Reconocimiento de Patrones

Informe de trabajo final

Identificación de proteínas con capacidad fusogénica

Estudiante:
Daniela Megrian

Orientadores:
Federico Lecumberry
Ignacio Ramírez

8 de diciembre de 2014

Índice

1. Introducción

La fusión de bicapas lipídicas de membranas celulares es termodinámicamente favorable pero existe una barrera cinética muy alta. Así, las membranas celulares no fusionan espontáneamente sino que dicho proceso es catalizado por proteínas denominadas fusasas que median esta fusión, actuando de manera precisa y bajo estricto control espacial y temporal.

Este trabajo nos enfocaremos en la identificación de fusasas de tipo viral porque son las únicas maquinarias conocidas al momento que catalizan la fusión de membranas en el exterior celular. Además los escasos casos de proteínas identificadas con capacidad de fusionar células presentan homología con fusógenos virales. Una de estas es la proteína Eff de C. elegans. Esta proteína es estructuralmente homóloga a las proteínas de fusión viral de clase II, conservando también su organización en hojas β [1]. A pesar de esta homología, la secuencia aminoacídica difiere en gran medida de las secuencias de proteínas de fusión viral de clase II.

En humanos, se identificaron proteínas denominadas sincitinas, las cuales participan en la formación de la placenta. Son proteínas de origen retroviral y se comportan como fusógenos virales de clase I [2].

Entre los fusógenos pertenecientes a una misma clase existe una alta conservación de la estructura, pero una gran divergencia entre las secuencias aminoacídicas. Se han definido tres clases de proteínas de fusión de membranas virales basadas en características estructurales claves.

Se conoce que la organización de la estructura secundaria sigue determinados patrones según la clase. Aunque no se han publicado datos cuantitativos, se conoce la estructura secundaria de las proteínas de fusión de clase I es predominantemente de α-hélice, mientras que para las de clase II es de hoja β.

2. Objetivo

Clasificar un conjunto de proteínas como fusógenos virales en clase I o clase II a partir de su secuencia aminoacídica.

3. Enfoque

El trabajo consta de 2 etapas principales. La primera etapa se basa en la clasificación de fusógenos virales en clase I o clase II con el fin evaluar una métrica que funcionen bien para los datos. Se trabajó con un clasificador de tipo Support Vector Machine (SVM).
La segunda etapa consistió en clasificar un conjunto de fusógenos virales como clase I o clase II, entrenando con fusógenos virales de clase I. Se utilizó el método One-Class Support Vector Machine, a partir de la métrica determinada en la primera etapa.

4. Pre-procesamiento de los datos

Se obtuvieron las secuencias aminoacídicas de las proteínas precursoras de fusógenos virales depositadas en la base de datos pública UniProt. Se seleccionaron aquellas proteínas etiquetadas como fusógeno viral de clase I o fusógeno viral de clase II. Estas secuencias aminoacídicas fueron la información inicial a partir de la cual se trabajó según se esquematiza en la Figura 1. Se obtuvieron 27846 secuencias de clase I y 1800 de clase II, de largos variables de hasta 691 aminoácidos.
Las proteínas de fusión viral se sintetizan como precursores inactivos que en determinadas condiciones se clivan, liberando una proteína transmembrana con capacidad fusogénica. Se extrajo la región fusogénica de cada proteína a partir de las anotaciones presentes en UniProt, y de aquí en adelante se trabajó solamente con este fragmento.
Se tenía conocimiento de la redundancia de secuencias en UniProt. Esto se observa particularmente para la proteína Hemaglutinina, fusógeno de clase I del virus de la Influenza, de la cual se encuentran depositadas secuencias de miles de cepas. Por esta razón, como paso previo a la predicción de estructuras secundarias se clusterizaron las secuencias al 99 % de identidad con la herramienta CD-HIT [3]. CD-HIT es ampliamente utilizado para el manejo de grandes bases de datos de proteínas como Uniprot y PDB. El algoritmo que implementa CD-HIT primero ordena las secuencias según su largo, en orden decreciente. La secuencia más larga corresponde a la secuencia representativa del primer cluster. Cada una de las secuencias restantes se compara aminoácido a aminoácido con las secuencias declaradas como

representativas. Si la identidad con una secuencia representativa de cluster supera un valor umbral, la secuencia en cuestión se incluye en el cluster. Para reducir el número de comparaciones se utilizan "short word filters". Dos proteínas que comparten determinada identidad de secuencia deben tener al menos cierta cantidad de péptidos (cadena de aminoácidos corta) de longitud fija idénticos. Por ejemplo, dos proteínas que tienen un 85 % de aminoácidos idénticos en una ventana de 100 aminoácidos, deben tener al menos 70 dipéptidos idénticos, 55 tripéptidos idénticos y 25 pentapéptidos idénticos. Los pares de secuencias que no satisfacen estas condiciones no se comparan. Aplicando este algoritmo se obtuvieron 1769 secuencias representativas de fusógenos virales de clase I y 1103 de clase II.

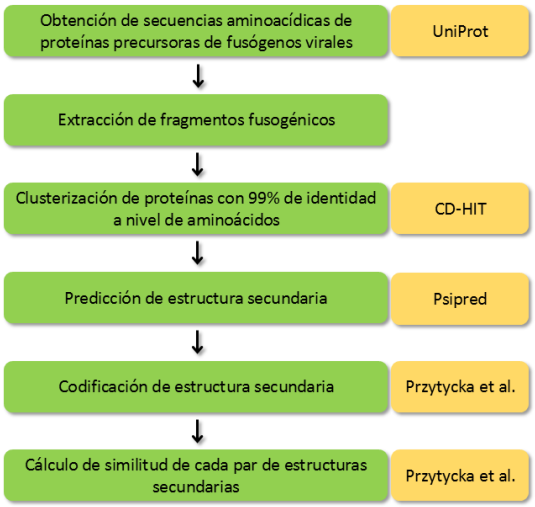

Figura 1 Esquema de etapas del pre-procesamiento de los datos. En amarillo se indica la herramienta que se utilizó para cada etapa

Dado que el tiempo computacional es limitante, se calcularon las predicciones de estructura secundaria solamente para una proteína representativa de cada cluster. Esto se realizó con la herramienta Psipred [4], a través del paquete HHSuite [5]. HHSuite incorpora información de secuencias de proteínas homólogas a la proteína de interés, generando predicciones más precisas. El software genera una matriz de score posición específica de 20 x M elementos, donde 20 son los distintos aminoácidos y M es el largo de la secuencia. Cada elemento de la matriz representa el logaritmo de la probabilidad de que suceda una sustitución de un aminoácido por

otro en esa posición. Esta matriz es el input del algoritmo back-propagation de dos redes neuronal feed-forward con una capa oculta.
Para cada posición aminoacídica se obtiene un caracter H, E o C, correspondiente a la estructura más probable para esa posición. H corresponde a hélice α, E a hoja β y C a loops o estructuras desorganizadas.

5. Desarrollo de una métrica de similitud

Para medir la similitud entre dos proteínas nos basamos en el método descripto por Przytycka et al. [6]. La predicción de estructura secundaria realizada por Psipred de cada fusógeno viral se representa como una secuencia resumida y ordenada de los elementos de su estructura secundaria (H, E y C). Los caracteres repetidos consecutivos, se colapsan y se almacena la longitud del elemento.
Como ejemplo, la secuencia aminoacídica del precursor del fusógeno viral Hemaglutinina del virus de la Influenza A es la siguiente:

MNTQILVFALVAIIPTNADKICLGHHAVSKGTKVNTLTERGVEVVNATETVE
RTNIPRICSKGKRTVDLGQCGLLGTITGPPQCDQFLEFSADLIIERQEGSDVC
YPGKFVNGEALRQILRESGGIDKETMGFTYSGIRTNGATSACRRSGSSFYAE
MKWLLSNTDNAAFPQTTKSYKNTRKDPALIIWGIHHSGSTTEQTKLYGSGN
KLITVGSSNYQQSFVPSPGARPQVNGQSGRIDFHWLMLNPNDTVTFSFNGA
FIAPDRASFLRGKSMGIQSGVQVDANCEGNCYHNGGTIISNLPFQNINSRAVG
KCPRYVKQESLLLATGMKNVPEIPKGR**GLFGAIAGFIENGWEGLIDGW**
YGFRHQNAQGEGTAADYKSTQSAIDQITGKLNQLIEKTNQQFGLI
DNEFTEVEKQIGNVINWTRDSMTEVWSYNAELLVAMENQHTIDL
ADSEMNKLYERVRRQLRENAEEDGTGCFEIFHKCDDDCMASIRN
NTYDHSKYREEAMQNRIQIDPVKLSSGYRDVILWFSFGASCFILLA
IAMGLVFICVKNGNMRCTICI

Mientras que la región fusógenica, denominada HA2, es la porción carboxiloterminal de la proteína precursora que está marcada en negritas.
La predicción de la estructura secundaria del fusógeno según Psipred es:

CCCHHHHHHHHCCCCCCCCCEECCEEECCCCCEECCCHHHHHHHHHHH

HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCHHHHHHHHHH
HHHHHHCCCCCCCCHHHHHHHHHHHHHHHHHHHCCCCCCCCEEEEEECCCH
HHHHHHHHCCCCCHHHHHHHHCCCCCCCCCEECCCCCEEEEEEEHHHHHH
HHHHHHHHHHHHHHHHCCCCEEEEEEC

Entonces, la representación de la estructura secundaria a partir de la cual se calculará la métrica es:

C, H, C, E, C, E, C, E, C, H, C ,H , C, H, C, E, C, H, C, H, C, E, C, E, H, C, E, C

Y las longitudes de cada elemento son:

3, 8, 9, 2, 2, 3, 5, 2, 3, 48, 2, 16, 8, 17, 8, 6, 3, 9, 5, 8, 8, 2, 5, 7, 21, 4, 6, 1

Para calcular la métrica de similitud realizamos alineamientos de las representaciones de las estructuras secundarias de a pares, a partir de un algoritmo de progamación dinámica. El algoritmo de programó en Python.
El algoritmo que proponen Przytycka et al. [6] es análogo al de Needleman-Wunsch [7], pero utilizando otro sistema de scores. Para alinear 2 elementos X e Y, el score se define como S(X,Y):

S(X,Y):	si X = Y:	min(L(X),L(Y))
	si X = H o E e Y = C; o X = C e Y = H o E:	0.5*min(L(X),L(Y))
	si X = H e Y = E; o X = E e Y = H:	0

Siendo L(X) y L(Y) el largo del elemento X e Y, respectivamente.

El score de similitud final se normaliza por el promedio los largos de las representaciones de la estructura secundaria. Este score final (d(X,Y)) vale entre 0 y 1, y cuanto mayor es el valor mayor es la similitud entre esas 2 proteínas según su estructura secundaria.

Przytycka et al. proponen la realización de alineamientos globales, pero nosotros implementamos la técnica para alineamientos locales (algoritmo análogo al de Smith-Waterman [8]) dado que nos interesa identificar posibles fragmentos fusogénicos de proteínas precursoras. Fontana et al. [9] proponen esta técnica, pero no la aplican. No hemos identificado otros artículos que lo utilicen.
Se han publicado diversos artículos aplicando la métrica propuesta por Przytycka. Particularme resulta interesante un trabajo reciente [10] , basado en la discriminación de proteínas de membrana externa utilizando SVM. El trabajo que proponen los autores presenta puntos en común con este trabajo y obtuvieron resultados satisfactorios.
Se obtuvo una matriz de similitud de fusógenos de 2872 x 2872 (1769 secuencias de clase I y 1103 de clase II), en la cual se ven representados los scores de similitud entre cada par de fusógenos virales seleccionados previamente (Figura 2).

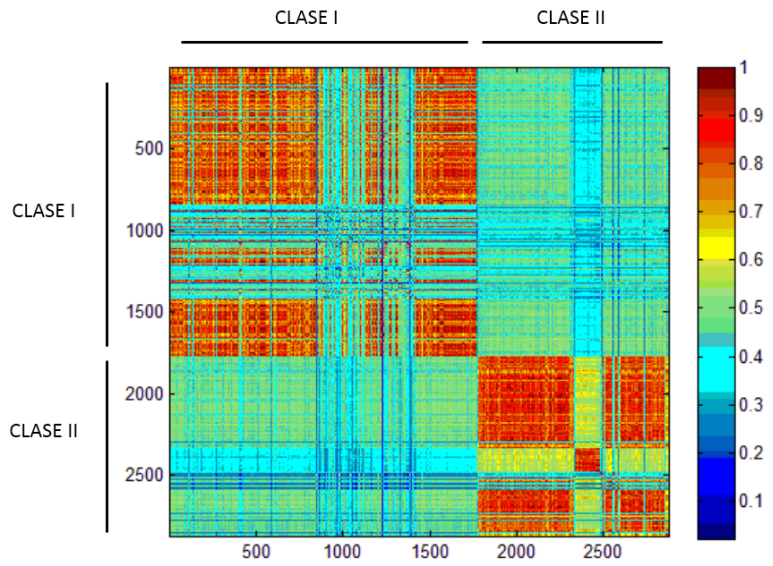

Figura 2 Mapa de color de la matriz de similitud entre fusógenos virales. Cuanto más similares son dos proteínas, más cerca de 1 se encuentra el score. Los fusógenos dentro de una misma clase presentan en general scores altos. Esto se observa en el recuadro predominantemente rojo en la esquina superior izquierda correspondiente a los fusógenos de clase I enfrentados contra sí mismos. Lo mismo se observa en la esquina inferior derecha para los fusógenos de clase II.

6. Entrenamiento y clasificación

6.1. C-SVM

El paso siguiente fue aplicar la métrica para clasificar fusógenos virales de clase I y de clase II. El método Support Vector Machines (SVM) ha sido ampliamente utilizado en el análisis de secuencias biológicas. Una particularidad del método es el uso de funciones kernel, que proyectan el problema en un espacio de características de mayor dimensión (incluso infinito). Esto permite la búsqueda del hiperplano que proporciona la máxima separación entre dos clases en el espacio transformado. En este trabajo transformamos las distancias con el kernel:
k(x,y) = exp(γ * d(X,Y))
Se trabajó con el paquete LIBSVM [11] para Python. LIBSVM permite generar un clasificador a partir de un kernel precalculado y estimar su desempeño. El desempeño del clasificador depende de los parámetros C y γ. El parámetro C aporta cierta flexibilidad en la clasificación, permite algunos errores pero también los penaliza. El parámetro γ define qué tan lejos llega la influencia de una muestra. La mejor combinación de C y γ se seleccionó a partir de una búsqueda "grid search" (Figura 3) utilizando "10-fold cross-validation" con:

$C \in \{2^{-15}, 2^{-13}, 2^{-11}, 2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}, 2^{7}, 2^{9}, 2^{11}, 2^{13}, 2^{15}\}$

$\gamma \in \{5.4\text{x}10^{-4}, 5.4\text{x}10^{-3}, 5.4\text{x}10^{-2}, 5.4\text{x}10^{-1}, 5.4, 54\}$

Se seleccionaron como parámetros óptimos $C = 2^{-15}$ y $\gamma = 5.4\text{x}10^{-2}$. Para estos parámetros la medida de "accuracy" de la clasificación de fusógenos virales de clase I y clase II fue del 100 %.

6.2. One-class SVM

Los clasificadores de tipo SVM clásicos están basados en el entrenamiento a partir de muestras de dos clases (por ejemplo positivas y negativas). Sin embargo, en diversas situaciones se dispone solamente de muestras positivas para el entrenamiento. Este es el caso del problema planteado originalmente por este proyecto, donde se dispone de un set de entrenamiento de proteínas que se conoce que son

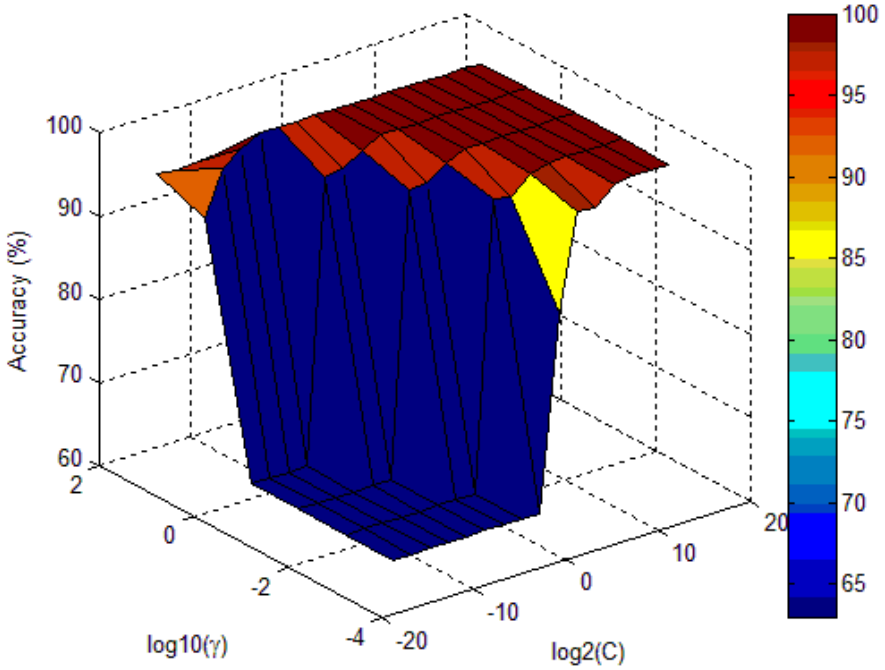

Figura 3 Gráfico del "accuracy"de la clasificación de fusógenos virales obtenido a partir de una búsqueda "grid search"para los parámetros C y γ. Se trabajó con el método SVM clásico.

fusogénicas, y se pretende seleccionar proteínas candidatas a fusógenos de otro set de proteínas muy diverso. Dada la variabilidad del set de proteínas candidatas, resulta complejo identificar muestras negativas representativas. Scholkopf et al. [12] describen el método "one-class classification" que permite entrenar el modelo con solamente muestras positivas.

La última etapa de este trabajo fue entrenar un clasificador con fusógenos virales del clase I como muestras positivas. Se evaluó la clasificación de un conjunto de fusógenos virales de clase I (distinto al conjunto de entrenamiento) y un conjunto de fusógenos de clase II. Para esto se aplicó el método one-class SVM.

Para esta etapa también se trabajó con el paquete LIBSVM que dispone de esta metodología. Se aplicó el mismo kernel a los datos y se seleccionó la mejor combi-

nación de parámetros. En este caso el parámetro ν sustituye al parámetro C. El significado del parámetro ν es análogo al significado de C, pero solamente puede tomar valores entre 0 y 1.
Análogamente a la parte anterior la mejor combinación de ν y γ se seleccionó a partir de una búsqueda "grid search" utilizando "10-fold cross-validation" (Figura 4) con:

$\nu \in \{0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1\}$

$\gamma \in \{5.4\text{x}10^{-4}, 5.4\text{x}10^{-3}, 5.4\text{x}10^{-2}, 5.4\text{x}10^{-1}, 5.4, 54\}$

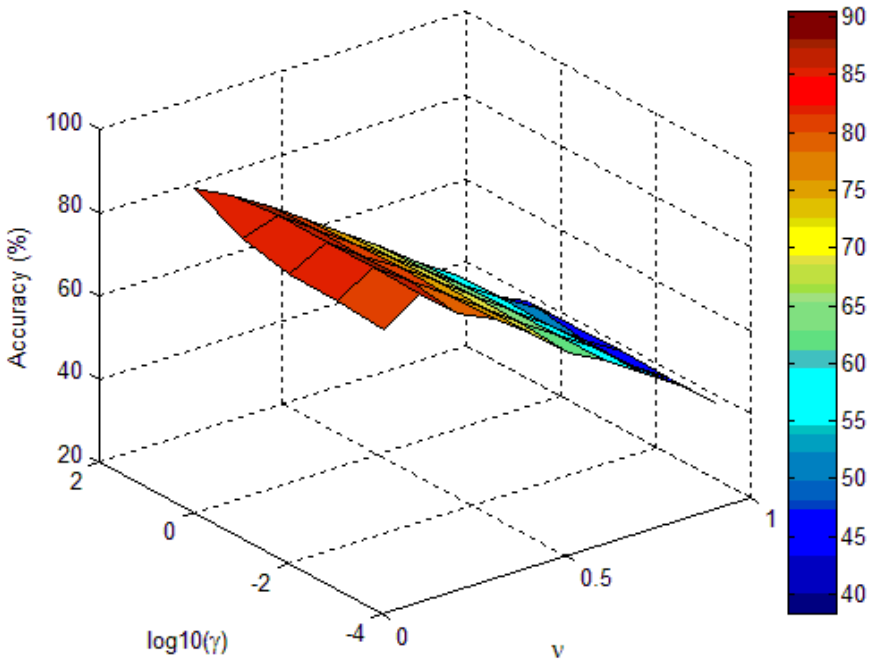

Figura 4 Gráfico del "accuracy"de la clasificación de fusógenos virales obtenido a partir de una búsqueda "grid search"para los parámetros ν y γ. Se trabajó con el método One-class SVM.

Se realizó un nuevo "grid search" para un rango más pequeño de ν:

$\nu \in \{0.02, 0.04, 0.06, 0.07, 0.1, 0.12, 0.14, 0.16, 0.18, 0.2, 0.22, 0.24, 0.26, 0.28, 0.3\}$

Se seleccionaron como parámetros óptimos $\nu = 0.12$ y $\gamma = 5.4\text{x}10^{-1}$. Para estos parámetros la medida de "accuracy" de la clasificación de fusógenos virales de clase I y clase II fue del 91.3 %.

7. Conclusiones y perspectivas

En este trabajo se implementó una nueva medida de similitud, basada en el alineamiento de estructuras secundarias propuesto por Przytycka et al. y el algoritmo alineamientos locales propuesto por Needleman et al. Esta medida de similitud permitió clasificar fusógenos virales de clase I y clase II con alta precisión. A pesar de que en la Figura 1 se puede ver algunas secuencias que parecen divergentes dentro de una misma clase (color cyan), la clasificación utilizando el método SVM clásico resultó en un "accuracy"del 100 %. El método One-class SVM no permitió discriminar de forma totalmente precisa fusógenos de clase I y de clase II. Sin embargo el "accuracy"del 91.3 % resulta prometedor para ir más allá de este problema e intentar clasificar un conjunto de proteínas en fusógenos y no-fusogenos.
El método One-class SVM resulta más apropiado que el método SVM clásico para resolver el problema biológico real. Es de interés identificar proteínas candidatas a ser fusógenos virales dentro de la enorme variedad de proteínas que existe en modelos biológicos. Por esta razón resulta poco razonable definir un conjunto de proteínas que definan el universo de los no-fusógenos.
Resulta interesante examinar en detalle la diversidad dentro de una misma clase, lo cual puede permitir mejorar la clasificación. Próximamente analizaremos en detalle cómo está definida cada clase filogenéticamente, lo cual creemos que puede llegar a explicar la divergencia. También nos interesa analizar si las proteínas mal clasificadas por el método One-class SVM corresponden a fusógenos de clase I clasificados como clase II o viceversa.
Adicionalmente, intentaremos clasificar un conjunto de proteínas virales que contiene fusógenos y no-fusógenos con el fin de identificar las proteínas fusogénicas. Este trabajo forma parte de un proyecto en el cual se pretende identificar proteínas candidatas a ser fusógenos en distintos modelos biológicos donde, si bien se sabe que existe una proteína fusogénica, aún no se ha identificado.

Referencias

[1] Perez-Vargas, J., T. Krey, C. Valansi, O. Avinoam, A. Haouz, M. Jamin, H. Raveh-Barak, B. Podbilewicz and F. A. Rey. (2014). Structural basis of eukaryotic cell-cell fusion. *Cell*, 157(2):407–419.

[2] Mi, S., X. Lee, X. Li, G. M. Veldman, H. Finnerty, L. Racie, E. LaVallie, X. Y. Tang, P. Edouard, S. Howes, J. C. Keith, Jr. and J. M. McCoy. (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, 403(6771):785–789.

[3] Li, W. and A. Godzik. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

[4] McGuffin, L. J., K. Bryson and D. T. Jones. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, 16(4):404–405.

[5] Soding, J., A. Biegert and A. N. Lupas. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res*, 33:W244-248.

[6] Przytycka, T., R. Aurora and G. D. Rose. (1999). A protein taxonomy based on secondary structure. *Nat Struct Biol*, 6(7):672–682.

[7] Needleman, S. B. and C. D. Wunsch. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453.

[8] Smith, T. F. and M. S. Waterman. (1981). Identification of common molecular subsequences. *Mol Biol*, 147(1):195–197.

[9] Fontana, P., E. Bindewald, S. Toppo, R. Velasco, G. Valle and S. C. Tosatto. (2005). The SSEA server for protein secondary structure alignment. *Bioinformatics*, 21(3):393–395.

[10] Ni, Q. and L. Zou. (2014). Accurate discrimination of outer membrane proteins using secondary structure element alignment and support vector machine. *J Bioinform Comput Biol*, 12(1):1450003.

[11] Chih-Chung Chang and Chih-Jen Lin. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 27:1–27:27.

[12] Scholkopf, B., J. C. Platt, J. Shawe-Taylor, A. J. Smola and R. C. Williamson. (2001). Estimating the support of a high-dimensional distribution. *Neural Comput*, 13(7):1443–1471.