El Problema de la Predicción de Sitios de Clivaje

Gloria Inés Alvarez V., Jorge Hernán Victoria M.

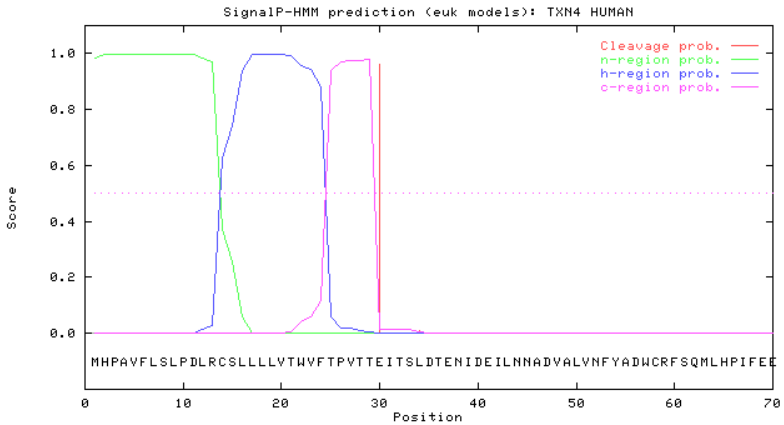
Proyecto *Técnicas de Inferencia Gramatical y Aplicación al Procesamiento de Biosecuencias*
Grupos de Investigación TLCC y Destino
Facultad de Ingeniería
Universidad Politécnica de Valencia
Pontificia Universidad Javeriana Cali

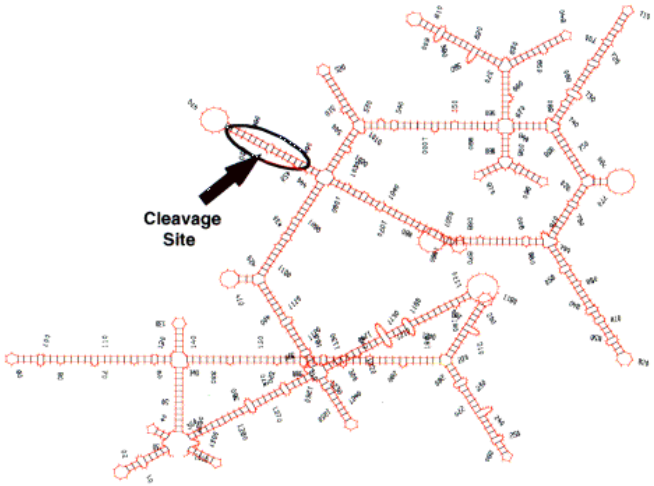
Octubre de 2009

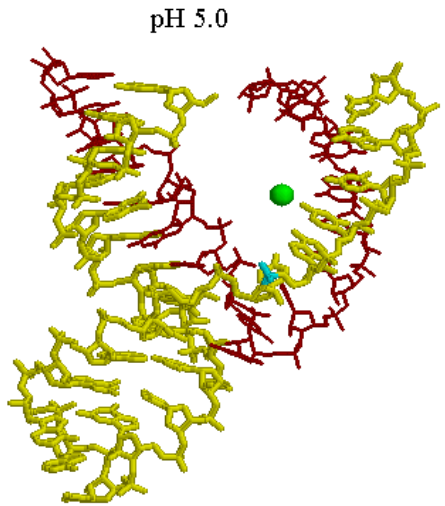
Contenido de la Presentación

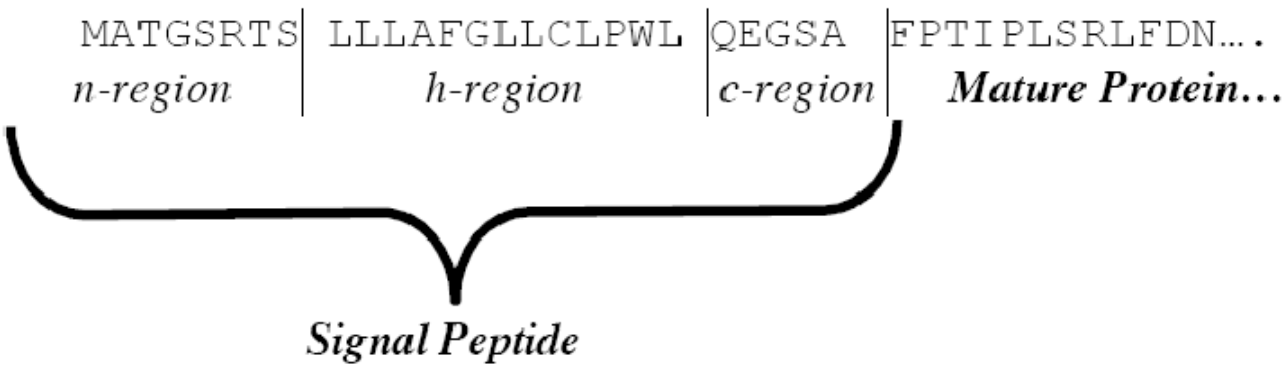
La Predicción de Sitios de Clivaje

Consiste en detectar el sitio exacto dentro de una cadena de aminoácidos donde comienza la traducción de una proteina funcional específica.



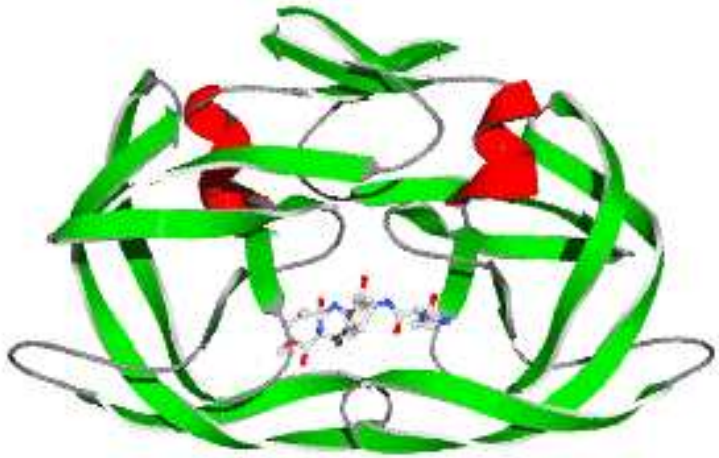





MATGSRTS
LLLAFGLLCLPWL
QEGSA
FPTIPLSRLFDN….
n-region
h-region
c-region
Mature Protein…
Signal Peptide

Aplicaciones

Conocer el sitio de clivaje de una proteina de un virus, permite construir drogas que eviten la traducción de dicha proteina, inhibiendo su producción y por lo tanto también sus efectos. Por ejemplo, la proteasa HIV-1 participa en la replicación del virus de HIV, inhibirla evitaría que el virus se propague.

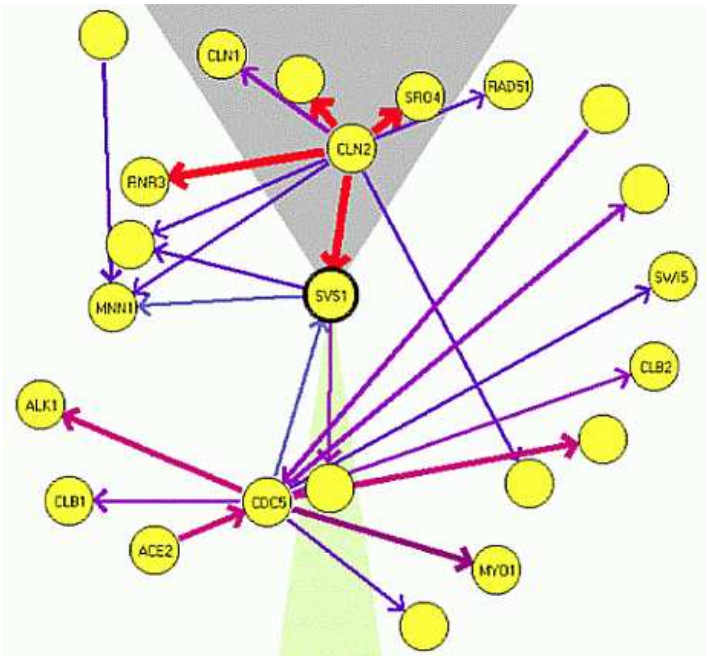


Soluciones Propuestas

- Modelos utilizados:
  - Redes Bayesianas.
  - Matrices de pesos por posición.
  - Redes Neuronales.
  - Máquinas de soporte vectorial.
  - Modelos Ocultos de Markov.
  - K-vecinos más cercanos.
  - Perceptrón simple.
  - Máquina de soporte vectorial lineal.
  - Técnicas de comité de expertos: votación, sistemas en cascada o jerárquicos.
- Medidas de error.

Redes Bayesianas

- Tiene desempeño comparable a las redes neuronales[a].
- Pero con la ventaja que se pueden obtener explicaciones para los resultados.
- A partir de una base de datos se calcula la frecuencia relativa de ciertos patrones y se usan esos datos como probabilidades a posteriori.
- Algunos patrones son absolutos y otros relativos en cuanto a su posición en la secuencia.



[a] Bayesian Sequence Learning for Predicting Protein Cleavage Points. Michael Mayo,University of Waikato, New Zeland. ????

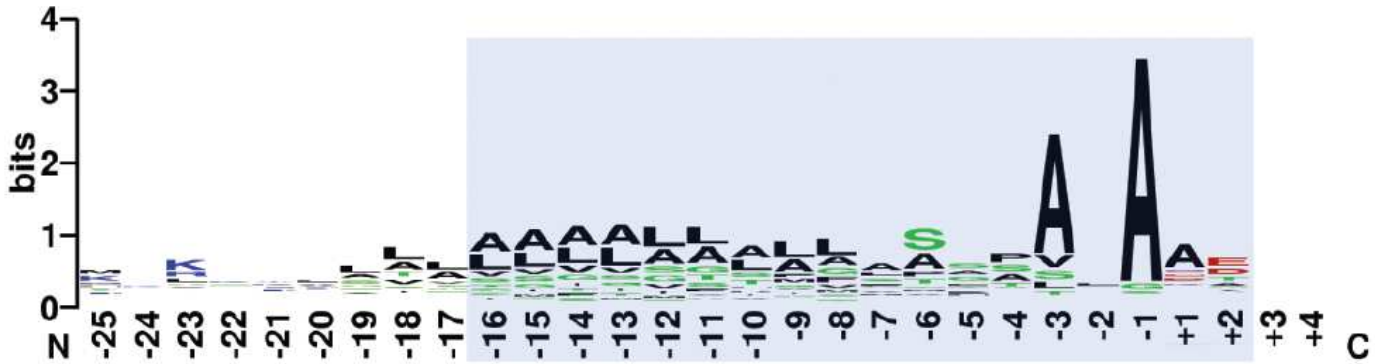
Matriz de Pesos

- A partir del alineamiento de secuencias etiquetadas, se generaron matrices de frecuencia[1].
- Se crearon tres matrices de frecuencia diferentes, para eukariotas, baterias Gram-positivas y Gram-negativas.
- Las matrices de pesos se basan en la frecuencia de algunas subsecuencias además de cuatro aminoácidos en la región N.
- Resultados levemente inferiores a SignalP2.0

[1] PrediSi:prediction of signal peptides and their cleavage positions. K. Hiller, A. Grote,M. Scheer, R. Munch, D Jahn. Nucleic acids Research. Vol 32. 2004.

Matriz de Pesos

Resultados obtenidos para bacterias Gram-positivas en la correspondiente matriz de pesos por posición.



Redes Neuronales

Principalmente se ha utilizado:

- Perceptrón multicapa.
- Máquinas de Soporte Vectorial.

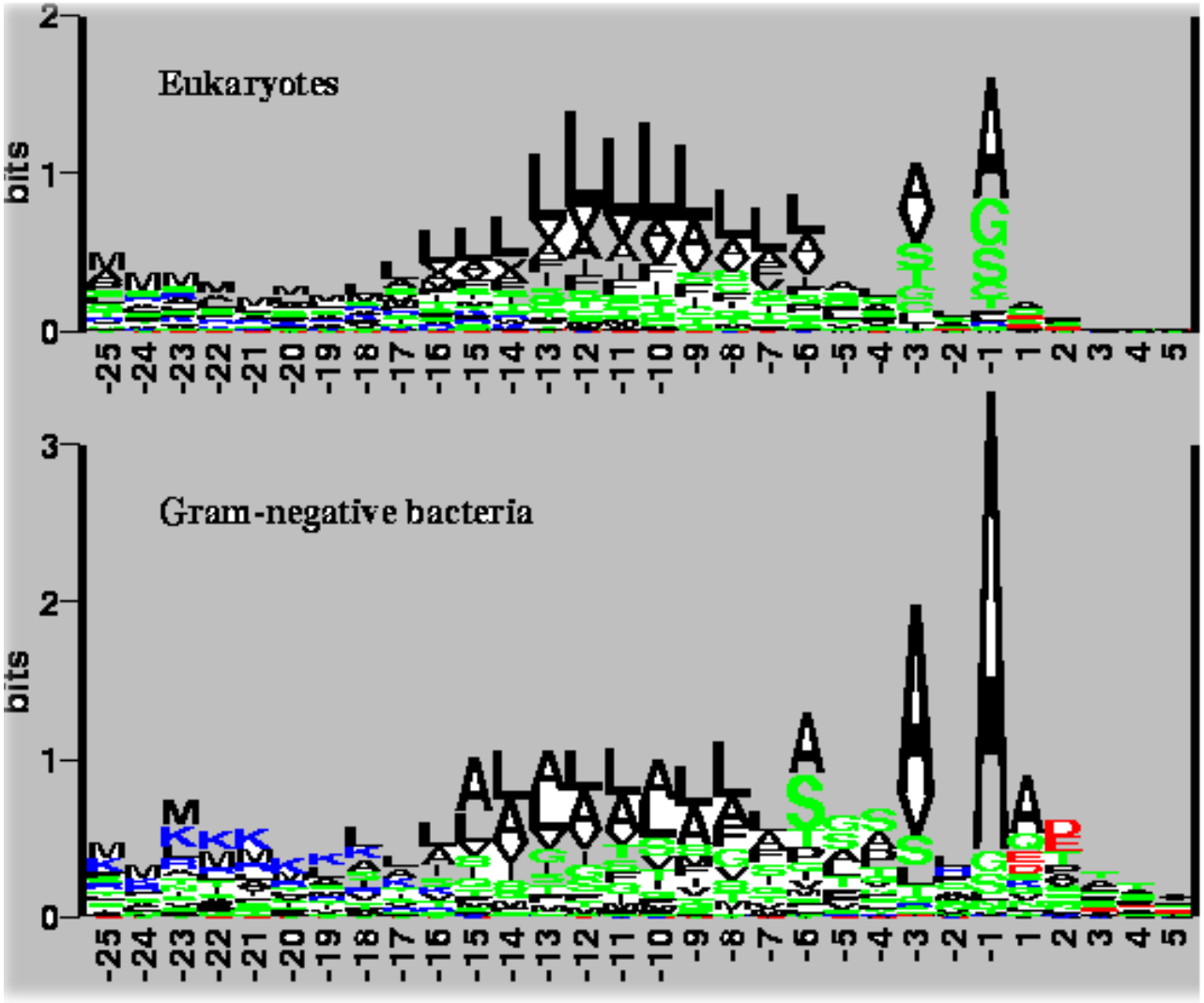
La herramienta más utilizada y de mejores prestaciones en el momento es SignalP 3.0[2], la cual usa perceptrones multicapa.

[2] Improved prediction of signal peptides: SignalP 3.0. J. Bendtsen, H. Nielsen, G.von Heijne, S. Brunak.Journal on Molecular Biology. vol 340. 2004.

SignalP 3.0

- Esta herramienta combina uso de perceptrones multicapa y modelos ocultos de Markov.
- Se puede usar para discriminar si en una secuencia existe algún sitio de clivaje o no y también para predecir el punto exacto del sitio, si lo hay.
- La versión 3.0 se ha mejorado tomando en consideración más información biológica.
  - Depuración de la base de datos de entrenamiento.
  - Afinamiento del tamaño de la ventana deslizante.
  - Adición de nuevas entradas a la red neuronal que indican la posición de la ventana deslizante en la secuencia y la composición de aminoácidos de la secuencia completa.

Resultados obtenidos con SignalP 3.0



Reporte Negativo sobre el uso de Redes Neuronales

Para el caso de la proteasa HIV-1, se ha encontrado que la base de datos disponible es linealmente separable[3], por lo que se desaconseja el uso de redes neuronales y otros métodos de separación no lineal del espacio de búsqueda, al considerarlos innecesariamente complejos. Este estudio logró resultados similares usando un perceptrón simple o máquinas de soporte vectorial lineales.

[3] Why neural networks should not be used for HIV-1 protease cleavage site prediction. T. Rognvaldsson, L You. Bioinformatics. Vol 20. No 11. 2004

Perceptrón Simple - Máquina de soporte vectorial lineal

Al trabajar con estos modelos lineales[4], se han explorado formas de representación de la secuencia y técnicas de comité de expertos para mejorar su desempeño:

- Codificación ortonormal.
- 2-gramas (parejas aminoacido, frecuencia).
- BLOSUM50 (basado en la matriz de sustitución que lleva ese nombre junto con información del orden de la composición de los aminoácidos).
- Vector de momentos de composición (incluye información de la composición y posición de los aminoácidos en la secuencia).

[4] Comparison among feature extraction methods for HIV-1 protease cleavage site prediction. L. Nanni. Pattern Recognition. Vol 39. 2006.

Herramientas Conocidas

Es importante aclarar que estas herramientas han sido entrenadas para predecir los sitios de clivaje de diversas proteinas en diversas especies de individuos, por lo que no son ellas necesariamente comparables.

- SignalP 3.0: redes neuronales y HMM.
- SigCleave, SPScan y PrediSi: enfoque de matriz de pesos de predicción.
- SigFind, NNPSL: redes neuronales.
- PSORTB, SPEPlip, Phobius.

Medidas de Desempeño

Las medidas más comunes para evaluar el desempeño de un programa de predicción de sitios de clivaje son:

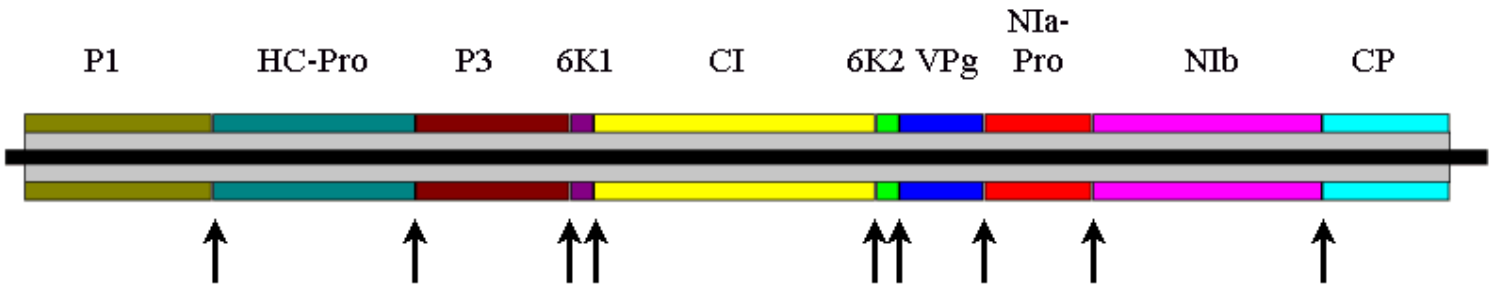
- Sensibilidad: $\frac{tp}{tp+fn}$
- Especificidad: $\frac{tp}{tp+fp}$
- Exactitud: $\frac{tp+tn}{tp+fp+tn+fn}$
- Coeficiente de correlación: $\frac{(tp*tn)-(fp*fn)}{\sqrt{(tp+fn)(tp+fp)(tn+fp)(tn+fn)}}$

Nuestro Problema de Predicción de Sitios de Clivaje

- La familia Potyviridae comprende virus de plantas entre los cuales está el mosaico del fríjol y de otras plantas como el tabaco, la sandía, etc.
- Los puntos de clivaje son los sitios en la poliproteina obtenida a partir del genoma del virus en los que inician y terminan los segmentos que dan origen a las proteinas funcionales.
- El problema de predicción de sitios de clivaje consiste en determinar la posición de dichos sitios sobre una cadena de aminoacidos.
- Se desea aplicar los algoritmos HyRPNI y OIL a resolver el problema de predicción de sitios de clivaje en secuencias correspondientes a virus de la familia Potyviridae.

Predicción de Sitios de Clivaje [5]

Mapa de la estructura de un miembro típico de la familia Potyviridae.



[5] Tomado de `www.dpvweb.net/potycleavage/index.html`

Porqué aplicar Inferencia Gramatical?

- Al ser un problema que ha intentado resolverse por varias técnicas, es útil para poder evaluar el desempeño de los algoritmos de inferencia gramatical.
- La inferencia gramatical no presupone el principio de independencia.
- La inferencia gramatical no requiere voluminosas representaciones de la secuencia de entrada.

Algoritmos Utilizados

- HyRPNI
  - Infiere DFAs.
  - Muy buen desempeño en tiempo y espacio.
  - Se debe fijar el tamaño de la primera fase del proceso de inferencia.
- OIL
  - Infiere NFAs.
  - Es un algoritmo no determinista.
  - Muy buen desempeño en espacio, requiere construir varias hipótesis de inferencia.
  - Se debe fijar el número de hipótesis a generar para el proceso de votación.

Cómo se va a solucionar el problema de predicción de sitios de clivaje

- Se construye una ventana deslizante que se mueve sobre la secuencia.
- Cada ventana alimenta un autómata previamente aprendido mediante inferencia gramatical que reconoce las cadenas que corresponden al sitio de clivaje del primer segmento.
- Cuando se detecta la presencia del primer sitio de clivaje, se reubica la ventana al comienzo del siguiente segmento y se empieza a procesar con el modelo del segundo sitio de clivaje y así sucesivamente.

Variables a considerar:

- Longitud de la ventana deslizante.
- Ubicación del sitio de clivaje dentro de la ventana.

Resultados Obtenidos HyRPNI, primer punto de clivaje

Table 1: Recognition rates of the Hy-RPNI. Cleavage P1-HCPro

Window Size	Group	Phase Size	Rate	%	Automaton Size	Time
4_1	a	5	151	0.974	5	0:04.43
4_1	a	10	152	0.980	3	0:13.36
4_1	a	25	152	0.980	4	0:26.26
4_1	b	5	154	0.993	7	0:06.00
4_1	b	10	152	0.980	5	0:13.34
4_1	b	25	151	0.974	6	0:27.39
4_1	c	5	155	1.0	6	0:07.13
4_1	c	10	154	0.993	5	0:12.03
4_1	c	25	154	0.993	6	0:29.59
4_1	d	5	150	0.967	5	0:07.18
4_1	d	10	154	0.993	4	0:17.79
4_1	d	25	153	0.987	5	0:37.83

Resultados Obtenidos OIL, primer punto de clivaje

Table 1: Recognition rates of OIL. Cleavage 1-P1-HCPro

Window Size	Group	Rate	%	Automaton Size	Time
4_1	a	151	0.974	5	0:01.04
4_1	b	151	0.974	5	0:02.13
4_1	c	154	0.993	5	0:03.56
4_1	d	151	0.974	4	0:03.63
Average		151.75	0.97875	4.75	0:2.59
Total					0:10.36
14_1	a	149	0.961	10	0:14.25
14_1	b	148	0.954	10	0:37.18
14_1	c	147	0.948	10	0:55.25
14_1	d	150	0.967	9	0:41.39
Average		148.5	0.9575	9.75	0:37.01
Total					2:28.07
10_10	a	148	0.954	9	0:51.29
10_10	b	144	0.929	13	1:02.66
10_10	c	151	0.974	8	1:48.84
10_10	d	150	0.967	7	0:53.71
Average		148.25	0.956	9.25	1:9.12
Total					4:36.5

Tareas actuales

- Implementar las medidas estandar de desempeño para nuestros resultados.
- Usar los datos de SignalP 2.0 para poder comparar nuestros algoritmos con otros, ya que en el problema de los potyvirus esto no parece posible.
- Depurar nuestras bases de datos para hacer más confiables los resultados obtenidos e intentar ejecutar otros programas de predicción de sitios de clivaje sobre ellos.