polinomial de grado dos.

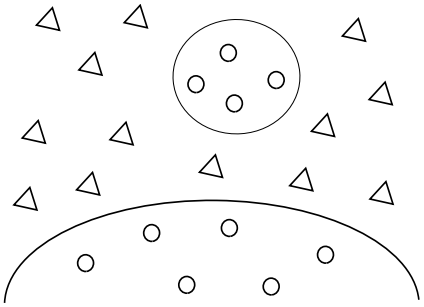

Figura 2.8: Conjunto de datos separables por un kernel gausiano.

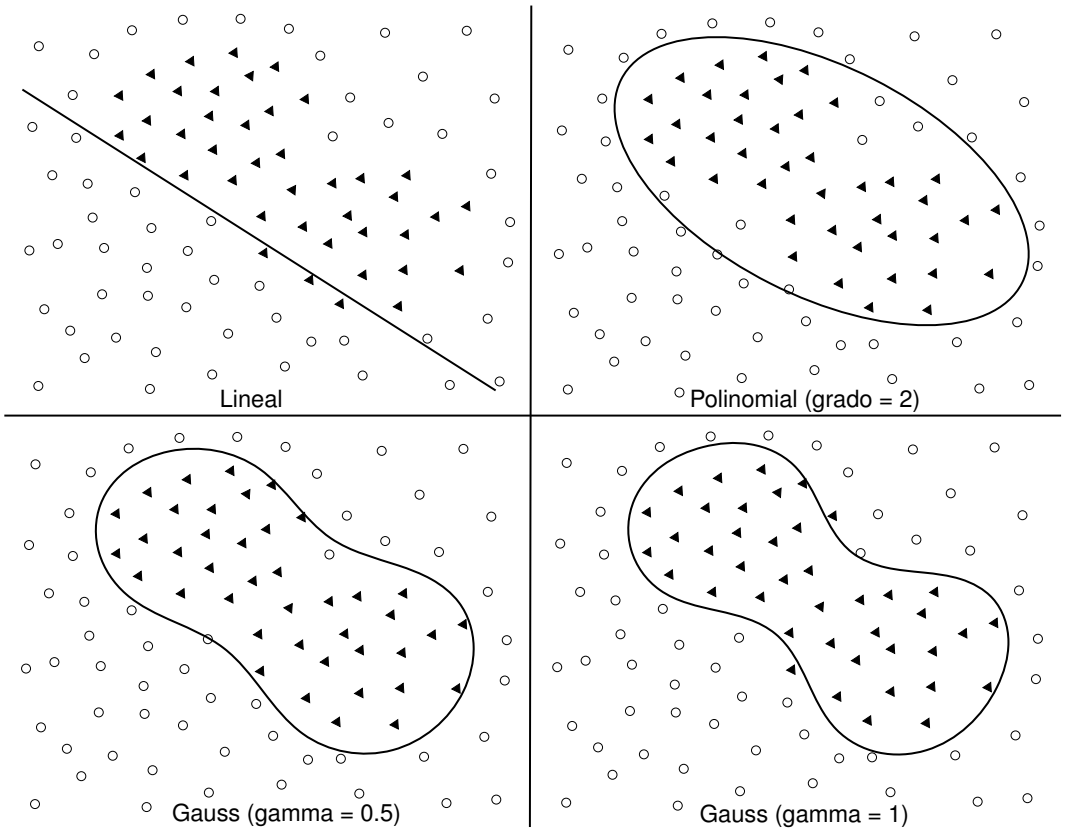

Figura 2.9: Comparación de cuatro kernels en la clasificación del mismo conjunto de datos.

2.3. F-measure

La medida de desempeño más utilizada en las tareas de clasificación es la llamada *F-measure*, que combina en un único valor dos estimadores: la precisión (*precision, P*) y la cobertura (*recall, R*). *F-measure* es la media armónica de P y R por lo que alcanza valores altos sólo cuando los valores de ambos son altos.

$$F\text{-}measure = \frac{2 \cdot P \cdot R}{P + R}$$

La precisión es la probabilidad de que un ítem que fue clasificado como positivo lo sea realmente, es decir:

$$P = Pr[y = 1 | f(x) = 1] = \frac{VP}{VP + FP}$$

donde *VP* significa *Verdaderos Positivos* (casos positivos que fueron clasificados como positivos), *FP* significa *Falsos Positivos* (casos negativos que fueron clasificados como positivos) y $f(x)$ es el resultado de la clasificación del elemento x.

La cobertura es la probabilidad de que un ítem positivo sea detectado como tal por el clasificador, es decir:

$$R = Pr[f(x) = 1 | y = 1] = \frac{VP}{VP + FN}$$

donde *FN* significa *Falsos Negativos* (casos positivos que fueron clasificados como negativos).

2.4. Validación cruzada

La validación cruzada (*cross-validation*) es una técnica que se utiliza para generalizar los resultados obtenidos en las clasificaciones. Consta de varias rondas en las que se particiona el conjunto de datos en subconjuntos disjuntos para utilizarlos en el entrenamiento y testeo del clasificador. En cada ronda se varían los subconjuntos de datos utilizados, y de esta manera los que en una ronda fueron utilizados para el entrenamiento del clasificador, en la siguiente pueden ser usados para el testeo, y viceversa. De aquí proviene el nombre de *validación cruzada*. Una vez concluidas todas las rondas, los resultados son promediados. Este resultado se asemeja más al real que el obtenido en una única clasificación, ya que deja de lado casos patológicos.

Un tipo particular de validación cruzada es la que recibe el nombre de *K-fold cross validation*. En ella el conjunto de datos se particiona aleatoriamente en K subconjuntos disjuntos. Se realizan K rondas de validación. En cada ronda se entrena el clasificador con $K-1$ subconjuntos y se lo testea con el restante. En cada ronda se utiliza un subconjunto distinto para el testeo, cubriendo así las K combinaciones posibles. Finalmente se calcula el promedio de los resultados obtenidos (figura 2.10).

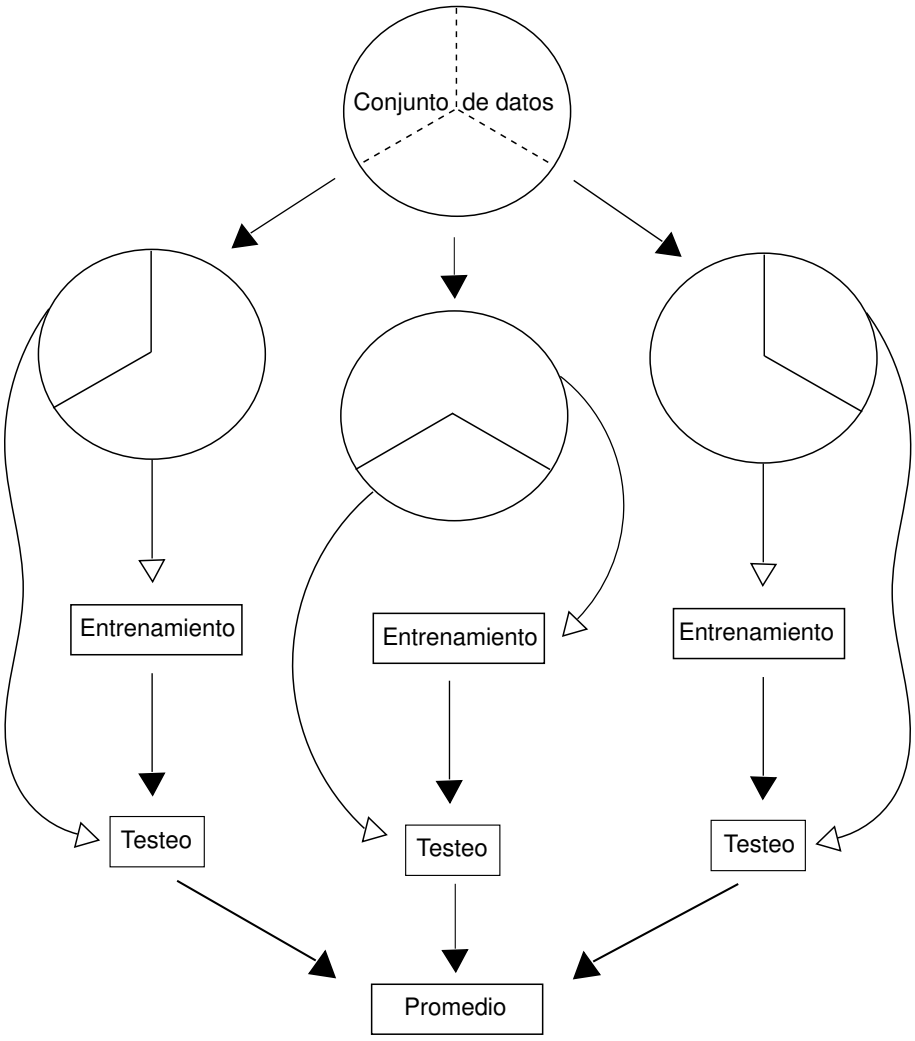

Figura 2.10: *K-fold cross validation* con $k = 3$.

2.5. PyML: Una implementación de SVM

PyML* es un *framework* de código abierto escrito en Python†. Fue creado por Asa Ben-Hur, un profesor del departamento de ciencias de la computación de la Universidad de Colorado (USA).

PyML provee una implementación de *Support Vector Machine*, y de otros métodos de clasificación. También cuenta con extensiones de estos algoritmos para clasificar conjuntos de más de dos clases (métodos *multi-class*). Para los experimentos de esta tesis utilizamos PyML versión 0.7.2.

*http://pyml.sourceforge.net
†http://www.python.org

Capítulo 3

Construcción del clasificador

En este capítulo explicamos el proceso de creación del clasificador de polipéptidos multi-criterio. Primero presentamos los datos que utilizaremos y planteamos una hipótesis sobre los mismos. Luego resumimos los criterios que decidimos usar para la generación de los vectores. Después enumeramos algunos detalles específicos de la clasificación relevantes a nuestro problema. Por último, presentamos los experimentos realizados y proveemos un análisis de sus resultados.

3.1. Datos utilizados

Para los experimentos que detallaremos en esta sección empleamos proteomas obtenidos de GenBank*. El contenido de cada uno de ellos fue dividido en dos archivos. Uno contiene los polipéptidos conocidos, definidos como tales en base a tener registrada una función, es decir que se los ha encontrado y se les ha asignado una función en algún baculovirus. El otro contiene aquellos cuya existencia aún se desconoce, no tienen comprobación experimental hasta el momento, ni asignación de función teórica. Para cada uno de los genomas también obtuvimos el proteoma teórico completo, eliminamos todos los polipéptidos anotados (conocidos y desconocidos) y construimos un tercer archivo que contiene los polipéptidos que no fueron anotados porque los autores consideraron muy poco probable su existencia. En el caso de estos últimos utilizamos sólo aquellos cuyo tamaño iguala o supera los 45 aminoácidos, ya que teniendo en cuenta lo descripto hasta la fecha para baculovirus, no se han encontrado proteínas funcionales de menor tamaño.

Cada uno de los tres conjuntos de polipétidos mencionados anteriormente (a los que denominaremos de ahora en más polipéptidos conocidos, desconocidos y no anotados - ver figura 3.1) será dividido en subconjuntos para entrenar y testear

*http://www.ncbi.nlm.nih.gov/genbank

la SVM. Los mismos serán completamente disjuntos.

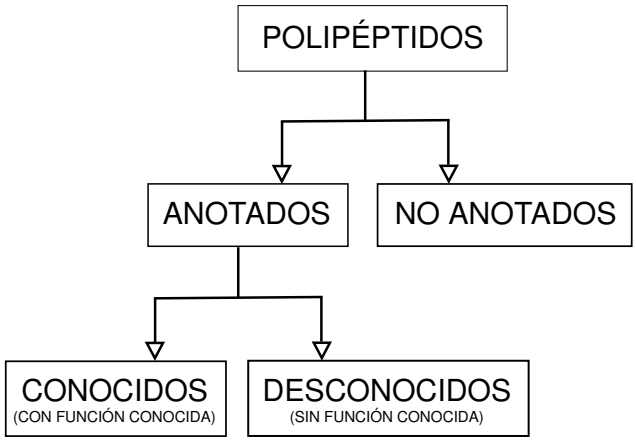

Figura 3.1: Clasificación de polipéptidos.

3.1.1. Hipótesis sobre el dominio

Si bien podemos afirmar que el conjunto de polipéptidos conocidos contiene en su totalidad casos positivos reales, no sucede lo mismo con los otros conjuntos de datos. La hipótesis que aquí planteamos es que los polipéptidos no anotados contienen un mayor número de casos negativos que los polipéptidos desconocidos, ya que los primeros fueron excluídos de la anotación por sus autores, mientras que los segundos no.

3.2. Criterios de clasificación

Para la clasificación de las cadenas nos basaremos en criterios composicionales. Esto incluye tanto la frecuencia de aparición como la ubicación relativa de los aminoácidos. Además emplearemos conceptos de teoría de la información para evaluar aspectos como la entropía de las cadenas.

A continuación enumeramos los criterios utilizados en los experimentos y proporcionamos la idea global de cada uno.

- **Frecuencia de aminoácidos:** cantidad de apariciones de cada aminoácido sobre el total de aminoácidos de la cadena.

- **Frecuencia de di-aminoácidos:** cantidad de apariciones de cada par de aminoácidos consecutivos sobre el total de posibles apariciones.

  El tipo y frecuencia de los aminoácidos presentes en las proteínas ha ido cambiando durante la evolución [Brooks *et al.* (2002)]. Durante el proceso evolutivo no sólo han ocurrido cambios en función de la aparición de nuevos aminoácidos sino también en función de procesos selectivos asociados a funciones estructurales, enzimáticas, o ambas. De esta manera, podemos afirmar que las frecuencias aminoacídicas son particulares de cada proteína [Urbina *et al.* (2006)] estando vinculadas a la esencia de la misma. Por otro lado, la frecuencia de los distintos di-aminoácidos presentes en un polipéptido tiene que ver con la existencia de asociaciones de a dos necesarias para una determinada estructura y/o función biológica.

- **Entropía de Shannon:** entropía de la cadena de aminoácidos utilizando la definición de Shannon [Shannon (1948)].

- **Entropía Zip:** entropía de la cadena de aminoácidos definida como el cociente entre el tamaño de la cadena comprimida en formato Zip y el tamaño original de la misma.

  La entropía es una medida de las redundancias informativas de los polipéptidos. A mayor redundancia, mayor entropía. En la naturaleza hay proteínas de todo tipo, algunas con muy baja redundancia y otras con muy alta redundancia. La distribución de dichas regiones de baja o alta entropía puede ser localizada o generalizada, esto puede estar relacionado con cuestiones estructurales [Liao *et al.* (2005)], o funcionales [Du *et al.* (2009)].

- **Frecuencia de pares de aminoácidos separados por N aminoácidos:** cantidad de apariciones de cada par de aminoácidos separados por cualquier secuencia de N aminoácidos sobre el total de posibles ocurrencias. Tomamos $1 \leq N \leq 5$.

  La medida de aminoácidos separados por un espaciador de longitud variable (pero corta) es una manera indirecta de evaluar periodicidad. La ventaja de utilizar este método es que se la detecta aún cuando sea un fenómeno local de un segmento pequeño. Un ejemplo típico de este tipo de motivos es el cierre de leucinas (*leucine zipper*) [Landschulz *et al.* (1988)].

- **Periodicidad de aminoácidos:** periodicidad de cada aminoácido medida mediante el método de Cornette [Cornette *et al.* (1987)].

  Hay ciertas estructuras presentes en las proteínas como las hélices alfa o las hojas plegadas beta que se caracterizan por una cierta periodicidad en la composición de las cadenas aminoacídicas que adoptan dicha estructura.

> Pero el concepto de periodicidad es mucho más general [Ivanov & Ivanov (1980)] y, al igual que las frecuencias de aminoácidos, está relacionado con la evolución de los polipéptidos [Gatherer & McEwan (2003)]. Por ello, en nuestro caso no intentaremos detectar ninguna estructura en particular, sino cualquier patrón de periodicidad de los aminoácidos dentro del polipéptido.

Evaluaremos el desempeño de cada uno de estos criterios utilizando diversos kernels para obtener las combinaciones que logren una mejor clasificación de los polipéptidos. Posteriormente definiremos una estrategia capaz de combinarlos para construir un clasificador de polipéptidos multi-criterio de alto desempeño.

3.3. Detalles específicos

En esta sección presentamos los kernels elegidos para la clasificación, las técnicas que utilizaremos para obtener resultados más representativos en los experimentos que así lo requieran y la medida de desempeño que emplearemos para comparar las distintas clasificaciones.

3.3.1. Kernels elegidos

Como no conocemos las distribuciones espaciales que adoptarán los vectores generados en base a los criterios que definimos en la sección 3.2, en los experimentos de este trabajo probaremos el desempeño de siete kernels diferentes: uno lineal, uno polinomial de grado dos y cinco gausianos, con $\gamma \in \{1, 0.5, 0.1, 0.05, 0.01\}$*.

3.3.2. Técnicas utilizadas

Para que los resultados sean más representativos, en los experimentos 2, 3, 4 y 5 utilizaremos la técnica de *K-fold cross-validation* que explicamos en la sección 2.4. En particular, tomamos $K = 7$ por considerarlo suficiente para obtener una generalización adecuada de los resultados.

3.3.3. Medida de desempeño del clasificador

En la sección 2.3 describimos las medidas de desempeño que se utilizan normalmente en clasificación. Las mismas funcionan adecuadamente cuando se cuenta con casos positivos y casos negativos. Cuando no se cuenta con casos negativos reales, es muy difícil estimar la cantidad de falsos positivos (FP). Por ello, en

*Para mayor información sobre los distintos kernels puede recurrirse a la sección 2.2.3.1

este trabajo utilizaremos una medida similar propuesta por Wee Sun Lee y Bing Liu en [Lee & Liu (2003)], a la que llamaremos *F'-measure* y se define como:

$$F\text{'-}measure = \frac{R^2}{Pr[f(x) = 1]}$$

Para que *F'-measure* esté definida para todos los valores posibles de R y de $Pr[f(x) = 1]$, la extendemos de la siguiente manera:

$$F\text{'-}measure = \begin{cases} \frac{R^2}{Pr[f(x)=1]} & \text{si } Pr[f(x) = 1] \neq 0 \\ 0 & \text{en caso contrario} \end{cases}$$

En [Lee & Liu (2003)] se demuestra la siguiente equivalencia:

$$\frac{R^2}{Pr[f(x) = 1]} = \frac{P \cdot R}{Pr[y = 1]}$$

F'-measure resulta entonces ser proporcional al cuadrado de la media geométrica de P y R, por lo cual se comporta de manera similar a *F-measure* (sólo toma valores altos cuanto los valores de P y R son ambos altos).

Cabe aclarar que, a diferencia de *F-measure* que alcanza su valor máximo en 1, *F'-measure* alcanza su máximo en 2.

3.4. Experimento 1: Selección de datos de entrenamiento

El objetivo de este experimento es definir que datos utilizar para el entrenamiento de las distintas intancias de SVM que construiremos a lo largo de este trabajo. Por un lado, sabemos que los casos positivos están representados por los polipéptidos conocidos, ya que se ha determinado mediante experimentación en mesada húmeda su existencia real. Por otro lado, deberíamos definir qué casos negativos utilizar para el entrenamiento.

Dada la imposibilidad de contar con casos negativos reales (ver sección 1.1) deberemos decidir qué subconjunto de datos utilizar. Disponemos de dos tipos de datos: polipéptidos desconocidos y no anotados. Los primeros son aquellos que si bien fueron anotados en su correspondiente proteoma, aún no se ha demostrado su presencia en baculovirus. Los segundos son aquellos que los autores excluyeron de la anotación por considerar poco probable su existencia en baculovirus. En este experimento probaremos utilizar estos dos conjuntos de datos, así como también su unión, para ver cual es el desempeño de la SVM en cada caso.

Por otro lado, el mismo problema que existe para el entrenamiento, también existe para el testeo de la SVM. Es decir, no contamos con casos negativos reales para medir el desempeño de la misma, por lo cual utilizaremos los mismos subconjuntos que para el entrenamiento: polipéptidos desconocidos y no anotados. La hipótesis inicial (ver sección 3.1.1) es que los polipéptidos no anotados contienen un mayor número de casos negativos que los polipéptidos desconocidos. Este experimento nos permitirá comenzar a analizar la veracidad de la misma.

Para realizar el experimento, cada uno de los tres conjuntos de polipéptidos mencionados en la sección 3.1 fue dividido en dos subconjuntos, uno de los cuales será utilizado para entrenar la SVM (conteniendo 44 proteomas) y el otro para testearla (conteniendo 5 proteomas).

Utilizamos cuatro de los kernels mencionados en la sección 3.3.1; uno lineal, uno polinomial de grado dos y dos gausianos (con $\gamma = 1$ y $\gamma = 0.5$) y dos de los criterios detallados en la sección 3.2; frecuencia de aminoácidos y frecuencia de di-aminoácidos. Consideramos que esta muestra de kernels y criterios es suficiente para testear el desempeño de la SVM usando los tres conjuntos de datos de entrenamiento mencionados anteriormente.

Dividimos este experimento en tres subexperimentos. En cada uno de ellos utilizamos diferentes conjuntos de datos para el entrenamiento de la SVM. A su vez, cada experimento fue dividido en dos subexperimentos, según el conjunto de datos que se haya utilizado para el testeo de la SVM. Si bien los explicamos en detalle a continuación, los hemos resumido en el cuadro 3.1 para mayor claridad.

Cuadro 3.1: Datos de entrenamiento y testeo utilizados en los distintos subexperimentos del experimento 1: Selección de datos de entrenamiento.

Sub-experimento	Datos utilizados			
	Entrenamiento		Testeo	
	+	-	+	-
1.1.1	conocidos	desconocidos	conocidos	desconocidos
1.1.2	conocidos	desconocidos	conocidos	no anotados
1.2.1	conocidos	no anotados	conocidos	desconocidos
1.2.2	conocidos	no anotados	conocidos	no anotados
1.3.1	conocidos	desconocidos $\cup$ no anotados	conocidos	desconocidos
1.3.2	conocidos	desconocidos $\cup$ no anotados	conocidos	no anotados

En el experimento 1.1 utilizamos como casos de entrenamiento positivo polipéptidos conocidos y como casos de entrenamiento negativo polipéptidos desconocidos. Dividimos este experimento en dos subexperimentos. En el experimento 1.1.1 (cuadro 3.2) testeamos el clasificador mediante polipéptidos conocidos y desconocidos mientras que en el experimento 1.1.2 (cuadro 3.3) lo testeamos mediante polipéptidos conocidos y no anotados.

Cuadro 3.2: Resultados del experimento 1.1.1: Selección de datos de entrenamiento. Entrenamiento y testeo con polipéptidos conocidos y desconocidos.

Criterio de clasificación	Kernel	Resultados obtenidos				
		Conocidos		Desconocidos		F'-measure
		+	-	+	-	
Frecuencia de aminoácidos	Lineal	0.66	0.34	0.37	0.63	0.85
	Polinomial (grado = 2)	0.66	0.34	0.35	0.65	0.86
	Gauss ($\gamma = 1$)	1.00	0.00	0.95	0.05	1.03
	Gauss ($\gamma = 0.5$)	0.93	0.07	0.69	0.31	1.07
Frecuencia de di-aminoácidos	Lineal	0.78	0.22	0.42	0.58	1.00
	Polinomial (grado = 2)	0.73	0.27	0.34	0.66	1.00
	Gauss ($\gamma = 1$)	0.93	0.07	0.56	0.44	1.15
	Gauss ($\gamma = 0.5$)	0.88	0.12	0.48	0.52	1.14

Cuadro 3.3: Resultados del experimento 1.1.2: Selección de datos de entrenamiento. Entrenamiento con polipéptidos conocidos y desconocidos. Testeo con conocidos y no anotados.

Criterio de clasificación	Kernel	Resultados obtenidos				
		Conocidos		No Anotados		F'-measure
		+	-	+	-	
Frecuencia de aminoácidos	Lineal	0.66	0.34	0.44	0.56	0.80
	Polinomial (grado = 2)	0.66	0.34	0.44	0.56	0.80
	Gauss ($\gamma = 1$)	1.00	0.00	0.89	0.11	1.06
	Gauss ($\gamma = 0.5$)	0.93	0.07	0.64	0.36	1.10
Frecuencia de di-aminoácidos	Lineal	0.78	0.22	0.43	0.57	1.00
	Polinomial (grado = 2)	0.73	0.27	0.38	0.62	0.96
	Gauss ($\gamma = 1$)	0.93	0.07	0.58	0.42	1.14
	Gauss ($\gamma = 0.5$)	0.88	0.12	0.52	0.48	1.10

En el experimento 1.2 utilizamos como casos de entrenamiento positivo polipéptidos conocidos y como casos de entrenamiento negativo polipéptidos no anotados. Dividimos este experimento en dos subexperimentos. En el experimento 1.2.1 (cuadro 3.4) testeamos el clasificador mediante polipéptidos conocidos y desconocidos mientras que en el experimento 1.2.2 (cuadro 3.5) lo testeamos mediante polipéptidos conocidos y no anotados.

Cuadro 3.4: Resultados del experimento 1.2.1: Selección de datos de entrenamiento. Entrenamiento con polipéptidos conocidos y no anotados. Testeo con conocidos y desconocidos.

Criterio de clasificación	Kernel	Resultados obtenidos				
		Conocidos		Desconocidos		F'-measure
		+	-	+	-	
Frecuencia de aminoácidos	Lineal	0.92	0.08	0.87	0.13	0.95
	Polinomial (grado = 2)	0.94	0.06	0.88	0.13	0.97
	Gauss ($\gamma = 1$)	1.00	0.00	0.98	0.02	1.01
	Gauss ($\gamma = 0.5$)	1.00	0.00	0.94	0.06	1.03
Frecuencia de di-aminoácidos	Lineal	0.96	0.04	0.89	0.11	1.00
	Polinomial (grado = 2)	0.97	0.03	0.87	0.13	1.03
	Gauss ($\gamma = 1$)	1.00	0.00	0.95	0.05	1.02
	Gauss ($\gamma = 0.5$)	1.00	0.00	0.95	0.05	1.02

Cuadro 3.5: Resultados del experimento 1.2.2: Selección de datos de entrenamiento. Entrenamiento y testeo con polipéptidos conocidos y no anotados.

Criterio de clasificación	Kernel	Resultados obtenidos				
		Conocidos		No Anotados		F'-measure
		+	-	+	-	
Frecuencia de aminoácidos	Lineal	0.92	0.08	0.07	0.93	1.72
	Polinomial (grado = 2)	0.94	0.06	0.06	0.94	1.77
	Gauss ($\gamma = 1$)	1.00	0.00	0.44	0.56	1.38
	Gauss ($\gamma = 0.5$)	1.00	0.00	0.19	0.81	1.67
Frecuencia de di-aminoácidos	Lineal	0.96	0.04	0.07	0.93	1.79
	Polinomial (grado = 2)	0.97	0.03	0.06	0.94	1.82
	Gauss ($\gamma = 1$)	1.00	0.00	0.22	0.78	1.64
	Gauss ($\gamma = 0.5$)	0.98	0.02	0.13	0.87	1.74

En el experimento 1.3 utilizamos como casos de entrenamiento positivo polipéptidos conocidos y como casos de entrenamiento negativo la unión de polipéptidos desconocidos y no anotados. Dividimos este experimento en dos subexperimentos. En el experimento 1.3.1 (cuadro 3.6) testeamos el clasificador mediante polipéptidos conocidos y desconocidos mientras que en el experimento 1.3.2 (cuadro 3.7) lo testeamos mediante polipéptidos conocidos y no anotados.

Cuadro 3.6: Resultados del experimento 1.3.1: Selección de datos de entrenamiento. Entrenamiento con polipéptidos conocidos y desconocidos $\cup$ no anotados. Testeo con conocidos y desconocidos.

Criterio de clasificación	Kernel	Resultados obtenidos				
		Conocidos		Desconocidos		F'-measure
		+	-	+	-	
Frecuencia de aminoácidos	Lineal	0.92	0.08	0.84	0.16	0.96
	Polinomial (grado = 2)	0.94	0.06	0.84	0.16	0.99
	Gauss ($\gamma = 1$)	1.00	0.00	0.99	0.01	1.00
	Gauss ($\gamma = 0.5$)	1.00	0.00	0.95	0.05	1.02
Frecuencia de di-aminoácidos	Lineal	0.95	0.05	0.85	0.15	1.00
	Polinomial (grado = 2)	0.96	0.04	0.83	0.17	1.03
	Gauss ($\gamma = 1$)	1.00	0.00	0.93	0.07	1.03
	Gauss ($\gamma = 0.5$)	0.98	0.02	0.90	0.10	1.03

Cuadro 3.7: Resultados del experimento 1.3.2: Selección de datos de entrenamiento. Entrenamiento con polipéptidos conocidos y desconocidos ∪ no anotados. Testeo con conocidos y no anotados.

Criterio de clasificación	Kernel	Resultados obtenidos				
		Conocidos		No Anotados		F'-measure
		+	-	+	-	
Frecuencia de aminoácidos	Lineal	0.92	0.08	0.07	0.93	1.71
	Polinomial (grado = 2)	0.94	0.06	0.05	0.95	1.77
	Gauss ($\gamma = 1$)	1.00	0.00	0.88	0.12	1.06
	Gauss ($\gamma = 0.5$)	1.00	0.00	0.35	0.65	1.48
Frecuencia de di-aminoácidos	Lineal	0.95	0.05	0.07	0.93	1.76
	Polinomial (grado = 2)	0.96	0.04	0.06	0.94	1.80
	Gauss ($\gamma = 1$)	1.00	0.00	0.27	0.73	1.57
	Gauss ($\gamma = 0.5$)	0.98	0.02	0.14	0.86	1.72

3.4.1. Análisis de los resultados

Luego de analizar con detenimiento los resultados del experimento 1 podemos concluir que los mejores resultados que obtuvimos en la clasificación de polipéptidos desconocidos fue utilizando para el entrenamiento polipéptidos conocidos y desconocidos (ver cuadro 3.2). Por otro lado, en el caso de polipéptidos no anotados obtuvimos mejores resultados utilizando para el entrenamiento polipéptidos conocidos y no anotados (ver cuadro 3.5). Además, en el experimento 1.3 puede observarse que la unión de polipéptidos desconocidos y no anotados para conformar un conjunto de datos negativos mixto para el entrenamiento de la SVM no mejora los resultados obtenidos en la clasificación de polipéptidos desconocidos ni tampoco de polipéptidos no anotados (ver cuadros 3.6 y 3.7).

Las tres observaciones anteriores sugieren que cuanto más se asemejan los conjuntos de datos de entrenamiento y los de testeo, mejores son los resultados que se obtienen (ver figuras 3.2 y 3.3). Por este motivo, en los siguientes experimentos utilizaremos como casos de entrenamiento negativo polipéptidos desconocidos para la clasificación de polipéptidos desconocidos y polipéptidos no anotados para la clasificación de polipéptidos no anotados.

Los experimentos 1.1 y 1.2 sugieren además que los polipéptidos conocidos se diferencian más de los no anotados que de los desconocidos (en relación a los criterios elegidos para este primer experimento), permitiendo de esta manera que la clasificación de mejores resultados. Esto es coherente con la hipótesis inicial

del dominio planteada en la sección 3.1.1.

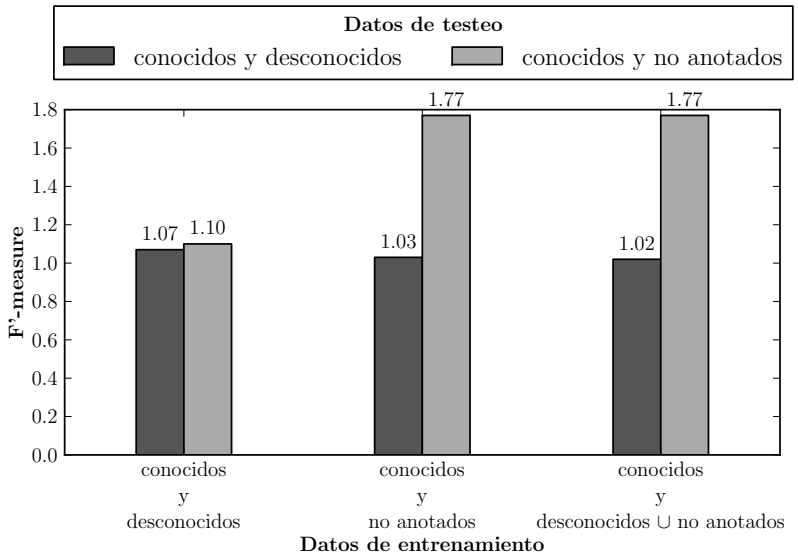

Figura 3.2: Cuadro comparativo de los resultados obtenidos para los distintos conjuntos de datos de entrenamiento. Criterio: Frecuencia de aminoácidos.

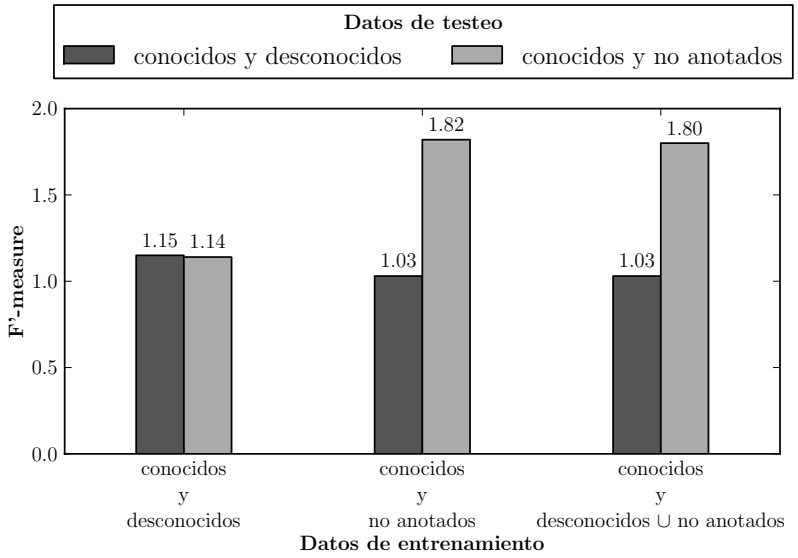

Figura 3.3: Cuadro comparativo de los resultados obtenidos para los distintos conjuntos de datos de entrenamiento. Criterio: Frecuencia de di-aminoácidos.

3.5. Experimento 2: Frecuencia de aminoácidos y di-aminoácidos

En este experimento elegimos como criterios de clasificación la frecuencia de aparición de aminoácidos y di-aminoácidos (es decir, pares de aminoácidos) ubicados de manera adyacente en la cadena.

Utilizaremos la técnica de *K-fold cross validation* descripta en la sección 3.3.2 y los siete kernels mencionados en la sección 3.3.1 (siendo éstas dos diferencias fundamentales con el experimento 1).

Dividimos el experimento 2 en dos subexperimentos. En el experimento 2.1 (cuadro 3.8) entrenamos y testeamos el clasificador mediante polipéptidos conocidos y desconocidos mientras que en el experimento 2.2 (cuadro 3.9) lo entrenamos y testeamos mediante polipéptidos conocidos y no anotados.

Cuadro 3.8: Resultados del experimento 2.1: Frecuencia de aminoácidos y di-aminoácidos. Entrenamiento y testeo con polipéptidos conocidos y desconocidos.

Criterio de clasificación	Kernel	Resultados obtenidos				
		Conocidos		Desconocidos		F'-measure
		+	-	+	-	
Frecuencia de aminoácidos	Lineal	0.66	0.34	0.38	0.62	0.84
	Polinomial (grado = 2)	0.67	0.33	0.38	0.62	0.86
	Gauss ($\gamma = 1$)	1.00	0.00	0.95	0.05	1.02
	Gauss ($\gamma = 0.5$)	0.92	0.08	0.70	0.30	1.05
	Gauss ($\gamma = 0.1$)	0.68	0.32	0.41	0.59	0.85
	Gauss ($\gamma = 0.05$)	0.66	0.34	0.40	0.60	0.82
	Gauss ($\gamma = 0.01$)	0.11	0.89	0.06	0.94	0.14
Frecuencia de di-aminoácidos	Lineal	0.74	0.26	0.42	0.58	0.95
	Polinomial (grado = 2)	0.73	0.27	0.37	0.63	0.98
	Gauss ($\gamma = 1$)	0.91	0.09	0.60	0.40	1.09
	Gauss ($\gamma = 0.5$)	0.84	0.16	0.52	0.48	1.04
	Gauss ($\gamma = 0.1$)	0.39	0.61	0.18	0.82	0.53
	Gauss ($\gamma = 0.05$)	0.19	0.81	0.09	0.91	0.25
	Gauss ($\gamma = 0.01$)	0.43	0.57	0.43	0.57	0.43

Cuadro 3.9: Resultados del experimento 2.2: Frecuencia de aminoácidos y di-aminoácidos. Entrenamiento y testeo con polipéptidos conocidos y no anotados.

Criterio de clasificación	Kernel	Resultados obtenidos				
		Conocidos		No anotados		F'-measure
		+	-	+	-	
Frecuencia de aminoácidos	Lineal	0.94	0.06	0.09	0.91	1.71
	Polinomial (grado = 2)	0.95	0.05	0.08	0.92	1.75
	Gauss ($\gamma = 1$)	1.00	0.00	0.45	0.55	1.38
	Gauss ($\gamma = 0.5$)	0.99	0.01	0.21	0.79	1.62
	Gauss ($\gamma = 0.1$)	0.95	0.05	0.11	0.89	1.70
	Gauss ($\gamma = 0.05$)	0.95	0.05	0.11	0.89	1.70
	Gauss ($\gamma = 0.01$)	0.95	0.05	0.13	0.87	1.67
Frecuencia de di-aminoácidos	Lineal	0.97	0.03	0.09	0.91	1.78
	Polinomial (grado = 2)	0.97	0.03	0.08	0.92	1.80
	Gauss ($\gamma = 1$)	0.99	0.01	0.24	0.76	1.61
	Gauss ($\gamma = 0.5$)	0.99	0.01	0.15	0.85	1.72
	Gauss ($\gamma = 0.1$)	0.96	0.04	0.13	0.87	1.70
	Gauss ($\gamma = 0.05$)	0.96	0.04	0.17	0.83	1.64
	Gauss ($\gamma = 0.01$)	0.92	0.08	0.26	0.74	1.44

3.5.1. Análisis de los resultados

Del análisis del experimento 2 podemos concluir que los mejores resultados que obtuvimos en la clasificación de polipéptidos desconocidos para ambos criterios fue utilizando kernels gausianos con $\gamma = 0.5$ para las frecuencias de aminoácidos y con $\gamma = 1$ para las frecuencias de di-aminoácidos (ver cuadro 3.8), mientras que en el caso de polipéptidos no anotados obtuvimos mejores resultados para ambos criterios utilizando kernels polinomiales de grado dos (ver cuadro 3.9).

Tanto para la clasificación de polipéptidos desconocidos como para la de no anotados obtuvimos mejores resultados utilizando como criterio la frecuencia de los di-aminoácidos que la de los aminoácidos. No obstante, ambos criterios parecen ser adecuados para la clasificación de polipéptidos.

Una última observación a realizar sobre los resultados de este experimento es que, al igual que en el experimento 1, los polipéptidos conocidos se diferencian más de los no anotados que de los desconocidos (en relación a los criterios de clasificación elegidos en este experimento). Esto nos permite reafirmar la hipótesis inicial del dominio planteada en la sección 3.1.1.

3.6. Experimento 3: Entropía de las cadenas

En este experimento buscaremos clasificar a los polipéptidos según su entropía. Mediremos la entropía de las cadenas de dos maneras distintas. La primera, y más clásica es basándonos en la definición de entropía provista por Claude E. Shannon en [Shannon (1948)]. La entropía de Shannon H^* de una variable discreta X cuyos posibles valores son $\Delta = \{x_1, x_2 \dots x_n\}$ se define como:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b(p(x_i))$$

donde b es la base del logaritmo usado.

Cuando $p(x_i) = 0$ se define $p(x_i) \log_b(p(x_i)) = 0$, lo que es consistente con el siguiente límite:

$$\lim_{p \to 0+} p \log_b(p) = 0$$

En particular, en el contexto del experimento, utilizaremos $b = 2$ y X tomará valores del conjunto de los posibles aminoácidos denotados por su código de una letra $\Delta = \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$.

La segunda forma de medición de la entropía que utilizaremos es el nivel de compresión de la cadena. Esto lo estimaremos como el cociente entre el tamaño de la cadena comprimida sobre el tamaño de la original. Existe una inmensa gama de formatos de compresión. En este experimento utilizaremos uno de los más conocidos, Zip.

Utilizaremos la técnica de *K-fold cross validation* descripta en la sección 3.3.2 y los siete kernels mencionados en la sección 3.3.1.

Dividimos el experimento 3 en dos subexperimentos: 3.1 y 3.2. En el experimento 3.1 (cuadro 3.10) entrenamos y testeamos el clasificador mediante polipéptidos conocidos y desconocidos mientras que en el experimento 3.2 (cuadro

* En esta tesis respetamos la simbología original propuesta por Shannon. Sin embargo, es de notar que en biología el símbolo H tiene un significado muy diferente. Para evitar las confusiones, desde el comienzo del empleo de la Teoría de la Información de Shannon en biología, el símbolo H fue cambiado por y [Schneider & Stephens (1990)]

3.11) lo entrenamos y testeamos mediante polipéptidos conocidos y no anotados.

Cuadro 3.10: Resultados del experimento 3.1: Entropía de las cadenas. Entrenamiento y testeo con polipéptidos conocidos y desconocidos.

Criterio de clasificación	Kernel	Resultados obtenidos				
		Conocidos		Desconocidos		F'-measure
		+	-	+	-	
Entropía de la cadena de aminoácidos (Shannon)	Lineal	0.96	0.04	0.93	0.07	0.97
	Polinomial (grado = 2)	0.92	0.08	0.81	0.19	0.97
	Gauss ($\gamma = 1$)	0.85	0.15	0.66	0.34	0.95
	Gauss ($\gamma = 0.5$)	0.95	0.05	0.90	0.10	0.97
	Gauss ($\gamma = 0.1$)	0.99	0.01	1.00	0.00	0.98
	Gauss ($\gamma = 0.05$)	0.88	0.12	0.87	0.13	0.88
	Gauss ($\gamma = 0.01$)	0.43	0.57	0.43	0.57	0.43
Entropía de la cadena de aminoácidos (Zip)	Lineal	0.83	0.17	0.52	0.48	1.02
	Polinomial (grado = 2)	0.78	0.22	0.44	0.56	1.01
	Gauss ($\gamma = 1$)	0.00	1.00	0.00	1.00	0.00
	Gauss ($\gamma = 0.5$)	0.00	1.00	0.00	1.00	0.00
	Gauss ($\gamma = 0.1$)	0.78	0.22	0.44	0.56	1.00
	Gauss ($\gamma = 0.05$)	0.83	0.17	0.52	0.48	1.02
	Gauss ($\gamma = 0.01$)	0.86	0.14	0.57	0.43	1.03

3.6.1. Análisis de los resultados

Del análisis del experimento 3 podemos concluir que los resultados que obtuvimos en la clasificación de polipéptidos desconocidos en el caso de la entropía de Shannon no fueron tan buenos como los obtenidos en la entropía medida mediante la compresión a Zip. Además, en este caso, a diferencia de lo sucedido hasta el momento, la clasificación con mayor valor de *F'-measure* (grisado claro) no resulta ser útil ya que todos los polipéptidos desconocidos son clasificados como positivos. Por esto, decidimos tomar como mejor resultado para el criterio de entropía de Shannon el obtenido con el kernel polinomial de grado dos (grisado oscuro). En el caso del criterio de compresión a Zip el mejor resultado obtenido fue utilizando un kernel gausiano con $\gamma = 0.01$. (ver cuadro 3.10).

Cuadro 3.11: Resultados del experimento 3.2: Entropía de las cadenas. Entrenamiento y testeo con polipéptidos conocidos y no anotados.

Criterio de clasificación	Kernel	Resultados obtenidos				
		Conocidos		No anotados		F'-measure
		+	-	+	-	
Entropía de la cadena de aminoácidos (Shannon)	Lineal	0.90	0.10	0.20	0.80	1.48
	Polinomial (grado = 2)	0.90	0.10	0.18	0.82	1.50
	Gauss ($\gamma = 1$)	0.81	0.19	0.07	0.93	1.50
	Gauss ($\gamma = 0.5$)	0.89	0.11	0.16	0.84	1.51
	Gauss ($\gamma = 0.1$)	0.94	0.06	0.36	0.64	1.35
	Gauss ($\gamma = 0.05$)	0.96	0.04	0.56	0.44	0.98
	Gauss ($\gamma = 0.01$)	0.99	0.01	0.99	0.01	0.98
Entropía de la cadena de aminoácidos (Zip)	Lineal	0.88	0.12	0.10	0.90	1.58
	Polinomial (grado = 2)	0.88	0.12	0.09	0.91	1.59
	Gauss ($\gamma = 1$)	0.29	0.71	0.01	0.99	0.57
	Gauss ($\gamma = 0.5$)	0.31	0.69	0.01	0.99	0.60
	Gauss ($\gamma = 0.1$)	0.87	0.13	0.09	0.91	1.59
	Gauss ($\gamma = 0.05$)	0.88	0.12	0.10	0.90	1.57
	Gauss ($\gamma = 0.01$)	0.89	0.11	0.11	0.89	1.57

En el caso de polipéptidos no anotados obtuvimos mejores resultados en el caso de la entropía de Shannon utilizando un kernel gausiano con $\gamma = 0.5$ y en el caso de la entropía medida mediante la compresión a Zip utilizando un kernel polinomial de grado dos (ver cuadro 3.11).

Una última observación a realizar sobre los resultados de este experimento es que al igual que en los experimentos 1 y 2 los polipéptidos conocidos se diferencian más de los no anotados que de los desconocidos (en relación a los criterios de clasificación elegidos en este experimento). Esto nos permite reafirmar la hipótesis inicial del dominio planteada en la sección 3.1.1.

3.7. Experimento 4: Frecuencia de di-aminoácidos separados

En este experimento utilizamos como criterio de clasificación la frecuencia de aparición de di-aminoácidos (es decir pares de aminoácidos) separados por cadenas de entre uno y cinco aminoácidos. Por ejemplo, si la cantidad de aminoácidos separadores es N, para el di-aminoácido AB, contamos todas las cadenas de la forma $AX_1X_2...X_NB$, donde los X_i representan aminoácidos cualquiera.

Utilizaremos la técnica de *K-fold cross validation* descripta en la sección 3.3.2 y los siete kernels mencionados en la sección 3.3.1. Dividimos el experimento 4 en dos subexperimentos: 4.1 y 4.2. En el experimento 4.1 (cuadro 3.12) entrenamos y testeamos el clasificador mediante polipéptidos conocidos y desconocidos mientras que en el experimento 4.2 (cuadro 3.13) lo entrenamos y testeamos mediante polipéptidos conocidos y no anotados.

3.7.1. Análisis de los resultados

Del análisis del experimento 4 podemos concluir que los mejores resultados que obtuvimos en la clasificación de polipétidos desconocidos para los cinco criterios fue utilizando kernels gausianos con $\gamma = 1$ (ver cuadro 3.12), mientras que en el caso de polipéptidos no anotados obtuvimos mejores resultados para los cinco criterios utilizando kernels polinomiales de grado = 2 (ver cuadro 3.13).

Otra cosa que podemos notar es que en el caso de los polipéptidos desconocidos obtuvimos mejores resultados utilizando N = 3 mientras que en el caso de polipéptidos no anotados fue con N = 1.

Una última observación a realizar sobre los resultados de este experimento es que al igual que en los experimentos 1, 2 y 3 los polipéptidos conocidos se diferencian más de los no anotados que de los desconocidos (en relación a los criterios de clasificación elegidos en este experimento). Esto nos permite reafirmar la hipótesis inicial del dominio planteada en la sección 3.1.1.

Cuadro 3.12: Resultados del experimento 4.1: Frecuencia de di-aminoácidos separados. Entrenamiento y testeo con polipéptidos conocidos y desconocidos.

Criterio de clasificación	Kernel	Resultados obtenidos				
		Conocidos		Desconocidos		F'-measure
		+	-	+	-	
Frecuencia de di-aminoácidos separados por 1 aminoácido	Lineal	0.77	0.23	0.40	0.60	1.00
	Polinomial (grado = 2)	0.77	0.23	0.37	0.63	1.04
	Gauss ($\gamma = 1$)	0.86	0.14	0.50	0.50	1.08
	Gauss ($\gamma = 0.5$)	0.86	0.14	0.50	0.50	1.08
	Gauss ($\gamma = 0.1$)	0.48	0.52	0.23	0.77	0.65
	Gauss ($\gamma = 0.05$)	0.01	0.99	0.00	1.00	0.02
	Gauss ($\gamma = 0.01$)	0.57	0.43	0.57	0.43	0.57
Frecuencia de di-aminoácidos separados por 2 aminoácidos	Lineal	0.78	0.22	0.42	0.58	1.01
	Polinomial (grado = 2)	0.77	0.23	0.39	0.61	1.02
	Gauss ($\gamma = 1$)	0.92	0.08	0.61	0.39	1.10
	Gauss ($\gamma = 0.5$)	0.86	0.14	0.53	0.47	1.07
	Gauss ($\gamma = 0.1$)	0.40	0.60	0.17	0.83	0.56
	Gauss ($\gamma = 0.05$)	0.01	0.99	0.00	1.00	0.02
	Gauss ($\gamma = 0.01$)	0.14	0.86	0.14	0.86	0.14
Frecuencia de di-aminoácidos separados por 3 aminoácidos	Lineal	0.75	0.25	0.40	0.60	0.98
	Polinomial (grado = 2)	0.76	0.24	0.38	0.62	1.02
	Gauss ($\gamma = 1$)	0.92	0.08	0.60	0.40	1.12
	Gauss ($\gamma = 0.5$)	0.85	0.15	0.50	0.50	1.07
	Gauss ($\gamma = 0.1$)	0.39	0.61	0.18	0.82	0.53
	Gauss ($\gamma = 0.05$)	0.02	0.98	0.00	1.00	0.03
	Gauss ($\gamma = 0.01$)	0.43	0.57	0.43	0.57	0.43
Frecuencia de di-aminoácidos separados por 4 aminoácidos	Lineal	0.72	0.28	0.37	0.63	0.96
	Polinomial (grado = 2)	0.72	0.28	0.34	0.66	0.97
	Gauss ($\gamma = 1$)	0.91	0.09	0.60	0.40	1.09
	Gauss ($\gamma = 0.5$)	0.82	0.18	0.48	0.52	1.03
	Gauss ($\gamma = 0.1$)	0.35	0.65	0.15	0.85	0.49
	Gauss ($\gamma = 0.05$)	0.04	0.96	0.01	0.99	0.06
	Gauss ($\gamma = 0.01$)	0.57	0.43	0.57	0.43	0.57
Frecuencia de di-aminoácidos separados por 5 aminoácidos	Lineal	0.75	0.25	0.40	0.60	0.97
	Polinomial (grado = 2)	0.76	0.24	0.38	0.62	1.01
	Gauss ($\gamma = 1$)	0.92	0.08	0.64	0.36	1.09
	Gauss ($\gamma = 0.5$)	0.84	0.16	0.50	0.50	1.06
	Gauss ($\gamma = 0.1$)	0.35	0.65	0.16	0.84	0.48
	Gauss ($\gamma = 0.05$)	0.02	0.98	0.01	0.99	0.03
	Gauss ($\gamma = 0.01$)	0.57	0.43	0.57	0.43	0.57

Cuadro 3.13: Resultados del experimento 4.2: Frecuencia de di-aminoácidos separados. Entrenamiento y testeo con polipéptidos conocidos y no anotados.

Criterio de clasificación	Kernel	Resultados obtenidos				
		Conocidos		No anotados		F'-measure
		+	-	+	-	
Frecuencia de di-aminoácidos separados por 1 aminoácido	Lineal	0.97	0.03	0.09	0.91	1.77
	Polinomial (grado = 2)	0.97	0.03	0.08	0.92	1.79
	Gauss ($\gamma = 1$)	0.99	0.01	0.25	0.75	1.59
	Gauss ($\gamma = 0.5$)	0.99	0.01	0.15	0.85	1.70
	Gauss ($\gamma = 0.1$)	0.96	0.04	0.13	0.87	1.69
	Gauss ($\gamma = 0.05$)	0.96	0.04	0.16	0.84	1.19
	Gauss ($\gamma = 0.01$)	0.99	0.01	0.65	0.35	1.19
Frecuencia de di-aminoácidos separados por 2 aminoácidos	Lineal	0.96	0.04	0.09	0.91	1.76
	Polinomial (grado = 2)	0.97	0.03	0.08	0.92	1.78
	Gauss ($\gamma = 1$)	0.99	0.01	0.25	0.75	1.59
	Gauss ($\gamma = 0.5$)	0.99	0.01	0.16	0.84	1.70
	Gauss ($\gamma = 0.1$)	0.96	0.04	0.13	0.87	1.69
	Gauss ($\gamma = 0.05$)	0.96	0.04	0.17	0.83	1.14
	Gauss ($\gamma = 0.01$)	0.99	0.01	0.73	0.27	1.14
Frecuencia de di-aminoácidos separados por 3 aminoácidos	Lineal	0.96	0.04	0.10	0.90	1.75
	Polinomial (grado = 2)	0.96	0.04	0.09	0.91	1.77
	Gauss ($\gamma = 1$)	0.99	0.01	0.28	0.72	1.55
	Gauss ($\gamma = 0.5$)	0.98	0.02	0.17	0.83	1.69
	Gauss ($\gamma = 0.1$)	0.96	0.04	0.14	0.86	1.68
	Gauss ($\gamma = 0.05$)	0.96	0.04	0.17	0.83	1.13
	Gauss ($\gamma = 0.01$)	0.99	0.01	0.74	0.26	1.13
Frecuencia de di-aminoácidos separados por 4 aminoácidos	Lineal	0.96	0.04	0.10	0.90	1.75
	Polinomial (grado = 2)	0.97	0.03	0.09	0.91	1.77
	Gauss ($\gamma = 1$)	1.00	0.00	0.29	0.71	1.54
	Gauss ($\gamma = 0.5$)	0.98	0.02	0.17	0.83	1.68
	Gauss ($\gamma = 0.1$)	0.96	0.04	0.13	0.87	1.68
	Gauss ($\gamma = 0.05$)	0.96	0.04	0.17	0.83	1.63
	Gauss ($\gamma = 0.01$)	0.99	0.01	0.65	0.35	1.19
Frecuencia de di-aminoácidos separados por 5 aminoácidos	Lineal	0.96	0.04	0.10	0.90	1.75
	Polinomial (grado = 2)	0.97	0.03	0.09	0.91	1.77
	Gauss ($\gamma = 1$)	1.00	0.00	0.30	0.70	1.53
	Gauss ($\gamma = 0.5$)	0.99	0.01	0.18	0.82	1.67
	Gauss ($\gamma = 0.1$)	0.96	0.04	0.14	0.86	1.68
	Gauss ($\gamma = 0.05$)	0.96	0.04	0.17	0.83	1.63
	Gauss ($\gamma = 0.01$)	0.99	0.01	0.80	0.20	1.09

3.8. Experimento 5: Periodicidad de aminoácidos

En este experimento emplearemos como criterio para la clasificación de los polipéptidos la periodicidad de los distintos aminoácidos presentes en el mismo. Para el cálculo de la periodicidad nos basaremos en la formulación presentada en [Cornette *et al.* (1987)]. Definimos la periodicidad de un polipéptido a frecuencia w como:

$$F = \sqrt{\sum_{k=0}^{l-1} h_k \cos(kw) + \sum_{k=0}^{l-1} h_k \sin(kw)}$$

donde $h_0 \dots h_{l-1}$ representa la cadena, en la que cada aminoácido fue reemplazado por un valor numérico único.

En este experimento calcularemos la periodicidad de los polipéptidos para las frecuencias w en el rango de los 0 a los 180 grados, con saltos de 10 grados. Utilizaremos la técnica de *K-fold cross validation* descripta en la sección 3.3.2 y los siete kernels mencionados en la sección 3.3.1.

Dividimos el experimento 5 en dos subexperimentos: 5.1 y 5.2. En el experimento 5.1 (cuadro 3.14) entrenamos y testeamos el clasificador mediante polipéptidos conocidos y desconocidos mientras que en el experimento 5.2 (cuadro 3.15) lo entrenamos y testeamos mediante polipéptidos conocidos y no anotados.

Cuadro 3.14: Resultados del experimento 5.1: Periodicidad de aminoácidos. Entrenamiento y testeo con polipéptidos conocidos y desconocidos.

Criterio de clasificación	Kernel	Resultados obtenidos				
		Conocidos		Desconocidos		F'-measure
		+	-	+	-	
Periodicidad de aminoácidos	Lineal	0.79	0.21	0.48	0.52	0.99
	Polinomial (grado = 2)	0.76	0.24	0.42	0.58	0.97
	Gauss ($\gamma = 1$)	0.00	1.00	0.00	1.00	0.00
	Gauss ($\gamma = 0.5$)	0.00	1.00	0.00	1.00	0.00
	Gauss ($\gamma = 0.1$)	0.73	0.27	0.38	0.62	0.95
	Gauss ($\gamma = 0.05$)	0.79	0.21	0.48	0.52	0.99
	Gauss ($\gamma = 0.01$)	0.84	0.16	0.58	0.42	0.99

Cuadro 3.15: Resultados del experimento 5.2: Periodicidad de aminoácidos. Entrenamiento y testeo con polipéptidos conocidos y no anotados.

Criterio de clasificación	Kernel	Resultados obtenidos				
		Conocidos		No anotados		F'-measure
		+	-	+	-	
Periodicidad de aminoácidos	Lineal	0.86	0.14	0.11	0.89	1.54
	Polinomial (grado = 2)	0.86	0.14	0.10	0.90	1.54
	Gauss ($\gamma = 1$)	0.00	1.00	0.00	1.00	0.00
	Gauss ($\gamma = 0.5$)	0.00	1.00	0.00	1.00	0.00
	Gauss ($\gamma = 0.1$)	0.85	0.15	0.09	0.91	1.54
	Gauss ($\gamma = 0.05$)	0.86	0.14	0.11	0.89	1.53
	Gauss ($\gamma = 0.01$)	0.88	0.12	0.14	0.86	1.52

3.8.1. Análisis de los resultados

Del análisis del experimento 5 podemos concluir que los mejores resultados que obtuvimos en la clasificación de polipéptidos desconocidos fue utilizando un kernel gausiano con $\gamma = 0.01$ (ver cuadro 3.12) mientras que en la clasificación de polipéptidos no anotados obtuvimos mejores resultados utilizando un kernel polinomial de grado = 2 (ver cuadro 3.13).

Otra observación a realizar es que al igual que en los experimentos 1, 2, 3 y 4 los polipéptidos conocidos se diferencian más de los no anotados que de los desconocidos (en relación a los criterios de clasificación elegidos en este experimento). Esto nos permite reafirmar la hipótesis inicial del dominio planteada en la sección 3.1.1.

3.9. Experimento 6: Combinación de criterios

El objetivo de este experimento es definir un criterio compuesto que resulte de la combinación de los diez criterios individuales presentados en este trabajo, cuyo desempeño iguale o supere a los obtenidos por cada criterio en forma individual y también definir una técnica de priorización de los polipéptidos según su probabilidad de representar una proteína funcional de baculovirus.

Para este experimento utilizaremos los clasificadores que dieron mejores resultados en cada experimento. Los detallamos en los cuadros 3.16 (polipéptidos conocidos y desconocidos) y 3.17 (polipéptidos conocidos y no anotados).

Al igual que en el experimento 1, cada uno de los tres conjuntos de polipéptidos mencionados en la sección 3.1 fue dividido en dos subconjuntos, uno de los cuales será utilizado para entrenar la SVM (conteniendo 44 proteomas) y el otro para testearla (conteniendo 5 proteomas).

Dividimos este experimento en dos subexperimentos: el 6.1 en el que compararemos las clasificaciones de polipéptidos conocidos y desconocidos (ver cuadro 3.18), y el 6.2 en el que compararemos las clasificaciones de polipéptidos conocidos y no anotados (ver cuadro 3.19). Usamos la notación $N+$, con $0 \leq N \leq 10$ para denotar que N criterios clasificaron a ese porcentaje de polipéptidos como positivos.

Cuadro 3.16: Clasificadores de polipéptidos conocidos y desconocidos utilizados en el experimento 6: Combinación de criterios.

Criterio de clasificación	Kernel
Frecuencia de aminoácidos	Gauss ($\gamma = 0.5$)
Frecuencia de di-aminoácidos	Gauss ($\gamma = 1$)
Entropía de la cadena de aminoácidos (Shannon)	Polinomial (grado = 2)
Entropía de la cadena de aminoácidos (Zip)	Gauss ($\gamma = 0.01$)
Frecuencia de di-aminoácidos separados por 1 aminoácido	Gauss ($\gamma = 1$)
Frecuencia de di-aminoácidos separados por 2 aminoácidos	Gauss ($\gamma = 1$)
Frecuencia de di-aminoácidos separados por 3 aminoácidos	Gauss ($\gamma = 1$)
Frecuencia de di-aminoácidos separados por 4 aminoácidos	Gauss ($\gamma = 1$)
Frecuencia de di-aminoácidos separados por 5 aminoácidos	Gauss ($\gamma = 1$)
Periodicidad de aminoácidos	Gauss ($\gamma = 0.01$)

Cuadro 3.17: Clasificadores de polipéptidos conocidos y no anotados utilizados en el experimento 6: Combinación de criterios.

Criterio de clasificación	Kernel
Frecuencia de aminoácidos	Polinomial (grado = 2)
Frecuencia de di-aminoácidos	Polinomial (grado = 2)
Entropía de la cadena de aminoácidos (Shannon)	Gauss ($\gamma = 0.5$)
Entropía de la cadena de aminoácidos (Zip)	Polinomial (grado = 2)
Frecuencia de di-aminoácidos separados por 1 aminoácido	Polinomial (grado = 2)
Frecuencia de di-aminoácidos separados por 2 aminoácidos	Polinomial (grado = 2)
Frecuencia de di-aminoácidos separados por 3 aminoácidos	Polinomial (grado = 2)
Frecuencia de di-aminoácidos separados por 4 aminoácidos	Polinomial (grado = 2)
Frecuencia de di-aminoácidos separados por 5 aminoácidos	Polinomial (grado = 2)
Periodicidad de aminoácidos	Polinomial (grado = 2)

Cuadro 3.18: Resultados del experimento 6: Combinación de criterios. Clasificación de polipéptidos conocidos y desconocidos.

Resultado	Conocidos	Desconocidos
10+	68.39 %	17.99 %
9+	8.43 %	12.19 %
8+	7.96 %	8.84 %
7+	7.50 %	9.45 %
6+	1.87 %	7.93 %
5+	2.34 %	7.32 %
4+	0.70 %	7.62 %
3+	1.87 %	11.59 %
2+	0.47 %	7.01 %
1+	0.47 %	4.88 %
0+	0.00 %	5.18 %

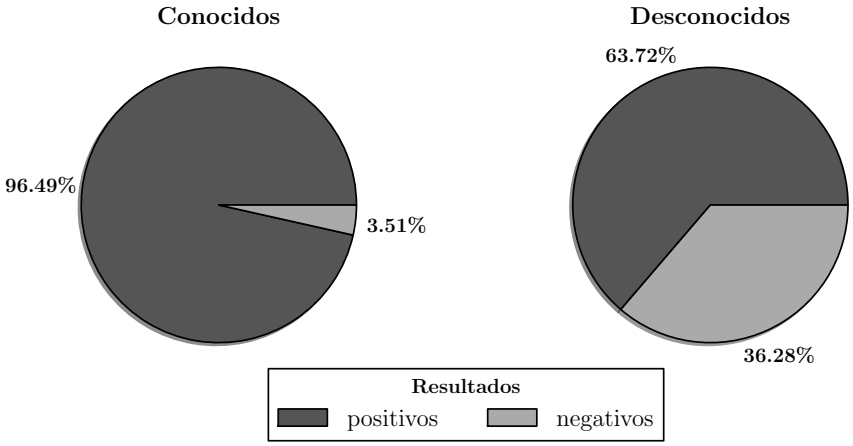

Figura 3.4: Resultados obtenidos en la clasificación de polipéptidos conocidos y desconocidos considerando positivos los que obtuvieron una puntuación $N+$ con $N \geq 5$. F'-measure: 1.16

Cuadro 3.19: Resultados del experimento 6: Combinación de criterios. Clasificación de polipéptidos conocidos y no anotados.

Resultado	Conocidos	No anotados
10+	78.46 %	0.00 %
9+	4.22 %	0.08 %
8+	6.32 %	0.69 %
7+	3.98 %	1.61 %
6+	2.11 %	0.77 %
5+	1.17 %	1.61 %
4+	0.47 %	1.61 %
3+	1.87 %	4.06 %
2+	1.17 %	8.05 %
1+	0.23 %	17.26 %
0+	0.00 %	64.26 %

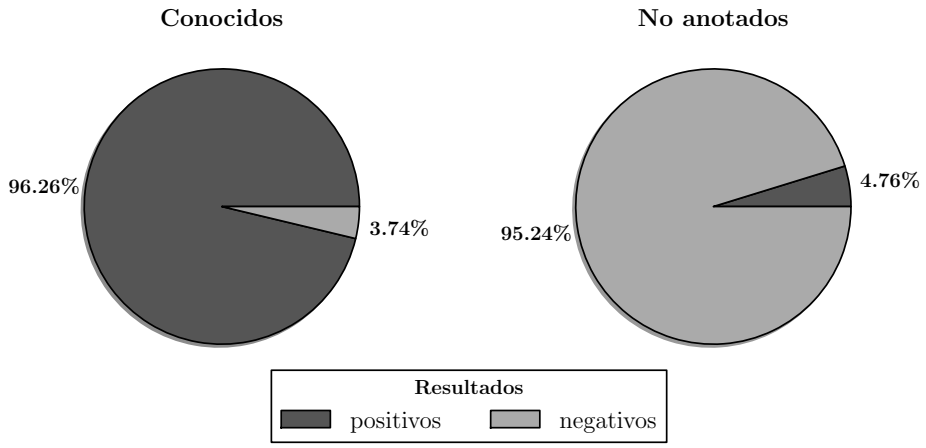

Figura 3.5: Resultados obtenidos en la clasificación de polipéptidos conocidos y no anotados considerando positivos los que obtuvieron una puntuación $N+$ con $N \geq 5$. F'-measure: 1.82

3.9.1. Análisis de los resultados

Uno de los resultados más notables de este experimento es que tanto en la clasificación de polipéptidos conocidos y desconocidos (cuadro 3.18) como en la de conocidos y no anotados (cuadro 3.19) todos los polipéptidos conocidos (que son en su totalidad casos positivos, como antes explicamos) son detectados como positivos por al menos uno de los diez criterios utilizados. Este resultado es muy alentador ya que implica que la combinación de criterios logra cubrir los distintos aspectos que caracterizan a las proteínas funcionales.

Una segunda observación, más importante aún, es que si trazamos una línea que divida los polipéptidos que fueron clasificados como positivos por al menos 5 criterios, es decir los que obtuvieron una puntuación $N+$ con $N \geq 5$, obtenemos un nuevo criterio de clasificación compuesto cuyo desempeño (medido con *F'-measure*) es superior a cualquiera de los criterios utilizados individualmente (ver figuras 3.4 y 3.5).

Por otro lado, los cuadros 3.18 y 3.19 revelan un posible criterio de priorización de los polipéptidos, ya que podemos considerar que los clasificados como positivos por N criterios son más probables de serlo en realidad, luego ubicaríamos a los clasificados como positivos por $N-1$ criterios, y así sucesivamente. Esto es muy útil a la hora de orientar los experimentos de mesada húmeda destinados a de-

tectar cuáles de los polipéptidos representan proteínas funcionales de baculovirus.

Otro uso que puede darse a los resultados mostrados en 3.18 y 3.19 es la selección de casos positivos y negativos más representativos del conjunto para ser usados en el entrenamiento de la SVM.

Capítulo 4

Puesta a prueba del clasificador

En este capítulo realizamos una simulación de uso real del clasificador de polipéptidos multi-criterio que hemos construido en el capítulo 3. Primero detallamos el algoritmo propuesto para la anotación de un nuevo genoma baculoviral y luego realizamos cuatro experimentos para medir el desempeño del clasificador. A continuación, evaluamos la eficiencia tomando como base la anotación de sus autores (aún sabiendo que esta puede contener errores).

4.1. Algoritmo de anotación propuesto

El análisis y la anotación de un genoma novedoso normalmente consta de una serie de pasos lógicos. A partir de la secuencia genómica obtenida en el laboratorio (genoma) se genera un listado de ORFs teóricos (orfeoma) que luego se traduce a su correspondiente proteoma teórico (ver sección 1.1). Finalmente se utiliza algún mecanismo para realizar su anotación en una base de datos. A continuación detallamos los pasos a seguir para la anotación de un genoma novedoso mediante el uso de nuestro clasificador, combinado con herramientas estándar para la asignación de identidades a los polipéptidos.

1. Obtención de la secuencia genómica (DNA)
2. Búsqueda de ORFs: ATG–(múltiplo de 3)–Codón de Stop (límite, no menos de 45 aminoácidos)
3. Eliminación de redundancias internas (múltiples inicios, igual posición del codón de stop)
4. Listado de ORFs teóricos
5. Traducción in silico
6. Proteoma máximo teórico (FASTA múltiple): no anotado, pero en principio anotable

7. Análisis con el clasificador multi-criterio
8. Subdivisión en dos proteomas teóricos, el anotable y el no anotable
9. Utilizando otras estrategias estándar (BlastP, Psi-Blast, Prosite, Pfam, ProDom, etc.), asignación de identidades a los miembros del proteoma anotable
10. Anotación

4.2. Experimento ciego: Análisis de proteomas baculovirales teóricos

Para verificar el uso del clasificador multi-criterio como una herramienta que permite un análisis rápido de los proteomas máximos* y la obtención del proteoma anotable con alta eficiencia, se simuló la situación de uso futuro realizando dos experimentos ciegos, empleando dos genomas novedosos de baculovirus. Una vez obtenidos los listados de proteomas anotables y no anotables se contrastaron con las anotaciones de los autores originales para evaluar el porcentaje de positivos detectados y cuántos polipéptidos adicionales se obtuvieron. Se consideraron como positivos los polipéptidos clasificados como tales por al menos cuatro criterios ($N+$ con $N \geq 4$).

Genoma A

1. DNA, circular, covalentemente cerrado, de 108592 bp
2. Búsqueda de ORFs teóricos: 853 ORFs (límite 138 pb)
3. Filtro de eliminación de redundancias (múltiples inicios, igual posición del codón de stop)
4. Listado de ORFs teóricos: 230 ORFs (ambas polaridades)
5. Traducción in silico usando el código genético universal
6. Proteoma máximo teórico: 230 polipéptidos
7. Análisis con el clasificador múltiple

Genoma B

1. DNA, circular, covalentemente cerrado, de 123876 bp
2. Búsqueda de ORFs teóricos: 1025 ORFs (límite 138 pb)
3. Filtro de eliminación de redundancias (múltiples inicios, igual posición del codón de stop)
4. Listado de ORFs teóricos: 281 ORFs (ambas polaridades)

*El proteoma máximo es un listado de todos los polipéptidos teóricos codificados en un genoma determinado considerando que todos los ORFs independientes detectados en el mismo podrían ser traducidos.

5. Traducción in silico usando el código genético universal
6. Proteoma máximo teórico: 281 polipéptidos
7. Análisis con el clasificador múltiple

Testeamos la clasificación de los polipéptidos de los proteomas teóricos obtenidos a partir de los genomas A y B con dos conjuntos de datos distintos para el entrenamiento positivo: polipéptidos anotados y polipéptidos conocidos. En ambos se utilizó como conjunto de datos para el entrenamiento negativo polipéptidos no anotados. Los mismos fueron extraídos de los 49 proteomas obtenidos del GenBank que utilizamos en los experimentos del capítulo 3, usando como anotados al conjunto de polipéptidos conocidos y desconocidos.

Cuadro 4.1: Resultados del experimento ciego: Análisis de proteomas baculovirales teóricos.

Conjunto de datos de entrenamiento		Genoma	Resultados obtenidos				
			Anotados		No anotados		F'-measure
+	-		+	-	+	-	
Anotados	No anotados	A	99.17	0.83	13.64	86.36	1.72
Conocidos	No anotados	A	97.50	2.50	10.91	89.09	1.75
Anotados	No anotados	B	94.87	5.13	29.88	70.12	1.44
Conocidos	No anotados	B	94.02	5.98	27.44	72.56	1.46

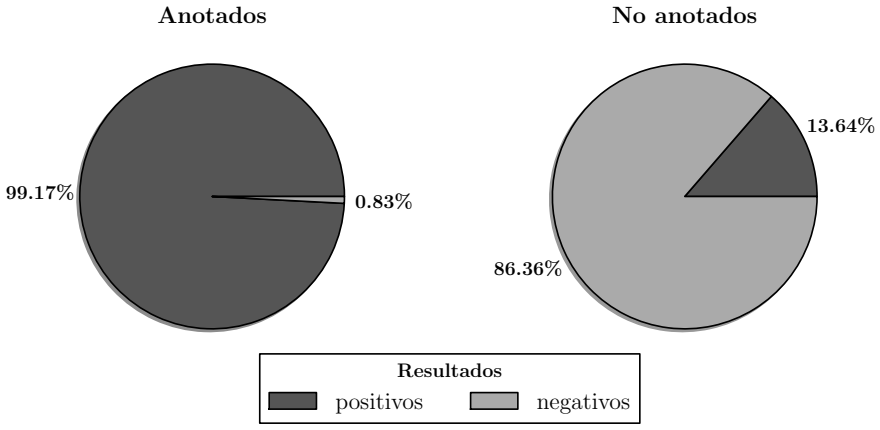

Figura 4.1: Resultados del experimento ciego. Genoma A. Entrenamiento con anotados y no anotados.

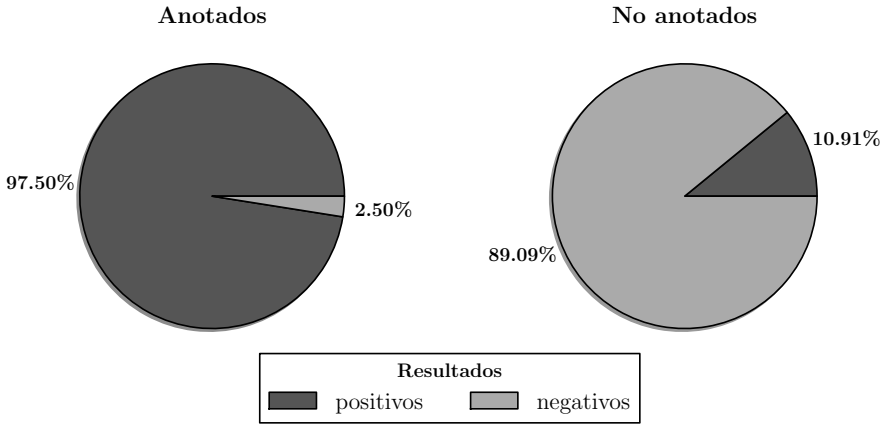

Figura 4.2: Resultados del experimento ciego. Genoma A. Entrenamiento con conocidos y no anotados.

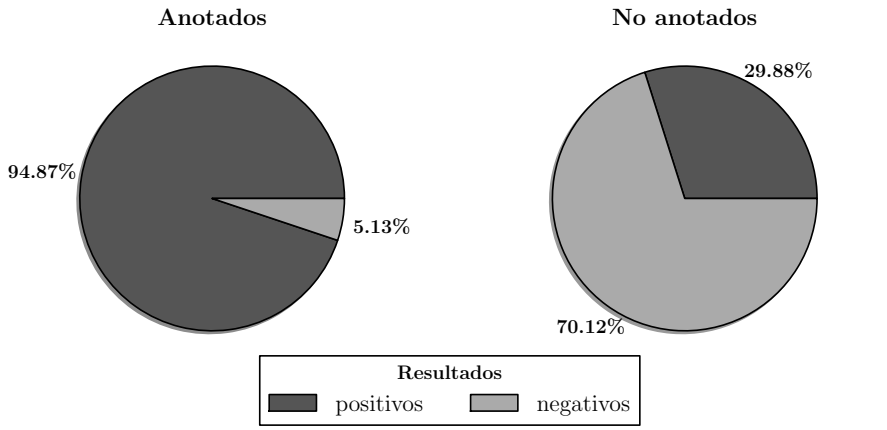

Figura 4.3: Resultados del experimento ciego. Genoma B. Entrenamiento con anotados y no anotados.

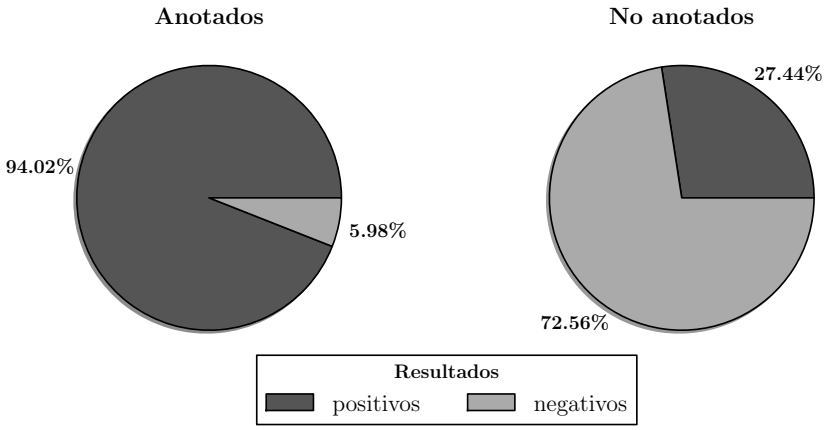

Figura 4.4: Resultados del experimento ciego. Genoma B. Entrenamiento con conocidos y no anotados.

4.2.1. Análisis de los resultados

En el cuadro 4.1 se observa que los resultados fueron suficientemente buenos para ambos conjuntos de datos de entrenamiento, siendo un poco mejores (según *F'-measure*) los obtenidos utilizando para el entrenamiento polipéptidos conocidos y no anotados. Para ambos proteomas ocurre que el porcentaje de polipéptidos anotados reconocidos como tales es superior en el caso de entrenamiento con polipéptidos anotados y no anotados, mientras que el porcentaje de polipéptidos no anotados reconocidos como tales es superior en el caso de entrenamiento con polipéptidos conocidos y no anotados.

Por otro lado, tomando como premisa de partida que los proteomas anotados en el GenBank son correctos, el método funciona con una eficiencia superior al 94 %. Sin embargo, dado que no todas las proteínas anotadas en el GenBank tienen identidad conocida, y la dificultad para definir un proteoma verdadero negativo, el porcentaje de eficiencia en realidad puede ser diferente. No obstante esto, el clasificador es muy útil para obtener de manera muy rápida un primer borrador de proteoma anotable, lo cual lo transforma en una herramienta valiosa para la anotación de nuevos genomas baculovirales.

Capítulo 5

Conclusiones

En este capítulo enumeramos algunas de las posibles líneas futuras de investigación y damos una reflexión final del trabajo realizado exponiendo las conclusiones obtenidas.

5.1. Trabajo futuro

Existen varias formas en que este trabajo puede ser continuado y mejorado.

Una de las mejoras más directas que podemos mencionar es el agregado de nuevos criterios que permitan mejorar el desempeño del clasificador construido. Los mismos podrán basarse en aspectos composicionales (como los que hemos utilizado en este trabajo) o evaluar otras características presentes en los polipéptidos.

Otra posibilidad de mejora es trabajar en la selección de los polipéptidos más representativos de los distintos conjuntos de datos para utilizarlos como casos de entrenamiento y de esta manera mejorar los resultados de la clasificación.

En nuestra implementación del clasificador todos los criterios son valorados igualmente para la priorización de los polipéptidos. Sin embargo, dado que el desempeño de algunos criterios es visiblemente mejor que el de otros, puede ampliarse la estrategia de priorización incluyendo un valorador de la importancia de cada criterio, de modo de darles prioridad a los polipéptidos elegidos por los criterios de mejor desempeño.

5.2. Reflexión

A lo largo de este trabajo logramos construir un clasificador capaz de separar aquellos polipéptidos que tienen una mayor chance de representar proteínas funcionales del resto con un error aceptable. Esto es muy útil tanto a la hora de anotar un proteoma como de definir qué polipéptidos priorizar para los experimentos de mesada húmeda que, como señalamos anteriormente, son muy costosos en términos de tiempo y dinero. Así, utilizando el clasificador, en algunas horas (y sin un costo monetario significativo) puede tenerse un borrador de anotación mientras que cada experimento de mesada húmeda lleva meses e implica un costo muy alto en reactivos, herramientas y personal capacitado.

Una apreciación importante es que la línea de separación de los polipéptidos puede variarse dependiendo de las circunstancias para incluir un mayor número de conocidos o excluir un mayor número de desconocidos o no anotados. En el primer caso se incluirían en la selección un menor número de falsos negativos y en el segundo se disminuiría la cantidad de falsos positivos. En general en el caso de una anotación es preferible evitar los falsos negativos, ya que se descartarían proteínas funcionales impidiendo su posterior detección en baculovirus, mientras que en el caso de la selección de polipéptidos anotados es preferible evitar los falsos positivos, ya que implicarían un costo económico inútil.

En síntesis, creemos que este trabajo representa un aporte significativo ya que, desde un punto de vista práctico, será de gran ayuda en las tareas de anotación y selección de polipéptidos en baculovirus (y puede extenderse para ser usado en otros organismos eucariotas), y desde un punto de vista teórico expone un mecanismo capaz de detectar proteínas funcionales basado únicamente en criterios composicionales.

Apéndice A

Sobre baculovirus

La intención de este apéndice es ampliar algunos de los conceptos importantes de la genética de los baculovirus. Para ello presentamos un breve resumen del material escrito por Víctor Romanowski y Pablo Daniel Ghiringhelli [Romanowski & Ghiringhelli (2001)] como autores invitados en *Los Baculovirus y sus Aplicaciones como Bioinsecticidas en el Control Biológico de Plagas* [Caballero *et al.* (2001)]. Las imágenes que incluimos también son de su autoría.

A.1. Introducción

Los baculovirus comprenden una familia numerosa de virus que infectan exclusivamente a artrópodos, exhibiendo, salvo contadas excepciones, un rango de huéspedes muy estrecho. Su patrón de replicación es complejo e incluye la producción de progenie con dos fenotipos diferentes, que poseen la misma información genética.

La producción de los fenotipos se encuentra regulada y ocurre en sitios y etapas diferentes de la infección. Los primeros viriones que se producen durante el ciclo de infección brotan de la membrana plasmática y reciben el nombre de virus brotantes (*budded virus, BV*), mientras que en etapas tardías se observa la formación de viriones rodeados por una matriz proteica pseudocristalina, que se acumula en la célula como cuerpos de inclusión que contienen una o más nucleocápsides envueltas (*occluded virus, OV*).

En función de la morfología de los cuerpos de inclusión y de algunas características diferenciales en la biología de la infección se distinguen dos géneros en la familia Baculoviridae: Nucleopolyhedrovirus (NPV), cuyos cuerpos de inclusión se encuentran en el núcleo de la célula infectada, son poliédricos y pueden con-

tener uno o más viriones, y Granulovirus (GV), cuyos cuerpos de inclusión son ovoides, de menor tamaño, y se detectan en las células infectadas que ya han perdido la integridad de la membrana nuclear.

A.2. Estructura molecular de los viriones

Los genomas de los diferentes baculovirus son de DNA circular de doble cadena de 78 a 200 kb y se encuentran asociados con proteínas básicas constituyendo nucleocápsides con forma de bastón de 40 x 200-400 nm. Las nucleocápsides se encuentran envueltas, a su vez, por una membrana lipídica que contiene proteínas codificadas por el virus.

En este punto cabe diferenciar los dos fenotipos virales que constituyen una característica única de la familia Baculoviridae: virus brotantes (BVs) y virus ocluidos (OVs). Los viriones brotantes se producen en la fase temprana de la infección, cuando las nucleocápsides brotan por la membrana plasmática modificada con proteínas codificadas por el genoma viral. Por el contrario, en la fase muy tardía de la infección se forman los cuerpos de inclusión rodeando a las nucleocápsides, que adquieren su membrana en el núcleo. La matriz proteica de estas inclusiones está formada por el acúmulo de una proteína mayoritaria denominada poliedrina o granulina, según el género de baculovirus, NPV o GV, respectivamente.

En la mayoría de los OVs de NPVs se observan varios viriones por poliedro y en algunas especies denominadas MNPVs (Multiple nucleocapsid Nucleopolyhedrovirus) se ven varios manojos de nucleocápsides envueltas en una membrana, mientras que los gránulos de los GVs, generalmente, poseen un solo virión. Eventualmente, pueden observarse algunos cuerpos de inclusión sin viriones (poliedros vacíos). La disolución de la matriz proteica de los cuerpos de inclusión libera viriones denominados ODV (occlusion derived virions) con una envoltura diferente a la de los BVs. En la figura A.1 se esquematizan los distintos tipos de viriones.

A.3. Infección primaria

Los cuerpos de inclusión (OVs, poliedros de NPV o gránulos de GV) se liberan al ambiente como consecuencia de la muerte y lisis de los insectos infectados, que ocurre en las etapas muy tardías de la infección. Estos OVs son eventualmente ingeridos junto con el follaje que es consumido por un nuevo huésped y disueltos en su intestino medio, donde el pH es alcalino. A pH 9-10 se activa una proteasa

de origen celular que acelera el proceso de disolución de la proteína de oclusión y liberación de los ODVs. Estos ODVs infectan las células epiteliales del intestino con gran eficiencia. Las evidencias de los estudios de microscopía electrónica y de inhibidores de la endocitosis sugieren que las nucleocápsides de los ODV entran por fusión de la membrana viral con la membrana plasmática de la célula epitelial [Granados (1978)]. En el caso de los GVs la entrada es facilitada por la actividad de una proteína denominada enhancin, presente en los gránulos.

La membrana de los ODVs contiene una serie de proteínas: ODV E25, ODV-E66, ODV-E56 y P74. Además, por debajo de la membrana del virión, en el llamado tegumento se localiza la GP41. No se conoce la función de las diferentes proteínas, aunque se ha demostrado que las mutaciones que inactivan al gen p74 de AcMNPV conducen a la formación de ODVs no infecciosos [Faulkner *et al.* (1997)]. Por otra parte, la membrana lipídica posee una estructura típica de bicapa, que posiblemente sea derivada de las microvesículas que se forman a partir de la membrana nuclear interna de las células infectadas.

A.4. Infección secundaria

Los BVs son el producto de la replicación viral en la etapa tardía y contienen una sola nucleocápside envuelta en una membrana, que proviene de la brotación de la membrana plasmática de la célula infectada. En la infección de una larva de lepidóptero el ciclo de replicación que inician los ODVs en la célula epitelial produce una progenie de BVs, que brota de la membrana basolateral e infecta otros tejidos (figura A.2). En uno de los extremos del virión se destacan unas estructuras protuberantes o peplómeros (*spikes*), formadas por la proteína de fusión GP64, que no se encuentra en los ODVs (figura A.1).

Granados y Lawler propusieron un mecanismo alternativo de producción de BVs, que no requiere de la replicación y consiste en una rápida conversión de ODVs en BVs [Granados & Lawler (1981)] (figura A.3). Esto es coherente con el hecho de que GP64 se expresa tanto en la fase temprana (antes de la replicación del DNA viral) como en la tardía. Este mecanismo puede haber sido seleccionado por presentar una ventaja evolutiva. De hecho, las células epiteliales del intestino medio infectadas se descaman más rápidamente y se eliminan. Para permitir que la infección pase rápidamente a otros tejidos, antes de que ocurra la descamación de las células, deben producirse en un tiempo muy corto los BVs capaces de infectar a otras células que se encuentran en las proximidades o pasar a los conductos traqueales y la hemolinfa. Por ello, es muy ventajoso para la propagación de la infección que una parte de las nucleocápsides que entraron a la célula se desnuden

y expresen GP64, mientras otras migran a la porción basal para brotar por esa membrana, en la que se inserta GP64 (y, eventualmente, otros productos génicos recién sintetizados). Este mecanismo no provee las ventajas de amplificación viral asociada a la replicación, pero asegura un rápido movimiento del virus desde la capa epitelial hacia otros tejidos internos. Parece razonable la coexistencia de ambos mecanismos para asegurar la propagación del virus.

La GP64 está presente en la superficie de las células infectadas por baculovirus en forma de trímero. Se ha documentado su actividad fusogénica que es esencial para liberar la nucleocápside al citosol de la célula [Hohmann & Faulkner (1983)]. Esta actividad ha sido mapeada a un hexapéptido hidrofóbico localizado en la parte central del polipéptido [Monsma & Blissard (1995)]. En contraste con los ODVs, los BVs entran a la célula huésped por endocitosis mediada, probablemente, por interacción de GP64 con un receptor aún no identificado (figura A.4).

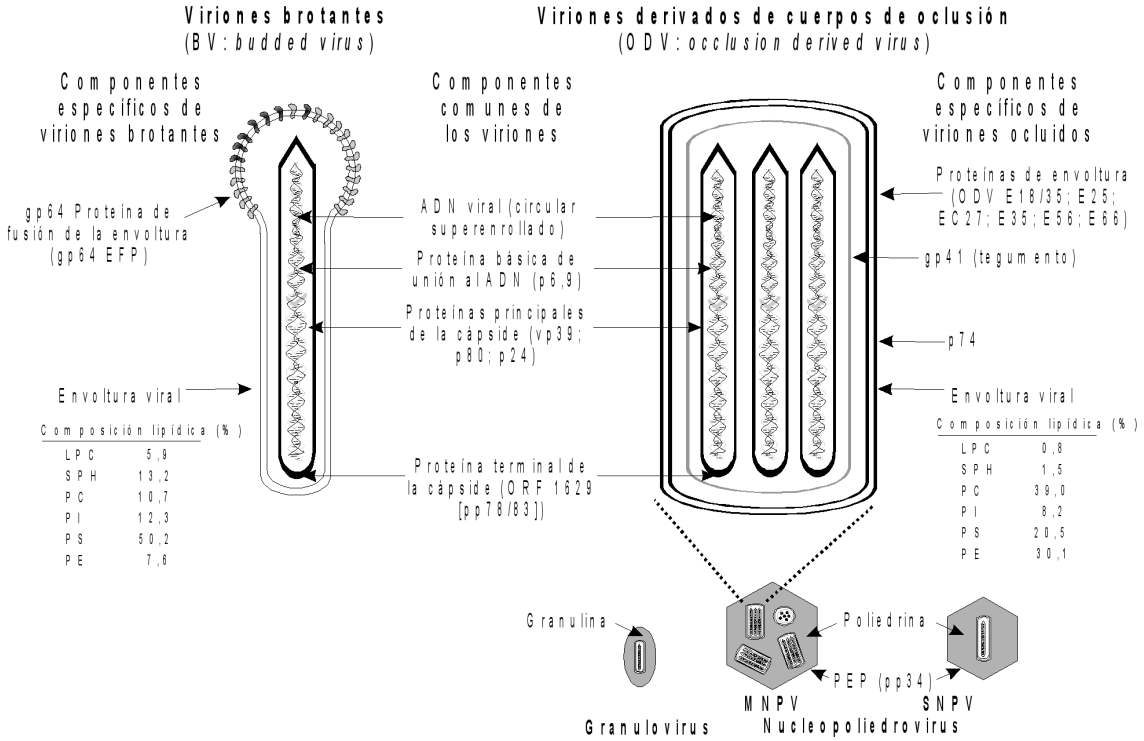

Figura A.1: Componentes estructurales básicos de los dos fenotipos virales.

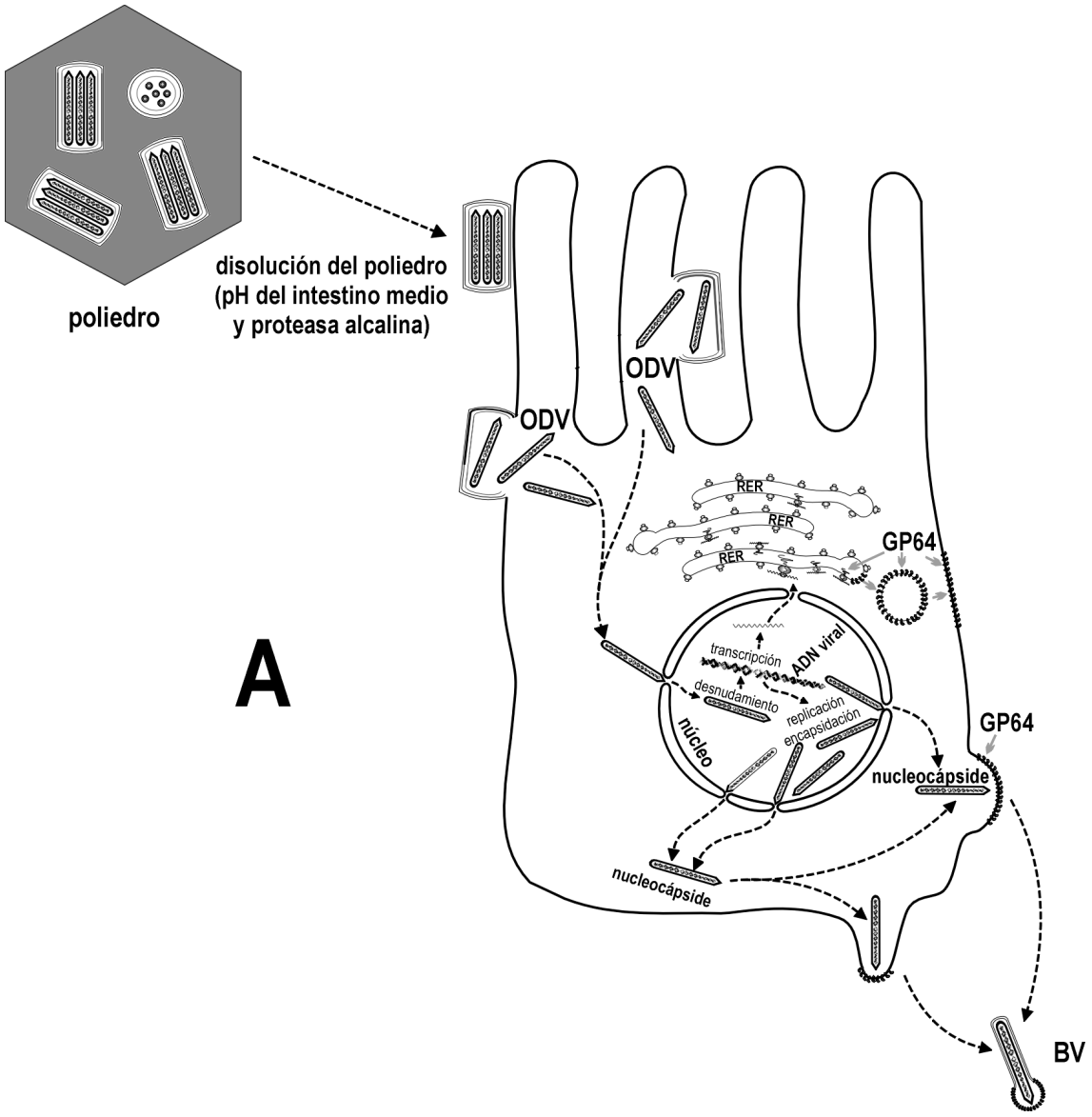

Figura A.2: Infección primaria y producción de BVs.

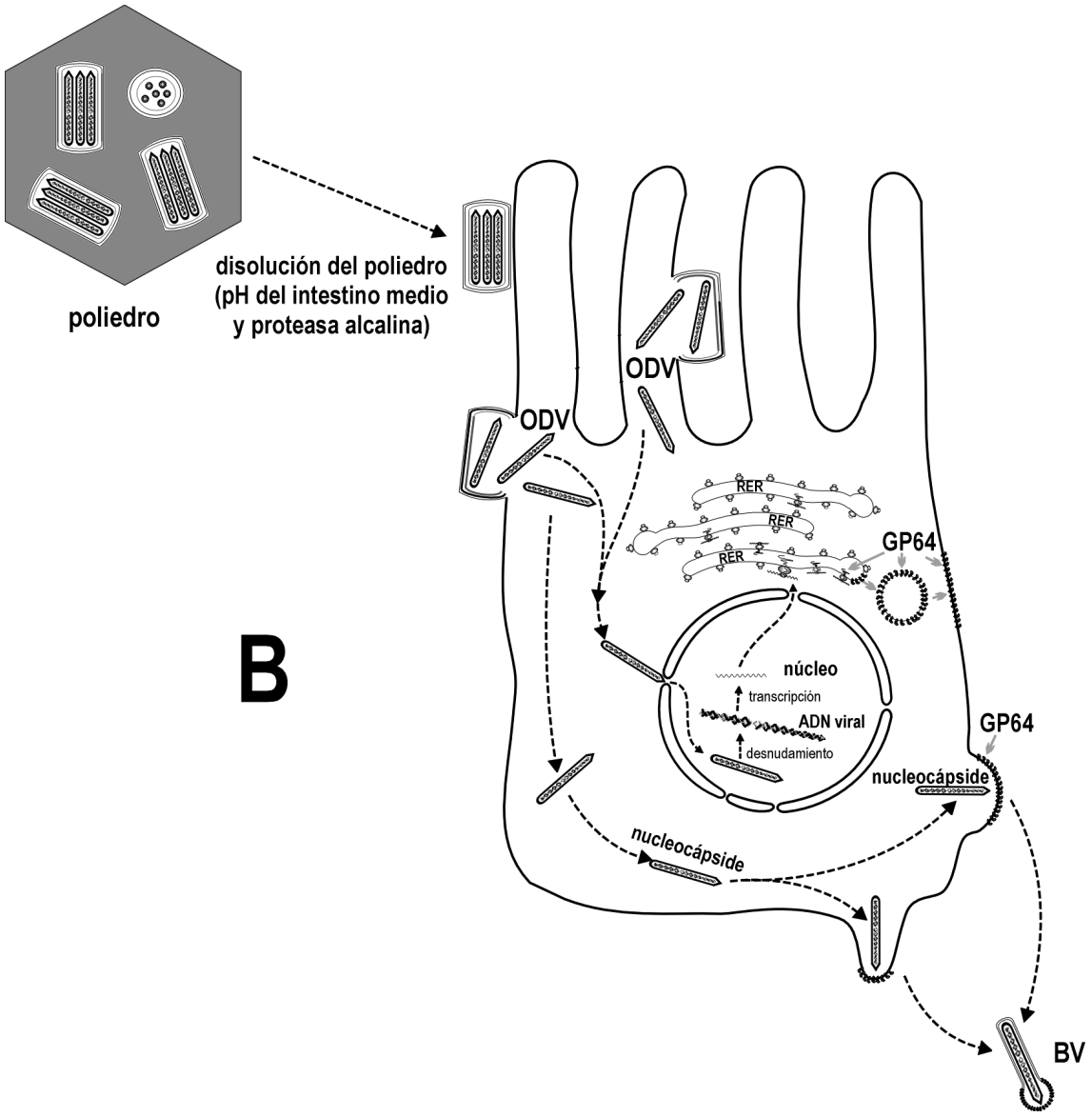

Figura A.3: Infección primaria y producción rápida de BVs.

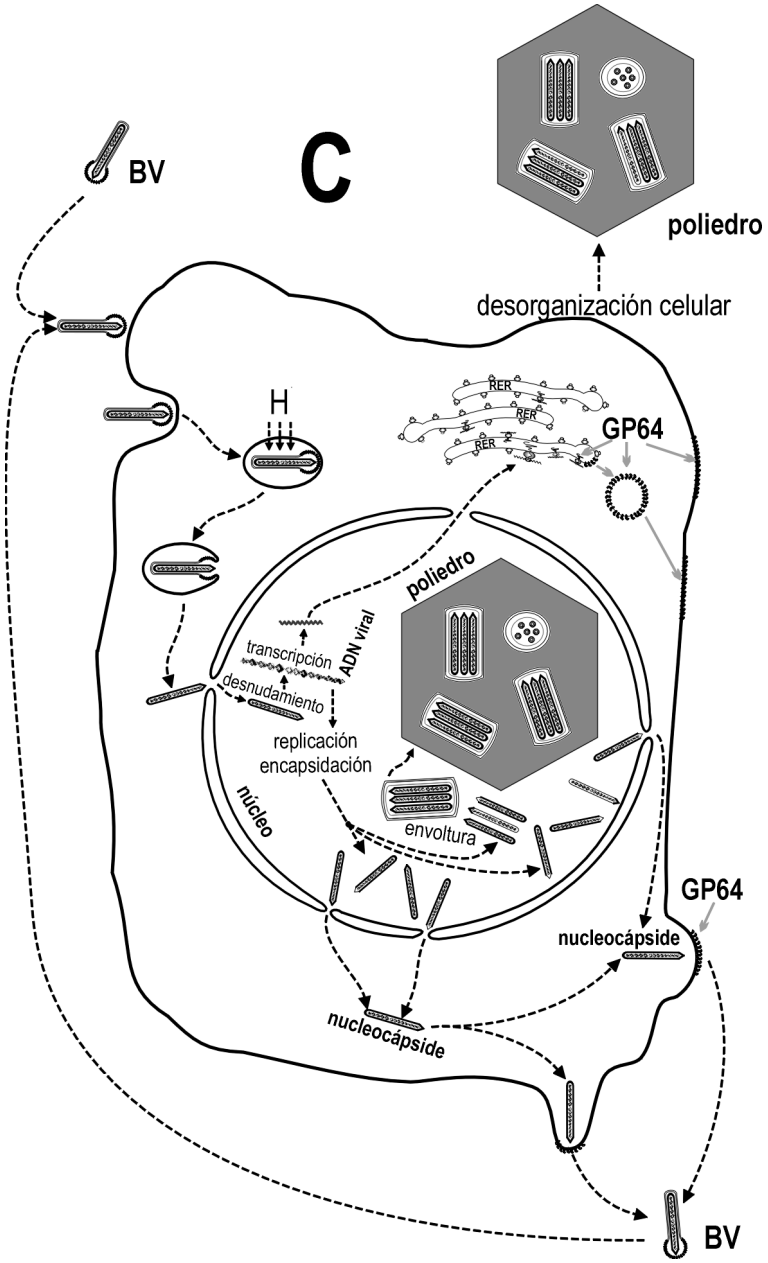

Figura A.4: Infección secundaria y producción de OVs.

Apéndice B

Sobre la aplicación

En este apéndice proveemos la información necesaria para la instalación, configuración y ejecución del clasificador construido en este trabajo.

B.1. Dependencias

Los requerimientos de la aplicación son:

- Python versión 2.6.6 o superior (de la línea 2.X)
- PyML* versión 0.7.2 o superior

La información necesaria para instalar Python y PyML puede encontrarse en *http://www.python.org/* y *http://pyml.sourceforge.net/* respectivamente.

B.2. Configuración

Previo a la ejecución de la aplicación, se deben configurar en el archivo `config.py` los siguientes parámetros:

- `proteinsDir`: directorio de polipéptidos a ser usados como casos positivos de entrenamiento
- `noProteinsDir`: directorio de polipéptidos a ser usados como casos negativos de entrenamiento
- `testProteinsDir`: directorio de polipéptidos a ser usados como casos positivos de testeo

*Tener en cuenta que PyML actualmente sólo funciona en plataformas Unix/Linux y Mac OS-X

- `testNoProteinsDir`: directorio de polipéptidos a ser usados como casos negativos de testeo
- `criteriaKernels`: conjunto de criterios y kernels a utilizar

Los archivos depositados en los directorios `proteinsDir`, `noProteinsDir`, `testProteinsDir` y `testNoProteinsDir` deben respetar el siguiente formato:

> *nombreDelPolipéptido1*
> *polipéptido1*
> *nombreDelPolipéptido2*
> *polipéptido2*
> ⋮
> *nombreDelPolipéptidoN*
> *polipéptidoN*

donde *nombreDelPolipéptidoI* es un identificador único de *polipéptidoI* pudiendo contener letras, números o cualquier otro caracter (distinto de `\n` o `\r`) y *polipéptidoI* es la secuencia de aminoácidos que representa a dicho polipéptido (denotados por su código de una letra).

B.3. Ejecución

Una vez finalizada la configuración en el archivo `config.py`, ya puede llevarse a cabo la clasificación de los polipéptidos. Para esto, debe ejecutarse el programa `main.py`. Este dejará en el directorio base los archivos que contienen el resultado de la clasificación de los polipéptidos según cada criterio y un archivo que contendrá todos los polipéptidos agrupados según la cantidad de criterios que lo clasificaron como positivo.

Bibliografía

Brooks, D.J., Fresco, J.R., Lesk, A.M. & Singh, M. (2002). Evolution of amino acid frequencies in proteins over deep time: Inferred order of introduction of amino acids into the genetic code. *Molecular Biology Evolution*, **19**, 1645–1655. 17

Caballero, P., López-Ferber, M. & Williams, T. (2001). *Los Baculovirus y sus Aplicaciones como Bioinsecticidas en el Control Biológico de Plagas*. Phytoma S.A. 48

Cornette, J.L., Cease, K.B., Margalit, H. & Spouge, J.L. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *Journal of Molecular Biology*, **195**, 659–685. 17, 34

Du, X., Cheng, J. & Song, J. (2009). Improved prediction of protein binding sites from sequences using genetic algorithm. *The Protein Journal*, **28**, 273–280. 17

Faulkner, P., Kuzio, J., Williamst, G.V. & Wilson, J.A. (1997). Analysis of P74, a PDV envelope protein of Autographa californica nucleopolyhedrovirus required for occlusion body infectivity in vivo. *Journal of General Virology*, **208**, 328–335. 50

Gatherer, D. & McEwan, N.R. (2003). Analysis of sequence periodicity in E. coli proteins: Empirical investigation of the "duplication and divergence" theory of protein evolution. *Journal of Molecular Evolution*, **57**, 149–158. 18

Granados, R.R. (1978). Early events in the infection of Heliothis zea midgut cells by a baculovirus. *Virology*, **90**, 170–174. 50

Granados, R.R. & Lawler, K.A. (1981). In vivo pathway of Autographa californica baculovirus invasion and infection. *Virology*, **108**, 297–308. 50

Heisele, B., Ho, P., Poggio, T., Ho, Y.P. & Poggio, P.T. (2001). Face recognition with support vector machines: Global versus component-based ap-

proach. In *Proceeding 8th International Conference on Computer Vision*, 688–694. 3

Hohmann, A.W. & Faulkner, P. (1983). Monoclonal antibodies to baculovirus structural proteins: determination of specificities by Western blot analysis. *Virology*, **125**, 432–444. 51

Huang, J., Blanz, V. & Heisele, B. (2002). Face recognition with support vector machines and 3D head models. In *Pattern Recognition with Support Vector Machines, First International Workshop, SVM 2002*, 334–341. 3

Ivanov, O.C. & Ivanov, C.P. (1980). Some evidence for the universality of structural periodicity in proteins. *Journal of Molecular Evolution*, **16**, 47–68. 18

Jaakkola, T., Diekhans, M. & Haussler, D. (1999). Using the Fisher kernel method to detect remote protein homologies. *American Association for Artificial Intelligence*. 3

Krause, L., McHardy, A.C., Nattkemper, T.W., Pühler, A. & Stoye, J. (2006). GISMO—gene identification using a support vector machine for ORF classification. *Nucleic Acids Research*, **35**, 540–549. 3

Landschulz, W.H., Johnson, P.F. & McKnight, S.L. (1988). The leucine zipper: a hypothetical structure common to a new class of DNA-binding proteins. *Science*, **240**, 1759–1764. 17

Lee, W.S. & Liu, B. (2003). Learning with positive and unlabeled examples using weighted logistic regression. *Proceedings of the Twentieth International Conference on Machine Learning*. 19

Liao, H., Yeh, W., Chiang, D., Jernigan, R. & Lustig, B. (2005). Protein sequence entropy is closely related to packing density and hydrophobicity. *Protein Engineering, Design & Selection*, **18**, 59–64. 17

Liao, L. & Noble, W.S. (2002). Combining pairwise sequence similarity and support vector machines for remote protein homology detection. In *Proceedings of the Sixth International Conference on Computational Molecular Biology*. 3

Monsma, S.A. & Blissard, G.W. (1995). Identification of a membrane fusion domain and an oligomerization domain in the baculovirus GP64 envelope fusion protein. *Journal of Virology*, **69**, 2583–2595. 51

Romanowski, V. & Ghiringhelli, P.D. (2001). Biología molecular de los baculovirus. replicación y regulación de la expresión génica. In P. Caballero, M. López-Ferber & T. Williams, eds., *Los Baculovirus y sus Aplicaciones como Bioinsecticidas en el Control Biológico de Plagas*, chap. 5, Phytoma S.A. 48

Schneider, T.D. & Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, **18**, 6097–6100. 28

Shannon, C.E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423,623–656. 17, 28

Urbina, D., Tang, B. & Higgs, P.G. (2006). The response of amino acid frequencies to directional mutation pressure in mitochondrial genome sequences is related to the physical properties of the amino acids and to the structure of the genetic code. *Journal of Molecular Evolution*, **62**, 340–361. 17

Vapnik, V.N. (1998). *Statistical Learning Theory*. Wiley-Interscience. 5

Vapnik, V.N. (2000). *The Nature of Statistical Learning Theory*. Springer, 2nd edn. 2, 3, 5, 9

Venkatesh, J. & Sureshkumar, C. (2009). Handwritten Tamil character recognition using SVM. *International Journal of Computer and Network Security*, **1**. 3

Vert, J.P. (2002). A tree kernel to analyze phylogenetic profiles. *Bioinformatics*, **18**, 276–284. 3

Wang, Y., Tian, Y. & Deng, N. (2008). Distinguishing enzymes from non-enzymes via support vector machine. *The Second International Symposium on Optimization and Systems Biology*, 166–173. 3

Wu, W., Chen, D. & Yang, J. (2005). Integrating co-training and recognition for text detection. In *IEEE International Conference on Multimedia & Expo*. 3

Zhu, B., Zhou, X.D., Liu, C.L. & Nakagawa, M. (2009). A robust model for on-line handwritten Japanese text recognition. *SPIE*, **7247**. 3