CURSO: CC471

Practica 11

MODELOS OCULTOS DE MARKOV PARA EL MODELAMIENTO DE ALINEAMIENTOS MULTIPLES DE SECUENCIAS.

Objetivo general
Aplicar técnicas heurísticas basadas en modelos estadísticos para efectuar el alineamiento múltiple de secuencias.
Objetivos específicos
Hacer una búsqueda en una base de datos de dominios conservados y obtener un alineamiento del alta calidad que pueda ser empleado como patrón o perfil.
Realizar un alineamiento múltiple empleando Modelos Ocultos de Markov guiados por patrones.
Utilizar Clustal Omega para realizar un alineamiento múltiple de secuencias guiado por patrones.

INTRODUCCIÓN
Cuando se puede disponer de un alineamiento múltiple de secuencias de calidad razonable, este puede emplearse como patrón o modelo para alinear otras secuencias pertenecientes a la misma familia. Para esto pueden emplearse métodos estadísticos, los cuales son algoritmos heurísticos que permiten construir perfiles o matrices (“profiles”) a partir de las frecuencias de aparición de todos los aminoácidos por cada columna del alineamiento. Dichos perfiles se pueden utilizar posteriormente para evaluar la probabilidad de que una nueva secuencia pueda ser incluida en el alineamiento, lo que puede ser aprovechado para alinear más secuencias o para realizar búsquedas altamente sensibles de otros homólogos en bases de datos.

Los perfiles de cada alineamiento pueden construirse automáticamente a partir de las secuencias no alineadas o bien puede derivarse de un alineamiento previo.

Dentro de los métodos estadísticos empleados para la representación de alineamientos múltiples se pueden citar las matrices de sitios de posición específica (Position-Specific Site Matrix, PSSM), como las que utiliza el programa PSI-Blast y los Modelos Ocultos de Markov (Hidden Markov Models, HMM). Los Modelos Ocultos de Markov son modelos estadísticos para la representación de procesos aleatorios basados en cadenas de Markov. Las cadenas de Markov describen procesos aleatorios que consisten de estados de un sistema en etapas sucesivas; el sistema puede realizar transiciones desde un estado a otro y la probabilidad de estos cambios depende solo del estado anterior.
Asimismo cada estado se puede considerar como un emisor de símbolos.
Tanto las transiciones como las emisiones del modelo tienen valores de probabilidad asociados. De esta forma se puede imaginar a un modelo de Markov como una máquina emisora de secuencias de símbolos, que trabaja internamente realizando transiciones entre los estados y emitiendo un símbolo cada vez que se alcanza uno de ellos. Cuando se desconoce la serie exacta de transiciones y emisiones que han ocurrido entre los estados se dice que el modelo es oculto.

En los modelos de Markov para la representación de alineamientos múltiples de una familia de secuencias, los estados del modelo se utilizan para representar las

columnas del alineamiento.

La construcción de un modelo de Markov requiere estimar los valores de sus parámetros desconocidos, que corresponden a las probabilidades de las emisiones y las transiciones a partir de parámetros observables. Para un alineamiento múltiple el cálculo de estas probabilidades puede realizarse partiendo un alineamiento previo o alineamiento semilla. El modelo Oculto de Markov así obtenido puede considerarse como un perfil estadístico que puede emplearse para incluir posteriormente otras secuencias al alineamiento o bien para realizar búsquedas de secuencias similares a las que lo componen.

Para la construcción de los perfiles se pueden utilizar alineamientos provenientes de bases de datos de dominios conservados para diferentes proteínas, o bien puede partirse de un alineamiento calculado previamiamente mediante otra estrategia, tal como las técnicas progresivas o las reiterativas. Existen diversas bases de datos en las cuales se pueden consultar alineamientos modelo de dominios conservados para diversas familias de proteínas, entre las cuales se puede citar la base de datos de dominios conservados del NCBI (Conserved Domain Database, o CDD, http://www.ncbi.nlm.nih.gov/cdd), la base de datos PFAM (actualmente mantenida por el EMBL, http://pfam.xfam.org), SMART (una base de datos para la identificación de dominios genéticamente móviles y para el análisis de la arquitectura de dominios, mantenida por el EMBL (http://smart.embl-heidelberg.de) y la base de datos de Clusters de Grupos Ortólogos del NCBI (Cluster of Orthologous Groups, COG, http://www.ncbi.nlm.nih.gov/COG/).

Por otro lado, las técnicas estadísticas para el alineamiento múltiple de secuencias, tales como los Modelos de Markov, son prácticamente las únicas herramientas bioinformáticas que permiten llevar a cabo el alineamiento de grandes cantidades de secuencias, por lo que en la actualidad revisten de gran interés para estudios de metagenómica.

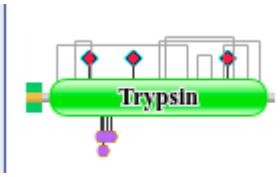
Dos de los paquetes de cómputo más sobresalientes para el alineamiento múltiple de secuencias, basado en el uso de HMMs son los programas HMMER y Clustal Omega que se revisaran en esta práctica.

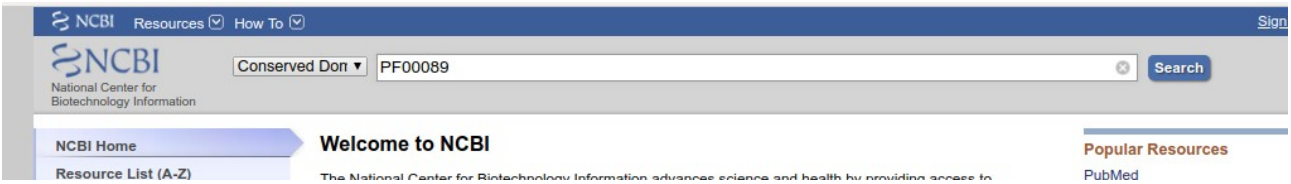
RECURSOS INFORMÁTICOS
Sitios WEB
EMBL – EBI https://www.ebi.ac.uk/
http://www.ncbi.nlm.nih.gov/blast
http://pfam.xfam.org/
http://pfam.xfam.orgProgramas
HMMER 3.2.1 (http://hmmer.org/)
HMMER – Guia del usuario: http://eddylab.org/software/hmmer/Userguide.pdf
Clustal Omega 1.2.0 (http://www.clustal.org/omega/)
Unipro UGENE 1.14 (http://ugene.unipro.ru)
HMMEditor 1.2 (http://sysbio.rnet.missouri.edu/multicom_toolbox/tools.html)

DESARROLLO
I) Obtención de un alineamiento múltiple de secuencias para generar el modelo.
1. Ingresar a la página principal del NCBI (http://www.ncbi.nlm.nih.gov/) y de ahí ir a la sección de programas BLAST
2. En la sección de aplicaciones especializadas de BLAST ingresar a la aplicación para realizar búsquedas dominios conservados. (CDART)
3. Utilizar para la búsqueda el número de acceso P00760 que corresponde a la tripsina

bovina y realizar la búsqueda en las bases de datos CDD (CDART) del NCBI y PFAM http://pfam.xfam.org/.

4. Observe los resultados de dominios conservados encontrados para esta proteína en ambas bases de datos. En particular para los resultados en la base de datos PFAM anote la clave de acceso del dominio encontrado (PF00089) y revise la descripción del registro Pfam de proteínas relacionadas a la tripsina bovina.





Se muestra un alineamiento representativo de este tipo de proteínas. Una forma más conveniente de guardarlo es dar clic en la sección “Links” y después en Pfam (Parte superior izquierda de la página)

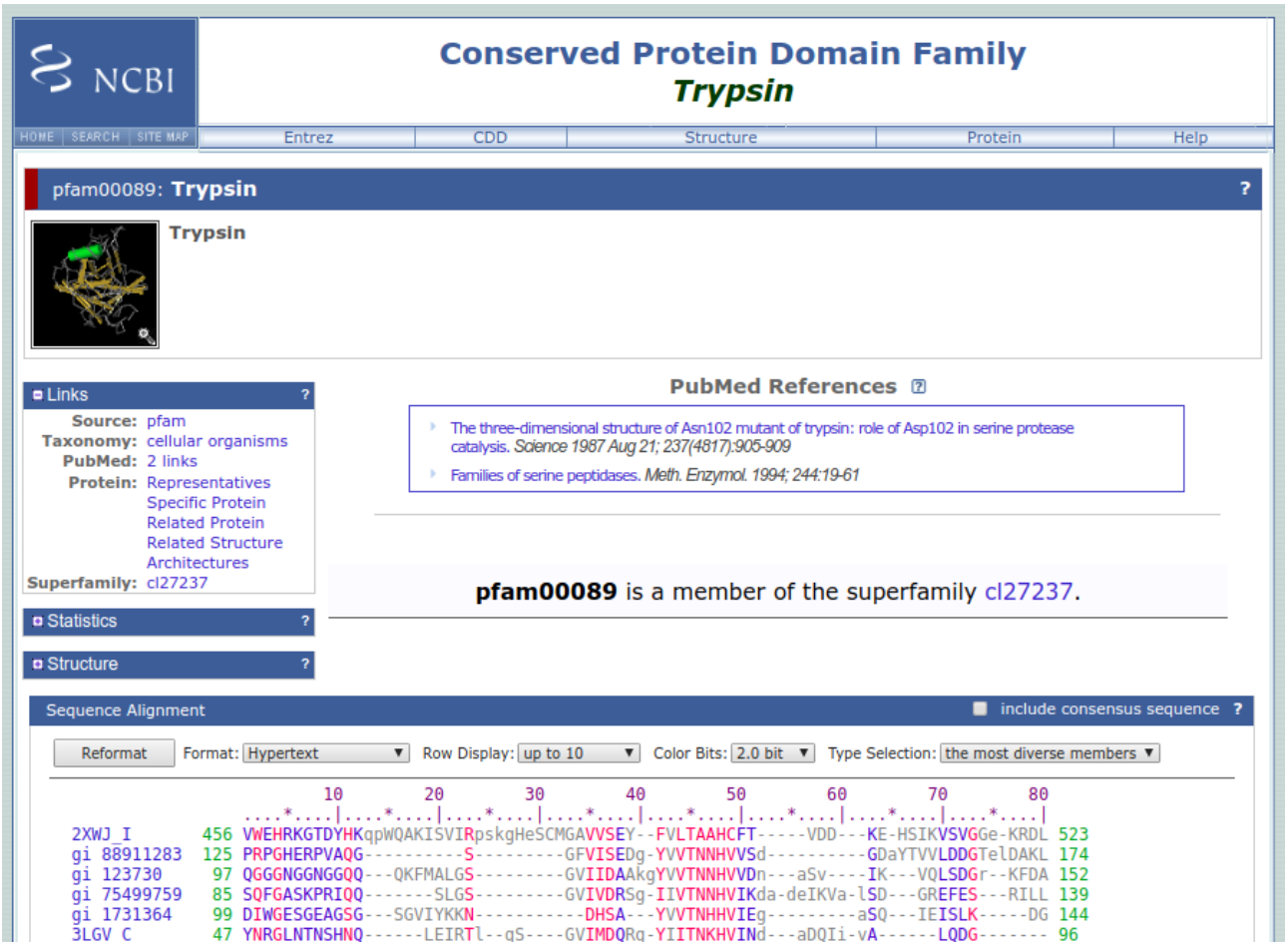


Figura 1: Información sobre la familia PFAM00089 en la base de datos de Dominios Conservados del NCBI:5.

Lo anterior conduce a la página de información de tripsina de la base de datos Pfam (http://pfam.xfam.org/family/PF00089).
6. Revise los alineamientos que pueden obtenerse para esta familia de proteínas en PFAM. En particular, el alineamiento denominado semilla (seed) es un alineamiento modelo que puede emplearse para la construcción de perfiles y HMMs. Revise la versión HTML de este alineamiento y preste atención a los datos anotados en el mismo. Por otro lado se llevará a cabo una descarga del alineamiento para construir patrones estadísticos. Para esto, en la sección "Format an Alignment / Alignment" elegir la opción semilla con 70 secuencias "seed (70)" y después el formato Stockholm. Preferentemente configure el alineamiento para mostrar únicamente guiones ("dashes") para los huecos. Descargue el alineamiento en una carpeta conveniente con el nombre PF00089.sto. Este alineamiento será empleado para crear el modelo oculto de Markov

II) Modelos ocultos de Markov de alineamientos múltiples de secuencias con HMMER
1. Una vez que se ha instalado HMMER (siga las instrucciones del archivo INSTALL) se recomienda copiar los archivos de las secuencias PFAM en formato Stockholm (PF00089.sto) y el de las secuencias de tripsinas sin alinear en formato FASTA para el alineamiento, en la carpeta BIN conteniendo las aplicaciones de HMMER.
2. La construcción del modelo de Markov se realiza desde la terminal de sistema mediante la instrucción 1 :
hmmbuild PF00089.hmm PF00089.sto
Esto permite crear el modelo oculto de Markov para el alineamiento de las secuencias y lo almacena en el archivo **PF00089.hmm**. Se puede abrir este archivo con un editor de textos. Observar su estructura ¿Qué secciones del archivo representan a las columnas del alineamiento? ¿Qué representan las letras mayúsculas del renglón marcado como HMM (cuéntelas)? ¿Qué representan los números presentes en esta sección? (ver el manual del usuario de HMMER – pags. 202 - 207)
3. Algunos editores permiten generar representaciones gráficas de modelos ocultos de Markov y logos. En particular el programa HMMEditor 1.2 2 , puede generar este tipo de representaciones gráficas del modelo. Esta herramienta requiere la instalación de Java para su ejecución. Nota: Para poder abrir el archivo con este editor, la primera línea de la cabecera del archivo PF00089.hmm debe modificarse para que coincida con el siguiente texto: **HMMER3/b [3.0 | March 2010]**
2
Puede obtener este programa en la página
http://www.mybiosoftware.com/hmmeditor-1-2-visual-editor-profile-hidden-markov-model.html
Utilice el formulario de descarga al final de la página y seleccione la descarga del programa HMMVE_1.2.tar.gz. Descomprima el archivo y ejecute el programa de extensión jar. (java -jar HMMVE_1.2.jar) Requiere la instalación de Java.

Abra el archvo con el HMMEditor y visualice el diagrama del modelo e identifique las partes correspondientes a los estados para representar columnas, inserciones, eliminaciones y transiciones entre los estados.
Compare esta información con el contenido del archivo de texto conteniendo el modelo.

4. Considerando que en este ejemplo el archivo que contiene a las tripsinas no alineadas se llama **TRIPSINAS.fasta**, ejecutar el siguiente comando para alinear a las secuencias:
hmmalign -o alineamiento_tripsinas.sto PF00089.hmm TRIPSINAS.fasta
Este comando almacena el alineamiento en el archivo alineamiento_tripsinas.sto en el

formato Stockholm.

5) Utilizar esl-reformat [-options] <format> <seqfile>
para convertir el formato stockholm a clustal en la secuencia de alineamiento obtenida.

Ejercicio: Encontrar el modelo HMM para la familia de proteinas PF00005 y nombrelo PF00005.hmm haga un screenshot de su modelo gráfico.

Al final Subirá al sitio web del curso, en la carpeta tareas un archivo CC471-LAB11-<Nombre-apellido>.zip con los archivos generados.

Bibliografía

1. Bateman, A., et al. (2004): The Pfam protein families database. Nucleic Acids Res 32(Database issue): D138-141.
2. Durbin R., S. R. Eddy, A. Krogh, G. Mitchison. (1998): Biological sequence analysis: Probabilistic models of proteins and nuclic acids. Cambridge University Press, London, England.
3. Eddy, S. R. (2004): What is a hidden Markov model? Nat Biotechnol 22(10): 1315-1316.
4. Eddy SR (2011). Accelerated Profile HMM Searches. PLoS Comput Biol. Oct;7(10):e1002195. doi: 10.1371/journal.pcbi.1002195.
5. Krogh A. (1998): An Introduction to Hidden Markov Models for Biological Sequences, en Computational Methods in Molecular Biology, Elsevier, editado por S. L. Salzberg, D. B. Searls and S. Kasif, pp: 45-63.
6. Mount D. W. (2001): Bioinformatics. Cold Spring Harbor Laboratory Press, New York, USA. pp:173-200.
7. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011 Oct 11;7:539.