CT 8 (2016) 179-197

Bioinformática
Análisis Bioinformático de la Major urinary protein 1

Víctor Rodríguez Pastor[1]
Universidad de Salamanca
victorrrp@usal.es

SUMARIO

El análisis bioinformático de las distintas proteínas es una técnica que se usa diariamente en muchos laboratorios de investigación. Recientemente, el número de proteínas conocidas ha aumentado exponencialmente gracias a los numerosos avances en los campos de la secuenciación masiva y la genómica funcional. Este trabajo se ha centrado en la "Major urinary protein 1", una proteína realmente poco conocida y que por tanto presenta un mayor interés de ser analizada en cuanto a su evolución filogenética entre las distintas especies.

Palabras clave: Bioinformática, árbol filogenético, Major urinary Protein 1, evolución.

SUMMARY

The analysis of bioinformatic proteins is employed daily in most research laboratories. Due to the amount of massive sequencing and breakthroughs in functional genomics, the number of known proteins is rising at an exponential rate. The major urinary protein 1 was chosen due to lack of information with regards to its function and phylogenetic relations. Therefore, this study will explore its origins and chronology.

Key words: Bioinformatics, phylogenetic tree, Major urinary Protein1, evolution.

1 Víctor Rodríguez Pastor es estudiante de 3º del Grado en Biotecnología en la Universidad de Salamanca.

1. INTRODUCCIÓN

El análisis bioinformático realizado tuvo como objetivo elegir una proteína de interés, la "Mayor urinary protein 1", e intentar averiguar sus relaciones filogenéticas en las distintas especies. Para ello se realizaron los siguientes análisis bioinformáticos: En primer lugar, se realizó un BLAST para ver las secuencias parecidas a ésta que existen en las distintas especies, para ello se retocaron los distintos parámetros hasta encontrar una configuración óptima. En segundo lugar, se guardaron y anotaron las distintas secuencias pertenecientes a cada especie y se hizo un alineamiento de múltiples secuencias (MSA) utilizando tres alineadores distintos para poder contrastar los resultados. Una vez realizado el alineamiento de múltiples secuencias, se utilizó el programa MEGA7 mediante tres algoritmos diferentes para la creación de un árbol filogenético que nos mostrara el emparejamiento de las distintas especies según la evolución de esta proteína y los antecesores predichos por el programa, con su correspondiente significación estadística. Por último, se estableció un patrón y una secuencia consenso, de manera que, si en el futuro se quiere introducir otra proteína en la misma familia que la "Mayor urinary protein 1", se podría comprobar si ésta se ajusta.

En cuanto a la función de la "Mayor urinary protein 1", solo se sabe que se une a feromonas que son excretadas en la orina de los machos y que, por tanto, tiene relación con el comportamiento sexual de las hembras. Además, es un receptor que se activa por insulina y éste está relacionado con muchas funciones biológicas como la respiración y la regulación del metabolismo.

El organismo en el que se ha identificado esta proteína es el *Mus musculus* (Ratón), su identificador en UniProt es P11588 y se abrevia como MUP 1.

2. MÉTODOS

2.1. BLAST

En primer lugar, hice un p-blast para tener una visión previa de las distintas especies que tenían proteínas con secuencias parecidas a la MUP 1. En los parámetros excluí todas aquellas proteínas provenientes del *Mus musculus* ya que sabemos que este es el organismo de procedencia de la proteína y queremos encontrar la relación entre distintas especies. Además, también se filtraron de la búsqueda todas aquellas secuencias que o bien solo hubieran sido predichas y no se hubiera obtenido en el laboratorio o bien se hubieran obtenido mediante metagenómica (*uncultured)*.

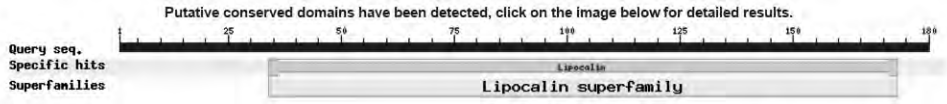


Figura 1. El BLAST nos informa de que tenemos un dominio perteneciente a la superfamilia de las lipocalinas.

Desde el primer momento, el BLAST ya predijo que la secuencia aminoacídica de la MUP1 contiene un dominio conservado de la superfamilia de las lipocalinas, a saber, varios miembros que comparten una estructura común, definida en general por la unión de ligandos lipofílicos.

Una vez realizado el BLAST, vemos que nos devuelve en primer lugar la misma proteína que no estaba bien anotada en la base de datos con el nombre de la especie y que por tanto no había sido filtrada al excluir *Mus musculus*. Una vez eliminada ésta, nos fijamos en que en el resto de secuencias encontradas, el porcentaje de identidad se encuentra entre el 40% y el 60%, a excepción de la perteneciente al organismo *Mus Spretus* que es otro tipo de ratón y por tanto también eliminaremos de la búsqueda.

El siguiente paso fue modificar los parámetros por defecto del BLAST ya que las identidades no son demasiado altas y por tanto podría haber secuencias que estuviéramos perdiendo por el camino al tener los parámetros demasiado estrictos: El *Word size* lo mantenemos en 6 ya que parece que bajarlo solo serviría para que el BLAST tardara más en realizarse, puesto que estamos encontrando un número suficiente de secuencias con este nivel de sembrado. El número máximo de secuencias lo subimos a 250 para poder tener mayor variedad de especies distintas y que nuestro árbol filogenético sea más rico. Además, desactivamos el botón de consultas cortas ya que no queremos que el BLAST nos modifique los parámetros. El *expected threshold* lo mantenemos en 10 ya que nos aporta un número suficiente de secuencias sin subirlo y la matriz la bajamos a una BLOSSUM más baja, en este caso, tras probar también varias PAM elegimos la BLOSSUM 45, para poder encontrar secuencias más divergentes que con la matriz por defecto. Por otro lado, mantenemos los valores por defecto de penalización de la creación y extensión de *gaps* ya que necesitaríamos tener un mayor conocimiento biológico sobre la proteína para poder retocar este parámetro de forma fiable.

Tras haber cambiado estos parámetros, nos encontramos que hay algunas secuencias nuevas que antes no estaban y además los e-valores de las que ya habían aparecido en la primera búsqueda son menores y por tanto estadísticamente más fiables (incluso hasta un valor de e-74).

Se recogen en la siguiente tabla las 36 secuencias elegidas para la realización del estudio bioinformático; si bien, de algunas especies cogimos más de una secuencia para posteriormente corroborar que los algoritmos de creación de árboles filogenéticos habían funcionado correctamente, ya que, *a priori*, las secuencias pertenecientes al mismo organismo deberían aparecer en ramas adyacentes.

ESPECIE	**Accession Number**
[Mus musculus]	P11588
[Rattus norvegicus]	NP_671745.1
[Rattus norvegicus]	NP_976070.1
[Mesocricetus auratus]	AGV07608.1
[Myotis davidii]	ELK24596.1
[Felis catus]	NP_001009233.1
[Canis lupus familiaris]	NP_001271390.1
[Oryctolagus cuniculus]	CCC15303.1
[Equus caballus]	NP_001075966.1
[Bos taurus]	DAA26480.1
[Bos mutus]	ELR50248.1
[Heterocephalus glaber]	EHB13722.1
[Sus scrofa]	NP_998979.1
[Cavia porcellus]	CAX62131.1
[Fukomys damarensis]	KFO26747.1
[Sus scrofa]	1GM6_A
[Capra hircus]	AHZ46504.1
[Camelus ferus]	EPY89972.1
[Cricetulus griseus]	ERE70560.1
[Tupaia chinensis]	ELW70591.1
[Macaca mulatta]	NP_001165312.1
[Heterocephalus glaber]	EHB11625.1
[Tachyglossus aculeatus]	ABJ90458.1
[Cavia porcellus]	NP_001192113.1
[Rattus norvegicus]	CAA29100.1
[Oncorhynchus masou formosanus]	ABY21331.1
[Fukomys damarensis]	KFO22774.1
[Oreochromis niloticus]	CCF55069.1
[Oncorhynchus mykiss]	CDQ80766.1
[Anoplopoma fimbria]	ACQ59033.1
[Salmo salar]	ACI69895.1
[Oreochromis niloticus]	CCF55070.1

[Pteropus alecto]	ELK07359.1
[Xenopus laevis]	NP_001081513.1
[Danio rerio]	NP_998799.1
[Cavia porcellus]	NP_001192113.1

Figura 2. Tabla con las secuencias escogidas.

Estos resultados se obtuvieron con los parámetros ya comentados anteriormente:

Search Parameters	
Program	blastp
Word size	6
Expect value	10
Hitlist size	250
Gapcosts	15,2
Matrix	BLOSUM45
Filter string	F
Genetic Code	1
Window Size	40
Threshold	21
Composition-based stats	2

Database	
Posted date	May 6, 2016 8:16 PM
Number of letters	31,916,365,256
Number of sequences	87,005,143
Entrez query	all [filter] NOT(txid10090 [ORGN] OR txid10096 [ORGN] OR (XP_000001:XP_999999[pacc] OR XP_000000001:XP_99999999[pacc]) OR environmental samples[organism] OR metagenomes[orgn] OR txid32644[orgn])

Karlin-Altschul statistics		
Lambda	0,231236	0,203
K	0,0940427	0,041
H	0,244528	0,12
Alpha	0,9113	1,7
Alpha_v	9,61106	44,8257
Sigma		45,0604

Figura 3. Parámetros del Blast.

También probé a hacer un PSI-BLAST con 5 interacciones, pero la mayoría de las secuencias nuevas respecto al p-BLAST eran sólo *hipotéticas* y, como ya tenía suficientes secuencias y especies distintas, decidí no incluir ninguna de ellas.

2.2. ALINEAMIENTO DE MÚLTIPLES SECUENCIAS (MSA)

El alineamiento de múltiples secuencias es la técnica bioinformática que nos permite aplicar un algoritmo y conseguir alinear secuencias aminoacídicas de proteínas que tienen una cierta homología. De esta manera podemos encontrar qué zonas están más conservadas en la evolución, lo cual indica que son regiones muy importantes para la función de la proteína, ya que las zonas menos importantes tienen mayor número de variaciones en la secuencia aminoacídica entre las distintas especies. Además, el alineamiento de múltiples secuencias sirve como paso previo para encontrar posibles sitios catalíticos de las enzimas que sean susceptibles de sufrir mutaciones dirigidas para aumentar su efectividad.

Para el alineamiento de múltiples secuencias usé tres herramientas bioinformáticas, de las cuales dos de ellas fueron *online* (Clustal Omega y T-coffee) y la otra fue el software informático Jalview.

En primer lugar, utilicé Clustal Omega en la página web del EBI (Instituto Europeo de Bioinformática). Esta herramienta no nos da demasiadas opciones en cuanto a cambio de parámetros se refiere, probé a aumentar el número de interacciones del HMM y el número de interacciones del árbol guía para así conseguir resultados más fiables. Aunque las zonas conservadas finales se presentarán con Jalview, aquí expongo una de las zonas conservadas obtenida con Clustal Omega (nuestra proteína de partida está identificada como "MUP1-Musmusculus" para así poder diferenciarla del resto y ver su homología):

Figura 4. Alineamiento con Clustal omega.

El siguiente método elegido fue la herramienta T-coffee. A pesar de que este programa *online* tampoco nos permite cambiar parámetros interesantes, sí que nos decidimos a usar su versión PSI-Coffee. La elección de esta versión del programa se debe a que las 36 secuencias que hemos elegido para nuestro estudio son de especies muy diversas y, además, sus porcentajes de identidad ronda en algunos casos el 30%. Por tanto, es mejor utilizar esta versión que, a pesar de ser más lenta (tardó en su ejecución 4 horas y 59 minutos mientras que el T-Coffe normal se ejecutó en 10 segundos), es mucho más precisa a la hora de encontrar homologías en proteínas que a priori están distanciadas. Aquí ejemplifico una de las zonas conservadas encontradas:

Figura 5. Alineamiento con Psi-Coffee.

En último lugar, utilicé el programa Jalview para visualizar el alineamiento. Jalview es un software que te permite utilizar distintos algoritmos alojados en servicios web, tales como Clustal Omega, T-coffee, etc. En este caso, como ya habíamos utilizado los dos anteriores en sus versiones *online,* utilicé el algoritmo Muscle. Como en el BLAST, decidí cambiar la matriz a una BLOSSUM45 ya que estas secuencias tienen relativamente poca identidad y por tanto se obtienen mejores resultados si se es más permisivo con las sustituciones de aminoácidos. Además le indiqué que se trataban de secuencias proteicas. Sin embargo, el programa me devolvió errores de ejecución y por tanto probé con las matrices BLOSSUM40, PAM170 y PAM140 (matrices destinadas a secuencias lejanas) pero todas daban error. Parece que la última versión del jalview debe tener algún tipo de error con el Muscle ya que solo me devuelve resultados si ejecuto el algoritmo por defecto, lo cual es una pena porque el alineamiento podría ser más preciso con una matriz más permisiva. Por tanto, ejecuté el Muscle por defecto y obtuve las siguientes zonas conservadas:

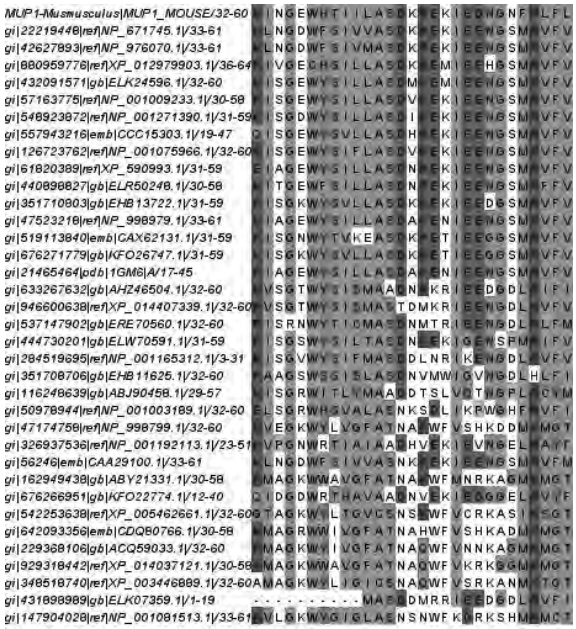


Figura 6. Zona conservada número 1.

Figura 7. Zona conservada número 2.

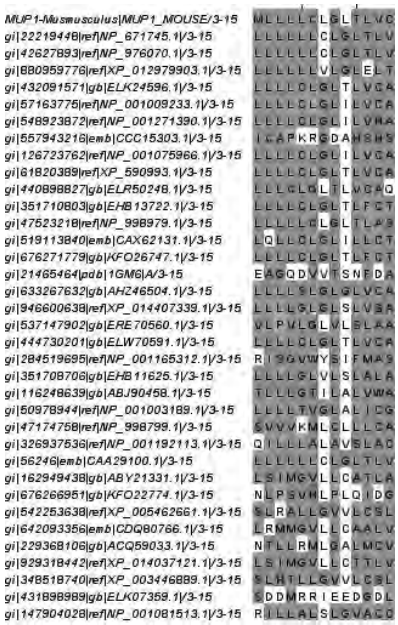


Figura 8. Zona Conservada número 3.

Figura 9. Zona conservada número 4.

He de comentar que los resultados obtenidos con los distintos algoritmos no variaron demasiado, las zonas conservadas quedaron prácticamente iguales y solo hubo cambios en las zonas de gaps. Esto se debe a que los porcentajes de identidad, aunque eran bajos (entre el 30 y el 40%), sí que se obtiene un mejor alineamiento que si se hubieran cogido secuencias con porcentajes de identidad cercanos al 20% que hubieran estado en el umbral de incertidumbre de si esas secuencias hubieran tenido ancestros comunes o no, y por tanto, los algoritmos informáticos no hubieran tenido un resultado claro y hubieran variado más.

En último lugar, se utilizó la herramienta WebLogo para ver las secuencias consenso en cada una de las cuatro zonas conservadas obtenidas anteriormente:

Figura 10. Logo zona conservada número 1.

Figura 11. Logo zona conservada número 2.

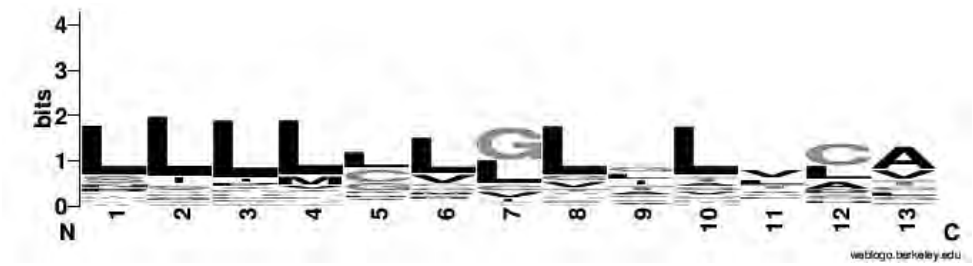


Figura 12. Logo zona conservada número 3.

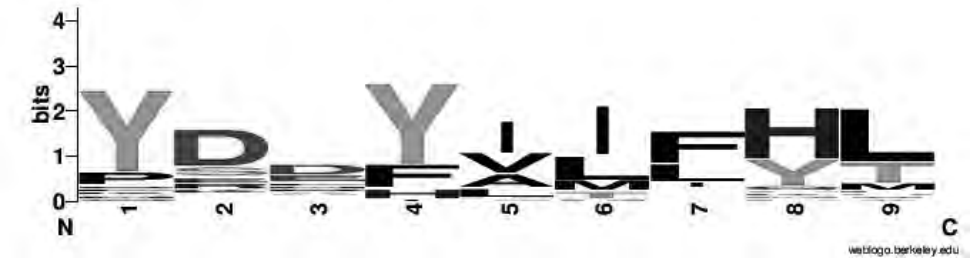


Figura 13. Logo zona conservada número 4.

La creación de estos Logos se basa en las proporciones relativas de cada aminoácido en cada columna del alineamiento de las regiones conservadas. De esta manera, las letras más grandes nos indican que hay una mayor proporción de ese aminoácido en esa columna del alineamiento de múltiples secuencias.

2.3. Árboles filogenéticos

En este paso reside la mayor parte del interés del trabajo, ya que es donde por fin vemos las relaciones filogenéticas que existen entre las distintas especies teniendo en cuenta cómo han evolucionado estas secuencias de proteínas que tienen una cierta identidad entre sí.

Se han utilizado tres algoritmos distintos para la creación del árbol: NJ, UPGMA y máxima parsimonia. Todos ellos disponibles en el programa gratuito MEGA 7.

Además, le he aplicado a los tres el bootstrapping, un análisis estadístico que nos dice cómo de fiable es cada una de las uniones entre ramas del árbol.

Para realizar los árboles utilicé la zona conservada número 1 ya que es la más larga y, por tanto, la más significativa estadísticamente.

Al intentar hacer los árboles por primera vez se encontró un error en 2 de los 3 algoritmos (NJ y UPGMA) ya que sólo nos permite hacer árboles si todas las secuencias tienen aproximadamente la misma longitud, por tanto, tuve que desactivar la secuencia de *Pteropus alecto* para que me permitiera calcular correctamente las distancias. El problema de esta secuencia era que en esta zona conservada entre proteínas tenía 9 gaps seguidos al principio.

También conviene recordar que para leer las distancias existentes entre especies mediante el árbol filogenético, no se considera la distancia en vertical, sino en horizontal, ya que dos ramas pueden rotar sobre el nodo sin que cambie el signifi-

cado del árbol. La distancia se calcula sumando toda la distancia horizontal hasta el nodo común a la distancia existente hasta el otro miembro con el que queremos ver la distancia filogenética.

En primer lugar, utilicé el método NJ (Neighbor-Joining). Marqué la casilla del Boostrap method, le informé de que se trataba de una secuencia de aminoácidos y fijé el número de interacciones del bootstrap en 500: (Ver ANEXO 1)

En este árbol podemos comprobar que el algoritmo se ha realizado correctamente ya que las secuencias pertenecientes a la especie Sus Crofa quedan muy próximas entre ellas, así como las de Rattus Novergicus y las de Oreochromis niloticus. Esto se debe a que, al pertenecer a la misma especie, han tenido menos tiempo evolutivo y por eso las secuencias son más parecidas. En el caso de Rattus norvegicus, a pesar de llamarse igual vemos que las secuencias son diferentes y por ello pensamos que han sido secuenciadas a partir de dos animales distintos pertenecientes a la misma especie. Por otro lado, vemos cómo las conservaciones a nivel de género también se mantienen a nivel de secuencia en los casos de Oncorhynchus y Bos. También es razonable que nuestra secuencia de partida de Mus musculus (es decir, ratón) se encuentre cercana a Rattus novergicus ya que los ratones y las ratas son muy parecidas.

Sin embargo, lo que más nos llama la atención son los casos de Canis lupus familiaris y de Cavia porcellus en los que a pesar de ser proteínas pertenecientes a la misma especie, están muy alejadas filogenéticamente. Esto se puede explicar porque se trata de alérgenos los cuáles varían bastante.

En siguiente lugar se realizó el método UPGMA también utilizando 500 interacciones de Bootstrapping y esta vez cambiamos el formato del árbol a straight: (VER ANEXO 2)

En este árbol vemos cómo se cumplen las mismas cosas que comentamos en el anterior. Esto confirma que los algoritmos se han llevado a cabo correctamente y aunque las significaciones estadísticas varíen, la predicción filogenética es buena.

Ahora en último lugar usamos el método de máxima parasimonia. Para la realización de este método sí que nos permite utilizar todas las secuencias y por tanto activamos otra vez la de *Pteropus alecto* y, además, como éste es el único en el que tenemos todas las secuencias, aumentamos el bootstrapping a 1000 para tener una mayor significación estadística: (VER ANEXO 3)

En este método lo que más se destaca es que Mus musculus aparece más alejado filogenéticamente de Rattus novergicus que en los casos anteriores y sobre todo el caso de las proteínas relativas a Oncorhynchus que aparecen muy alejadas entre ellas. Sin embargo, como este método tuvo un boostrapping más grande que

los anteriores y tardó en realizarse casi 5 horas el árbol, entendemos que debería ser a priori más correcto que los dos anteriores. Aun así, sabemos que todo este tipo de predicciones filogenéticas no son dogmas, sino que se basan en predicciones estadísticas y por tanto los resultados varían entre métodos y además no asegura que la evolución ocurriera así realmente.

2.4. Búsqueda de patrones y pHMMs

En este caso nos centramos en buscar un patrón para cada una de las zonas conservadas encontradas durante el alineamiento de múltiples secuencias para que así, en un futuro, si queremos ver si una nueva proteína se ajusta a esta familia de proteínas y por tanto tiene identidad con las proteínas de este grupo, solo tendremos que ver si se ajusta al patrón. Además, este patrón al introducirse en PROSITE nos deja ver qué proteínas relacionadas lo cumplen.

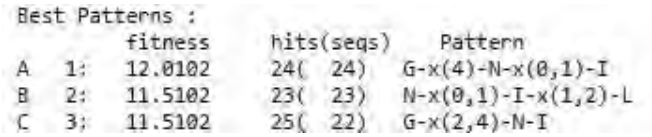
En cuanto a la primera secuencia conservada, como no todas las secuencias varían en las mismas posiciones y tienen porcentajes relativamente bajos de identidad entre ellas, fue difícil encontrar un patrón válido. En primer lugar, subí la *max flexibility* y el *máx num flex spaces* a 5 cada uno para que hubiera más flexibilidad a la hora de hacer el patrón. Además, bajé el porcentaje mínimo de coincidencia con el patrón al 70%. El patrón resultante de estos cambios era demasiado corto y por tanto volví a bajar el porcentaje mínimo a 60% y nos dio lo siguiente:

```
Best Patterns :
          fitness     hits(seqs)    Pattern
A   1:    27.6904     22(  22)   A-x-D-x(4)-I-x(0,1)-E-x(1,2)-G-x(2)-R-x(1,2)-F
B   2:    27.6904     22(  22)   A-x(1,2)-D-x(3,4)-I-E-x(2)-G-x(2)-R-x(1,2)-F
C   3:    27.1904     22(  22)   A-x-D-x(4)-I-x(0,1)-E-x(1,2)-G-x(2)-R-x(0,2)-F
```

Figura 14. Patrones zona conservada número 1.

De los tres patrones con mayor significancia estadística cogí el primero y lo introduje en PROSITE para buscar proteínas que cumplieran este patrón. Nos salieron 16 secuencias encontradas, además, sabemos que el patrón funcionó porque alguna de estas secuencias encontradas en PROSITE estaban ya incluidas en nuestro análisis de múltiples secuencias como Major allergen Equ_c 1.*Equus caballus (Horse)*. Además, también podemos confirmar que nos encontró proteínas relacionadas ya que nos salían en la búsqueda otras Mayor Urinatory Proteíns del Mus musculus como la 11, la 6 y la 5.

En cuanto a la secuencia número 2, primero probé con un porcentaje mínimo del 80% pero me dio un patrón muy corto. Después probé con el 60% y me dio este resultado:

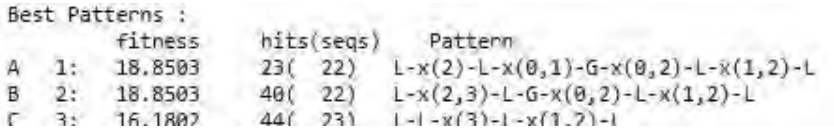


```
Best Patterns :
          fitness     hits(seqs)    Pattern
A   1:    12.0102     24(  24)   G-x(4)-N-x(0,1)-I
B   2:    11.5102     23(  23)   N-x(0,1)-I-x(1,2)-L
C   3:    11.5102     25(  22)   G-x(2,4)-N-I
```

Figura 15. Patrones zona conservada número 2.

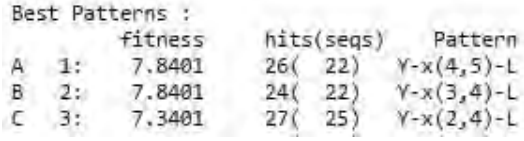
El problema de esta y de las dos secuencias conservadas que nos quedan por mostrar es que son más cortas que la primera y por tanto son menos específicas de esta familia de proteínas y el PROSITE no es capaz de encontrar específicamente este tipo de proteínas al tener patrones cortos y con poca puntuación.

De todas formas, mostraremos cuáles fueron los patrones de las secuencias conservadas número 3 y 4:



```
Best Patterns :
          fitness     hits(seqs)    Pattern
A   1:    18.8503     23(  22)   L-x(2)-L-x(0,1)-G-x(0,2)-L-x(1,2)-L
B   2:    18.8503     40(  22)   L-x(2,3)-L-G-x(0,2)-L-x(1,2)-L
C   3:    16.1802     44(  23)   L-L-x(3)-L-x(1,2)-L
```

Figura 16. Patrones zona conservada número 3.



```
Best Patterns :
          fitness     hits(seqs)    Pattern
A   1:     7.8401     26(  22)   Y-x(4,5)-L
B   2:     7.8401     24(  22)   Y-x(3,4)-L
C   3:     7.3401     27(  25)   Y-x(2,4)-L
```

Figura 17. Patrones zona conservada número 4.

2.5. Búsqueda en PFAM de la secuencia consenso

Por último, descargamos la secuencia consenso del alineamiento con JALVIEW obtenida mediante el algoritmo MUSCLE y eliminamos toda la parte intermedia que solo existía en una proteína y por tanto no era representativa de la familia. Además, quitamos los signos "+" que no los reconoce como secuencia proteica el PFAM. La secuencia es la siguiente:

MMKLLLLLLCLGLLTLVCAHAEEAEDVVRSNFDLEKISGEWYSILLASDKK
EKIEENGSMRVFVESIVLENS+SLSFKFAHTKVNGECTEVSLVCDKTEKDGV
YTVEYQRWFDGENKFRIVETDYDDYIIFHLINFKNGETFQLLELYGRPDLSP
ESLKEKFVQFCQKYGIVKENIIDLTKVDLRCLQARGSGVAQASSAETSD

Al introducirlo en PFAM, efectivamente nos dice (como ya nos había adelantado el BLAST al principio del trabajo) que nuestra secuencia consenso pertenece a la familia de las lipocalinas: "Lipocalin / cytosolic fatty-acid binding protein family".

3. CONCLUSIONES

Nuestra investigación sobre la MUP1 (Major Urinatory Protein 1) de Mus musculus se ha llevado a cabo correctamente. Se han variado los parámetros de los distintos algoritmos y procedimientos de forma óptima para obtener resultados interesantes sobre nuestra proteína que pertenece a la familia de las lipocalinas.

Se ha obtenido un árbol bastante rico con 35/36 secuencias (dependiendo del algoritmo usado) pertenecientes a 30 especies distintas en las que se refleja la evolución filogenética de proteínas parecidas a la nuestra de interés.

Además, se ha conseguido satisfactoriamente un patrón que encuentra otras proteínas parecidas en PROSITE y una secuencia consenso que se puede utilizar en el buscador de PFAM.

4. BIBLIOGRAFÍA

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool". J. Mol. Biol. 215: 403-410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". Nucleic Acids Res. 25: 3389-3402.

T-Coffee: A novel method for multiple sequence alignments. Notredame, Higgins, Heringa, JMB, 302 (205-217) 2000.

Edgar, Robert C. (2004), MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research 32 (5), 1792-97.

Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator, Genome Research, 14:1188-1190, (2004).

Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal OmegaSievers F, WILM A, DINEEN DG, GIBSON TJ, KARPLUS K, LI W, LOPEZ R, MCWILLIAM H, REMMERT M, SÖDING J, THOMPSON JD, HIGGINS D. Molecular Systems Biology 7 Article number: 539 doi:10.1038/msb.2011.75

MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets (Kumar, Stecher, and Tamura 2015).

WATERHOUSE AM, PROCTER JB, MARTIN DMA, CLAMP M, BARTON GJ (2009) Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191. doi:10.1093/bioinformatics/btp033

MICHENER, C.D., SOKAL, R.R. (1957): A quantitative approach to a problem of classification. Evolution, 11: 490-499.

SAITOU, N., NEI, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees, Molecular Biology Evolution 4: 406-425.

"Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous". Nature 431 (7011): 980-984

SIGRIST CJA, DE CASTRO E, CERUTTI L, CUCHE BA, HULO N, BRIDGE A, BOUGUELERET L, XENARIOS I. New and continuing developments at PROSITE. Nucleic Acids Res. 2012; doi: 10.1093/nar/gks1067

The Pfam protein families database: towards a more sustainable future: R.D. FINN, P. COGGILL, R.Y. EBERHARDT, S.R. EDDY, J. MISTRY, A.L. MITCHELL, S.C. POTTER, M. PUNTA, M. QURESHI, A. SANGRADOR-VEGAS, G.A. SALAZAR, J. TATE, A. BatemanNucleic Acids Research (2016) Database Issue 44:D279-D285.

5. ANEXOS

5.1. Anexo 1: Neighbor-joining

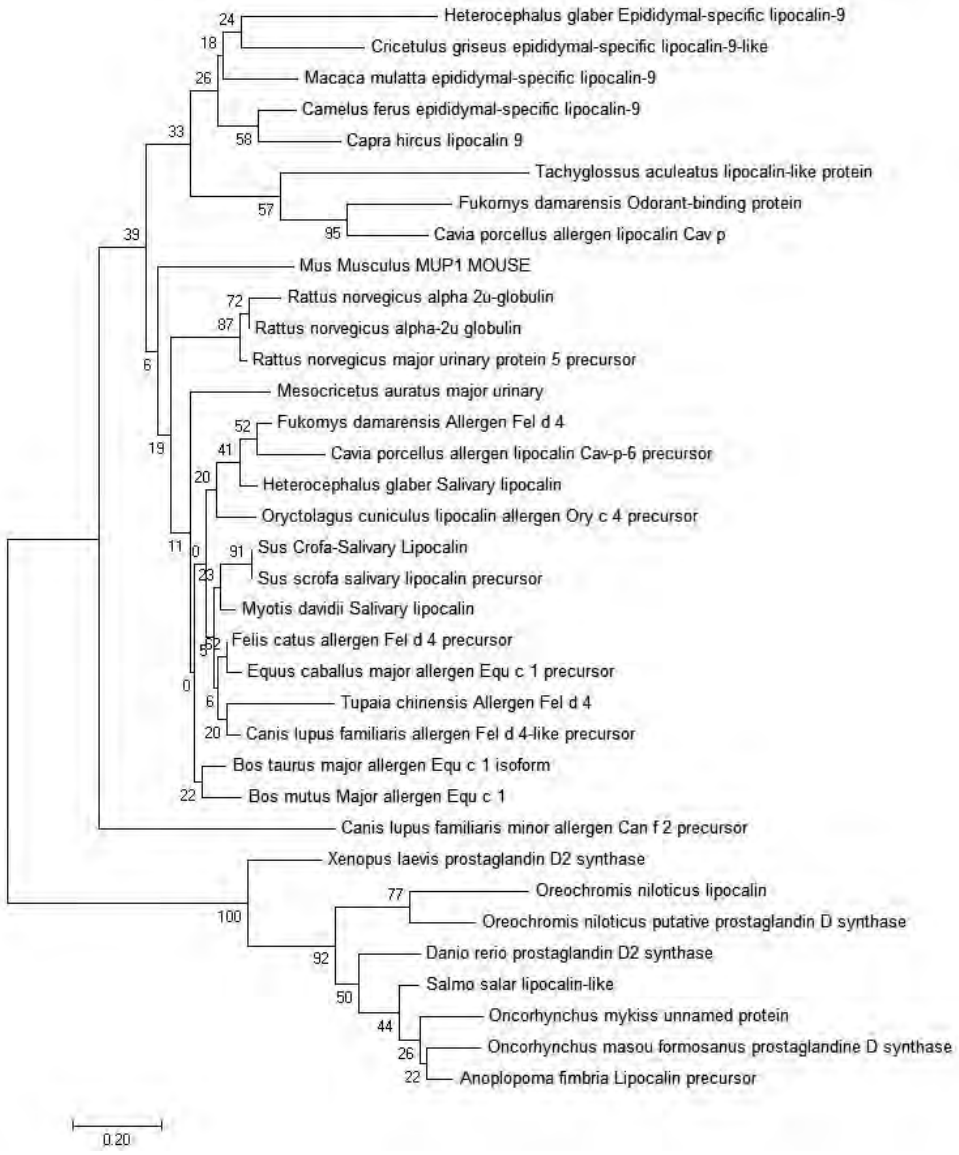


Figura 18. Árbol Neighbor-Joining.

5.2. ANEXO 2: UPGMA

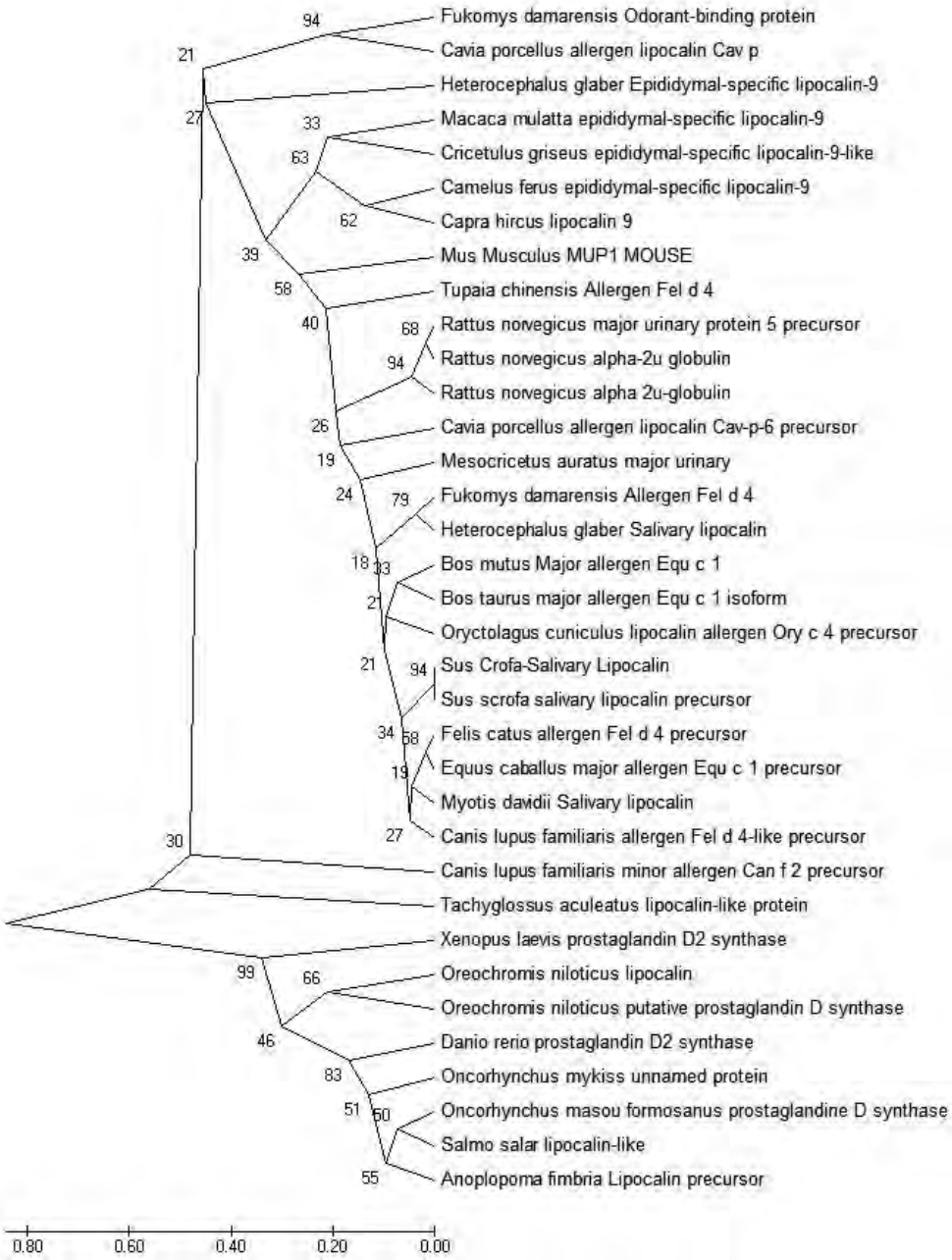


Figura 19. Árbol UPGMA.

5.3. Anexo3: Máxima Parasimonia

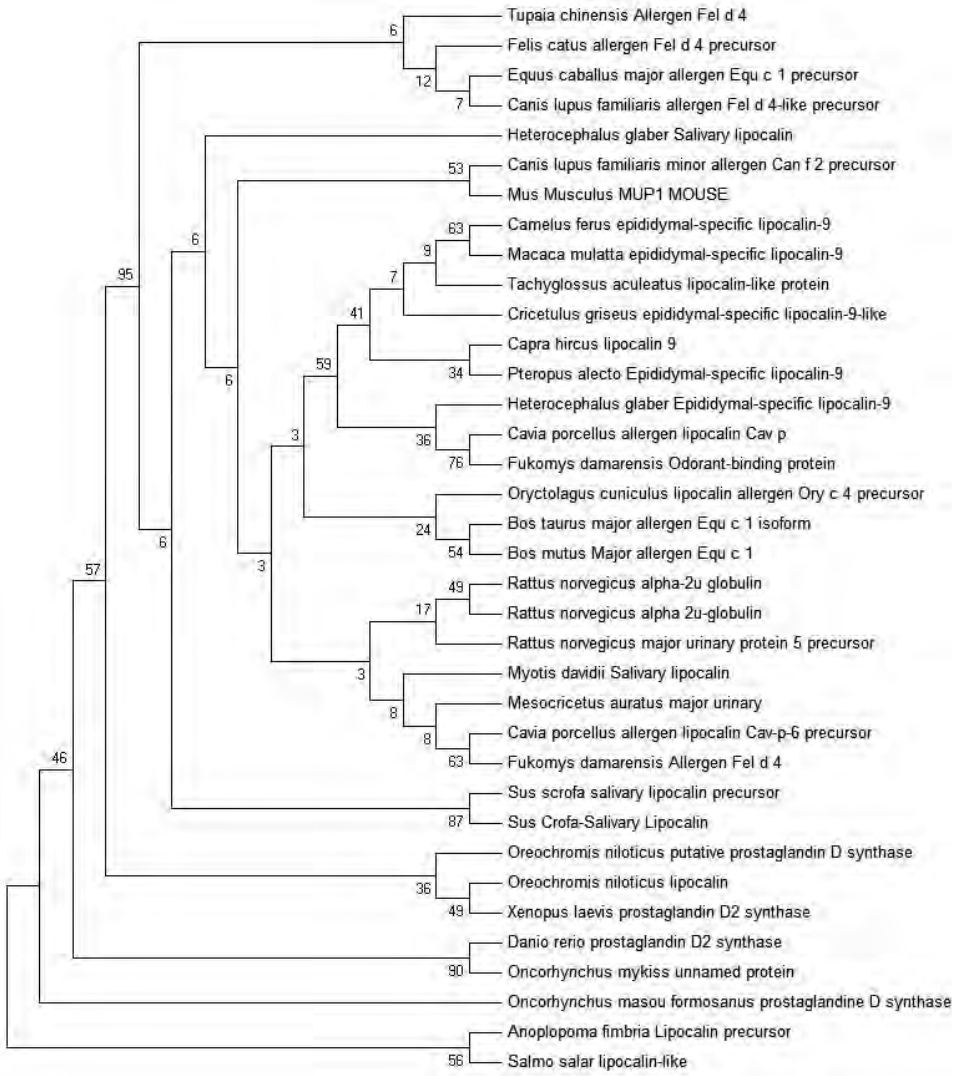


Figura 20. Árbol máxima parasimonia.