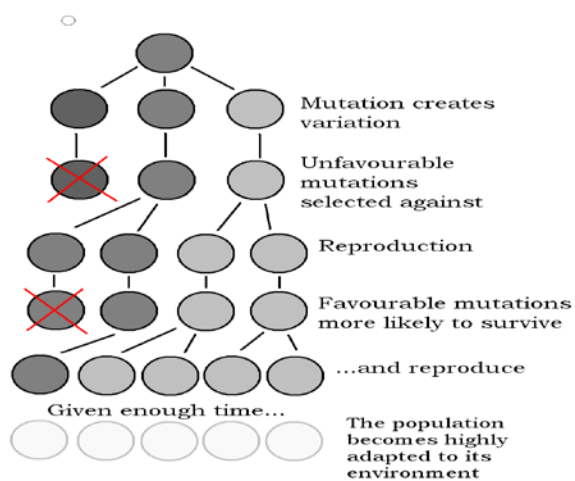


Universidad Andrés Bello
Facultad de Ciencias Biológicas
Center for Bionformatics & Integrative Biology
Chile

Evolución Molecular



Romina V. Sepúlveda

Jonathan Canan

Daniel Aguayo

2017

1. Proteínas y evolución

La función de las moléculas biológicas están directamente ligadas a sus propiedades fisicoquímicas y como estas son moduladas por el solvente, estructura, interacciones, relaciones y por como estas evolucionaron.

La forma que adoptan las moléculas biológicas están influenciadas por las propiedades del agua; por ejemplo, los fosfolípidos adaptan su estructura para alejar sus zonas hidrofóbicas (cadenas aciles) del agua, formando estructuras más complejas que han permitido la compartimentalización celular. Por su parte, las estructuras de las proteínas también están definidas por la interacción de sus componentes –los aminoácidos– con el solvente, ya que residuos hidrofóbicos tienden a empaquetarse para minimizar su interacción con el agua (ver cuadro Entropía y el plegamiento proteico), mientras que los grupos polares son estabilizados por sus propias interacciones y con la formación de enlaces de hidrógeno intra e intermoleculares.

Las estructuras que adoptan las biomoléculas también cumplen un rol importante y, como veremos más adelante, muchas veces son la clave para entender su función. Las proteínas adoptan patrones estructurales, por ejemplo las α -hélices y las láminas- β , que finalmente se condensan adoptando estructuras tridimensionales más complejas. Esta condensación o plegamiento entre ellas favorece la formación de enlaces de hidrógeno en el backbone y maximiza las interacciones entre residuos polares y cargados, en reemplazo de las interacciones con el agua, perdidas durante la compactación de núcleo hidrofóbico.



Entropía y el plegamiento proteico. Desde un punto de vista simple, una proteína desplegada tiene una alta entropía configuracional, asociada a la alta cantidad de configuraciones teóricas que esta puede adoptar, pero también una alta entalpía ya que su estructura está estabilizada por un bajo número interacciones. Por otro lado una proteína plegada tiene una entropía considerablemente menor, pero con una alta entalpía. A partir de la ecuación $\Delta G = \Delta H - T\Delta S$, es posible inferir que el plegamiento proteico esta dominado por un “juego” entre la entropía, la entalpía y la temperatura, sin embargo, esta explicación del plegamiento proteico omite algo fundamental, el proceso ocurre en presencia de un solvente particular, el agua. Cada vez que un dominio hidrofóbico esta expuesto a solvente, este interrumpe la red de enlaces de hidrógeno y constriñe las configuraciones que las moléculas de agua adyacentes pueden adoptar. De esta forma, su compactación en el llamado núcleo hidrofóbico, permite que las moléculas de agua aumenten su entropía con respecto al estado desplegado. Aun más, la red de enlaces de hidrógeno de los residuos polares y el esqueleto de la proteína (backbone) en agua es máxima tanto para el estado plegado como desplegado, por lo cual la diferencia de entalpía entre ambos estados es nula ($\Delta H = 0$) y el cambio esta casi en su totalidad guiado por efectos entrópicos.

Las estructuras que adoptan las biomoléculas están directamente ligadas con su función, siendo ambas seleccionadas durante la evolución. El mejor ejemplo de la relación entre estructura, función y evolución es a través del estudio de las enzimas. Las enzimas catalizan reacciones químicas a través de mecanismos complejos que les confieren una alta eficiencia con respecto a las reacciones químicas regulares. Habitualmente las reacciones enzimáticas ocurren en una región cerca de la superficie proteica, denominado “**sitio activo**” (Figura 1), una región pequeña comparada con el tamaño de una proteína. Cabe destacar que no tan solo el sitio activo es necesario para que ocurra la reacción enzimática, sino que todo el resto de la proteína. Esto habra una pregunta interesante ¿Cómo llegaron las proteínas a ser una máquina tan sofisticada? La respuesta es que fueron desarrolladas a través de un proceso temporal, en los cuales los cambios en la secuencia

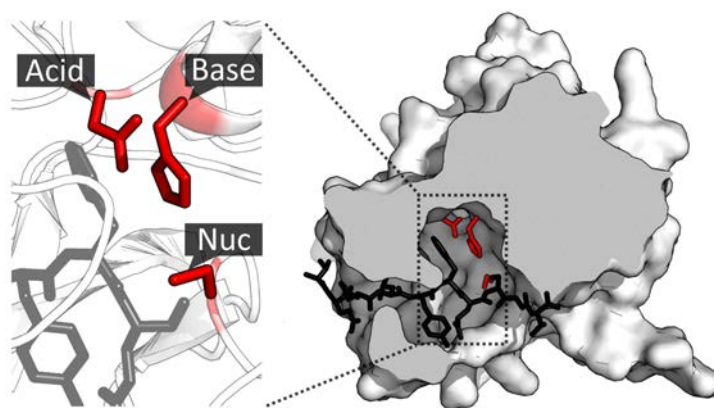


Figura 1 Sitio activo y triada catalítica de la enzima Serine Proteasa TEV (PDB: 1lvm). Los residuos aspartato (Acido), histidina (base) y serina (nucleófilo) están destacados en rojo, mientras que el sustrato se encuentra en negro. Adaptado de Shafee (2014).

aminoacídica modificaron las propiedades fisicoquímicas de la cadena polipeptídica, lo cual influye directamente en la estructura que estas últimas pueden adoptar y, por ende, la reacción química que estas catalizan. A este proceso se le denomina “evolución molecular” y para entenderlo es necesario conocer como evolucionan tanto la secuencia aminoacídica como la estructura proteica.

Ancestro común

A pesar que el ADN es quien se ve afectado por mutaciones (ver cuadro Evolución y ADN), estas pueden o no tener un efecto sobre la secuencia aminoacídica, lo cual no necesariamente afecta la función proteica; ya que esta última depende de un bajo número de residuos críticos entre los que destacan lo que conforman el sitio activo, el sitio de unión a sustrato o que contribuyen a la estabilidad de la estructura tridimensional. Debido a su importancia, estos residuos presentan una baja variabilidad ya que cumplen un rol relevante para mantener la función, a diferencia de otros en los que su cambio por otro residuo – de igual o distinta propiedad fisicoquímica – no produce un efecto significativo en el arreglo tridimensional de la cadena polipeptídica y en la función que ésta desarrolla. El hecho que las mutaciones pueden ocurrir sin perturbar en gran medida las propiedades fisicoquímicas que definen la estructura secundaria y terciaria de una cadena proteica, tiene como consecuencia que las estructuras evolucionan a un ritmo diferente de las secuencias de las que derivan, lo cual se traduce en que las estructuras están más conservadas que las secuencias.

Hoy en día, gracias a los métodos experimentales y teóricos disponibles, es posible observar estos procesos evolutivos a nivel molecular, la llamada “evolución molecular”. En este ámbito, el concepto de **evolución convergente** implica que NO hay una relación ancestral cercana o apreciable, solo una convergencia en un arreglo estructural estable, mientras que **evolución divergente** si implica una relación ancestral cercana, en la cual los cambios no se manifiestan en variaciones de gran magnitud de la estructura.

“La evolución selecciona la estructura proteica, NO la secuencia aminoácídica, ya que la estructura es la que determina la función”

De esta forma, es posible encontrar proteínas que no tienen una secuencia similar o función relacionada, pero que si comparten una estructura tridimensional similar. Esta diferencia en la velocidad evolutiva entre las secuencias y la estructura hacen que esta última sea un mejor marcador evolutivo, ya que el estudio de sus cambios permite reconocer mejor ancestros relacionados

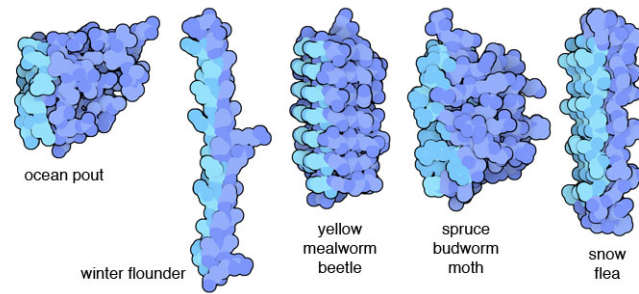


Figura 2 Proteínas anticongelantes (AFP). InterPro:IPR003460,SUPERFAMILY 1ezg,Pfam PF02420SCOP 1ezg, Pfam structures.

y determinar si la diferenciación entre las proteínas de interés proviene de procesos de evolución divergente o convergente.

Evolución Convergente

La divergencia y convergencia describen procesos evolutivos por los cuales los organismos se adaptan a sus ambientes. Es difícil encontrar casos representativos de episodios de evolución convergente. En general hablamos de evolución convergente cuando organismos no cercanamente relacionados entre sí han desarrollado características similares para adaptarse a un ambiente similar. Cabe destacar que para cada problema hay infinitas soluciones posibles, sin embargo estas emergen una y otra vez de manera independiente, aunque no siempre a través de un mecanismo similar. Para ejemplificar observe lo siguiente, los peces que habitan en la Antártica han evolucionado para sobrevivir en ambientes de baja temperatura, para ello utilizan glicoproteínas que les permiten evitar el crecimiento de cristales de agua y disminuir el punto de congelación de sus fluidos corporales.¹. Al mismo tiempo, al otro lado de la tierra los peces del Ártico también presentan proteínas que actúan de igual manera similar para sobrevivir a las frías condiciones ambientales. Es evidente que estos peces tienen relaciones evolutivas distantes, lo cual se traduce en que los genes de estas proteínas “anticongelantes” provienen de genes distintos y proteínas de estructura diferente (Figura 2). Esta evidencia sugiere episodios de “evolución molecular” independientes, que resultaron en funciones similares ya que ambos peces, Antárticos y Árticos, evolucionaron en nichos ambientales de similares características.

Evolución divergente

Las Serine proteasas están presentes virtualmente en todos los organismos, cumpliendo funciones intra e extracelulares. En general se describen como dos familias: “Trypsin-like” y “subtilisin-like” (traducido como parecidas a tripsina o a subtilisina). Se describen como Serine proteasas por dos razones principales: hidrolizan proteínas y tienen un residuo de Serina en el sitio activo, esencial para realizar la catálisis. Esta serina es significativamente más reactiva que otras serinas de la proteína. En mamíferos, las Serine proteasas conforman una familia de proteínas con diversas funciones, pero de estructura similar. Ejemplos de ella son:²:

¹Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod PNAS 1997 94 (8) 3817-3822

²se han conservado sus nombres en inglés

- Chymotrypsin
- Trypsin
- Elastase
- Plasmin
- Thrombin
- Acrosomal protease
- Complement C1
- Keratinase, Collagenase, Fibrinolysin, Cocoonase, etc.

En este práctico Ud. deberá resolver distintas actividades y preguntas, las cuales aparecerán a lo largo del texto. Para diferenciar las actividades de las preguntas, éstas aparecerán como una letra destacada con un círculo rojo, similares a las que aparecen a continuación y que Ud. debe constestar.

P.1 Traduzca los nombres de las serina proteasas descritas anteriormente.

P.2 Describa su función de forma general.

Como fue mencionado anteriormente, las serine proteasas tienen una estructura tridimensional similar³, sin embargo, tienen diferencias en su superficie – reflejo de sus diferentes secuencias – para mejorar la interacción con los sustratos involucrados en las diferentes actividades fisiológicas en que participan. A pesar de estas diferencias **todas ellas comparten un mismo mecanismo catalítico**. Las diferencias entre sus secuencias y la similitud entre sus estructuras indican sus relación evolutiva. La alta similitud entre chymotrypsin, trypsin y elastase indican que estas proteínas evolucionaron por de eventos de duplicación génica, a partir de una Serine proteasa ancestral, donde sus funciones fueron definidas a través de un proceso de **evolución divergente**. Por otro lado, en la Serine Proteasa Subtilisina (aislada de *Bacillus subtilis*, de secuencia y estructura diferente a la chymotrypsina) los grupos del sitio activo son similares a los descritos para Chymotripsina, sin embargo su posición tridimensional no lo es. De esta manera, Chymotripsina y Subtilisina son ejemplos claros de un proceso de evolución convergente (ver figura 1).

P.3 Explique con sus palabras los conceptos de evolución convergente y divergente

P.4 Investigue si hemoglobina, glutathion peroxidasa y cytochrome c son ejemplos de evolución convergente o divergente.

Cuando se investigan las relaciones evolutivas entre diferentes organismo, es luego importante observar tanto los genes como las proteínas que los diferencian.

En este práctico utilizaremos diferentes herramientas diseñadas para estudiar la evolución de los mecanismo enzimáticos a través de la información genética disponible, como la conservación evolutiva, en conjunto con la información estructural de las proteínas que estos genes codifican, lo cual es una herramienta útil al momento de entender los aspectos evolutivos que llevan a la función proteica y como esta puede ser modulada.

Homología y similitud

El concepto de homología – presencia de un ancestro evolutivo común – se encuentra mencionado en múltiples ocasiones en este texto, siendo central para el análisis computacional de proteínas y secuencias nucleotídicas. Sin embargo, la relación entre homología y similitud no es tan clara y atención debe prestarse a ella. En términos analíticos, decimos que existe homología cuando el grado de similitud entre dos elementos, ya sea secuencias o estructuras, excede el valor esperado para

³lo cual aun no hemos definido

elementos de la misma naturaleza (secuencias aleatorias de nucleótidos, aminoácidos de un mismo largo) elegidos al azar. Cuando existe un exceso de similitud mayor al esperado, la explicación más simple (parsimoniosa) es que ambas secuencias no fueron generadas de forma independiente, sino que provienen de un ancestro en común. Este ancestro en común permite explicar el exceso de similitud, ya que otras explicaciones requieren que las estructuras proteicas resultante de la expresión de estos genes evolucionen de manera independiente.

Sin embargo, secuencias homologas no siempre no siempre significan similitud de secuencias. Los antecedentes disponibles muestran que existen alineamientos de secuencias que no son significantes, pero que al mirar las estructuras proteicas utilizando alineamientos estructurales éstas presentan una similitud estadísticamente significativa, ya sea entre ellas o hacia una secuencia intermedia.

P.5 Discuta el concepto de “homología” bajo el contexto de evolución convergente y divergente.

Actividades a Realizar

La disciplina denominada “enzyme evolution” involucra la aplicación de conceptos bioinformáticos en al área de la enzimología para la describir el proceso de evolución de la estructura y función. En esta actividad práctica Ud.

- Investigará los distintos niveles de la comisión de enzimas, desde el punto de vista de las estructuras cristalinas que hay en cada una de ellas.
- En la base de datos UNIPROT, investigará relevante información sobre proteínas y enzimas.
- Utilizará PROSITE, Pfam e InterPro para encontrar proteínas homólogas a partir de secuencias aminoacídicas.
- Utilizará alineamientos múltiples para encontrar proteínas homólogas a partir de perfiles obtenidos de secuencias aminoacídicas.
- Utilizará los datos e información recopilada para plantear un experimento relacionado con la evolución de enzimas.
- Utilizará VMD y sus herramientas para relacionar la estructura y función con los conceptos de evolución convergente y divergente.

Comprenda y realice todas las instrucciones (destacas con números arábigos), conteste las distintas preguntas (destacadas con letras en mayusculas), y entregue un informe escrito con los resultados y discusión de los temas abordados. A lo largo del texto se entregará material complementario en forma de “cuadros” independientes, los que incluyen información y preguntas de distintos aspectos como importancia biológica, parámetros a profundizar y atajos para VMD.



Evolución y ADN. El ADN esta siempre sujeto a mecanismos de reparación que permiten la mantención del código genético, sin embargo, los sistemas de reparación están sujetos a fallas. Si consideramos que el ser humano tiene una tasa de error de $1 \cdot 10^{-8}$ mutaciones por nucleótido por generación y que el genoma humano tiene alrededor de $3 \cdot 10^9$ bases, entonces cada persona tiene en promedio alrededor de 70 mutaciones en su genoma (mutaciones por generación). Esto es una diferencia promedio de 0.1 % en los genoma de dos personas cualquiera (1 de cada 1000 bases).

Programas Requeridos

Los siguientes programas serán requeridos en este práctico:

- **VMD**⁴ (para todas las plataformas)
- **NAMD**⁵
- **Jalview**⁶
- **Programa de graficación matemática:** Será necesario utilizar programas para graficar las salidas de VMD y NAMD. VMD tiene un programa básico de graficación incorporado. Algunos ejemplos de otros programas son:
 - Unix/Linux: xmgrace ⁷
 - Windows: Excel ⁸ (Pago)
 - Mac/Multiples Plataformas: Mathematica ⁹ (Pago); gnuplot ¹⁰(Descarga gratuita)

Los archivos de este práctico se encuentran en el directorio `practico-evolucion`.

2. Clasificación funcional y estructural de enzimas

Nomenclatura y clasificación de Enzimas

Las enzimas se clasifican de acuerdo a la nomenclatura asignada por el Nomenclature Committee de la International Union of Biochemistry (descrita por primera vez en 1961, mejorada por última vez en 1992). Bajo esta nomenclatura, las enzimas **NO** son clasificadas por su su mecanismo, secuencia aminoacídica (ie. homología) o por la estructura que estas adoptan, **sino a través de la reacción química que estas catalizan**. La llamada *enzyme commission* clasifica las enzimas en 6 grupos o clases principales, de acuerdo a la reacción que estas catalizan:

- | | |
|-------------------------|--------------------|
| ■ EC 1. Oxidoreductasas | ■ EC 4. Liasas |
| ■ EC 2. Transferasas | ■ EC 5. Isomerasas |
| ■ EC 3. Hidrolasas | ■ EC 6. Ligasas |

P.1 A partir de los datos de la página de la Enzyme comission¹¹ describa de manera general las clases en las cuales se agrupan las enzimas y las reacciones que estas catalizan.

La enzyme commission asigna a cada enzima un código recomendado (Enzyme classification, EC) que contiene 4 cuatro secciones o números a,b,c y d; “a” es la clase, “b” la subclase y “c” es la sub-subclase. Mientras “b” y “c” describen la reacción, “d” es utilizada para distinguir enzimas que tienen una función similar sobre el sustrato de la reacción.

⁴<http://www.ks.uiuc.edu/Research/vmd/>

⁵<http://www.ks.uiuc.edu/Research/namd/>

⁶<http://www.>

⁷<http://plasma-gate.weizmann.ac.il/Grace/>

⁸<http://office.microsoft.com>

⁹<http://www.wolfram.com/>

¹⁰<http://www.gnuplot.info/>

¹¹<http://www.chem.qmul.ac.uk/iupac/jcban/>

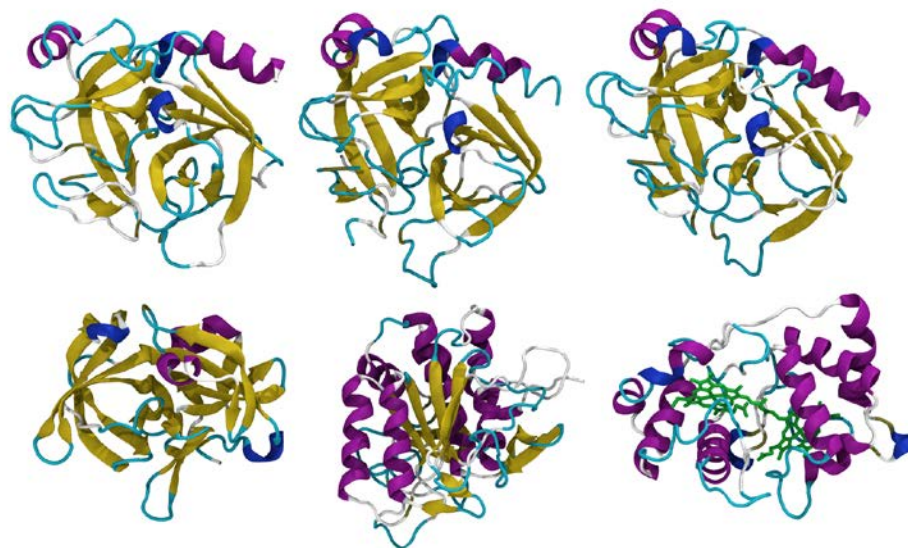


Figura 3 Estructura de cuatro Serine Proteasas y dos proteínas no relacionadas. (A) Tripsina Bovina (pdb 3OTJ) (B) Chymotripsina bovina (pdb 1YPH) (C) Tripsina de *S. griseous* (pdb 1SGT) (D) Proteasa A *S. griseous* (E) Subtilisina (pdb 1SBT) (F) Citocromo c4 (pdb 1ETP).

Por ejemplo, las “Serine proteasas” (también llamadas serine endopeptidasas, Referenciasscb:SerProt) son enzimas capaces de romper los enlaces peptídicos de proteínas. Su nombre se debe a que en todas ellas, un residuo conservado de Serina sirve como nucleófilo en el sitio activo. Además, todas las serine proteasas contienen tres residuos en su sitio activo: Serina, histidina y aspartato. Habitualmente, para identificar los residuos de una proteína se utiliza la numeración de los residuos de un miembro característico de una familia; en el caso de las serine proteasas es habitual utilizar el esquema de numeración de la Chymotripsina, por ejemplo la Chymotrypsin-C Humana¹² (EC:3.4.21.2), donde los residuos que componen la llamada “**triada catalítica**” son His⁷⁴, Asp¹²¹ y Ser²¹⁶, (figura 1)¹³. Estos residuos, también llamados “residuos catalíticos”, se encuentran alejados en la secuencia aminoacídica (cuadro 4), pero cercanos espacialmente en la conformación tridimensional que adopta esta proteína (Figura 1).

Las serines proteasas pertenecen a la subsubclase EC 3.4.21. El cuadro 2 señala 5 de los 121 miembros entre los descritos para esta subsubclase de enzimas según la enzyme commission.

P.2 A partir de la clasificación de la enzyme commission, identifique y justifique cada clase, subclase y subsubclase hasta encontrar el número identificador EC de la enzimas **Serine Proteasas** listadas anteriormente en el cuadro 2.

Recuerde que esta jerarquización esta realizada en base a la reacción catalizada por la enzima. De esta forma cuando dos enzimas que pertenecen a una misma subsubclase, el número final “EC a.b.c.d” o identificador esta relacionado con los metabolitos y cofactores involucrados en la reacción.

P.3 A partir de la clasificación de la enzyme commission, identifique y diferencie las reacciones de la enzimas de las subsubclases listadas en el cuadro 2.

¹²<http://www.uniprot.org/uniprot/Q99895>

¹³Thomas, Shafee, (2014). “Evolvability of a viral protease: experimental evolution of catalysis, robustness and specificity”. PhD Thesis. University of Cambridge.

Cuadro 1 Diferentes Serine proteasa y su triada catalítica.

[illegible]


La tabla 3 se encuentran los diferentes códigos que permiten acceder a información y a las secuencias de diferentes miembros de estas familias a partir del acceso a distintas bases de datos. Por ejemplo, los códigos de las estructuras cristalográficas (PDB id) y el organismo a cual pertenece cada secuencia.

P.4 A partir de lo aprendido e incluyendo la información del organismo del cual provienen las proteínas estudiadas discuta sobre el proceso de evolución de las serine proteasas.

Uniprot

La base de datos Uniprot¹⁴ provee anotación detallada de secuencias biológicas, incluyendo: estructura, función, clasificación en familias de proteínas, dominios estructurales, sitios catalíticos, cofactores, modificaciones postraduccionales, vías metabólicas, asociación a enfermedades. También provee links a otros recursos de interés, y es muy poco redundante. La información de secuencias deriva de TrEMBL, una base de datos de secuencias aminoacídicas traducidas desde secuencias de



 **Serine Proteasas.** Cerca de un tercio de las proteínas conocidas pertenecen a la familia de las Serine proteasas. Entre ellas las tripsinas son las más predominantes, con funciones que participan en la digestión, coagulación, fibrinólisis, desarrollo, fertilización, apoptosis e inmunidad, entre otras.^a Por otro lado, cabe destacar que los residuos que componen la triada catalítica son “altamente” superponibles entre diferentes serine proteasas. ¿Que se puede inferir de esto?

^aDi Cera, Serine Proteases. *Life*. 2009 May; 61(5): 510–515.

¹⁴<http://www.uniprot.org/>

Cuadro 2 Ejemplo de miembros de la subsubclase EC Serine Proteasas. Se resalta los residuos catalíticos de la Chymotripsina C humana.

- Trypsin-like
- Chymotrypsin-like
- Thrombin-like
- Elastase-like
- Subtilisin-like

Cuadro 3 Miembros de la familia de Serine Proteasas

PDB-Id	cadena	residuos	resolución	largo	PM	Fuente
4H4F	A	249				Homo sapiens
3OTJ	E	281		223	23324.3	Bos taurus
1YPH	C	482	1.34	131	13934.6	Bos taurus
1SGT	A	223	1.7	223	23076.8	Streptomyces griseus
2SGA	A	181	1.5	181	18016.6	Streptomyces griseus
1SBT	A	275	2.5	275	27552.5	Bacillus amyloliquefaciens

PDB-Id	uniprot Id	Nombre	cath	cathId	pfam
4H4F	Q99895	Chymotrypsin-C			
3OTJ	P00760	Cationic trypsin	Trypsin-like SP	2.40.10.10	PF00089
1YPH	P00766	Chymotrypsinogen A	Trypsin-like SP	2.40.10.10	PF00089
1SGT	P00775	Trypsin	Trypsin-like SP	2.40.10.10	PF00089
2SGA	P00776	Streptogrisin-A	Trypsin-like SP	2.40.10.10	PF00089
1SBT	P00782	Subtilisin BPN'		3.40.50.200	PF00082

ácidos nucleicos. La anotación de cada entrada es cuidadosamente curada por expertos que obtienen información de la literatura científica, contrastando diferentes datos experimentales y, por tanto, es de buena calidad.

A continuación buscaremos información biológica sobre las proteínas asociadas a las secuencias utilizadas para generar el alineamiento mostrado en 4y las estructuras de la figura 3. SerineProtsCatTriad, con el fin de vincular el proceso de evolución con la función.

- 1 Ingrese y describa que es Uniprot
- 2 Ingrese en el cuadro buscar “Q9985”

La base de datos de Uniprot contiene información curada sobre diferentes proteínas y es considerada como base de información para múltiples sitios y estudios.

P.5 ¿Describa a manera general la información biológica que puede obtener en UNIPROT?.

Para simplificar el trabajo a realizar, complete la siguiente tabla para los códigos que aparecen en la tabla 3

- Código UNIPROT
- Nombre
- Organismo
- Función
- Código PDB
- Código Pfam
- Código Supfam
- Código EC
- Secuencia aminoacídica
- Dominios
- otros datos que considere relevantes
- residuos del sitio activo

Entry	Entry name	Protein name	Gene name	Organism	Length	Temperature dependence	EC number
3OTJ	TRY1_BOVIN	Cationic trypsin		Bos taurus (Bovine)	246		3.4.21.4
3OTJ	BPT1_BOVIN	Pancreatic trypsin inhibitor		Bos taurus (Bovine)	100		
1YPH	CTRA_BOVIN	Chymotrypsinogen A		Bos taurus (Bovine)	245		3.4.21.1
1SGT	TRYP_STRGR	Trypsin	sprT	Streptomyces griseus	259		3.4.21.4
2SGA	PRTA_STRGR	Streptogrisin-A	sprA	Streptomyces griseus	297		3.4.21.80
1S8T	SUBT_BACAM	Subtilisin BPN'	apr	Bacillus amyloliquefaciens (Bacillus velezensis)	382	Optimum temperature is 48 degrees Celsius. 1 Publication	3.4.21.62
4H4F	ICIC_HIRME	Eglin C		Hirudo medicinalis (Medicinal leech)	70		
4H4F	CTRC_HUMAN	Chymotrypsin-C	CTRC CLCR	Homo sapiens (Human)	268		3.4.21.2

Figura 4 Información disponible en Uniprot a partir de los códigos pdb de serine proteasas. Adicionalmente se puede generar un alineamiento (1) de las secuencias seleccionadas, como también descargar las secuencias (2) en formato fasta.

Uniprot permite la búsqueda masiva y, además, realizar conversiones entre los códigos de una u otra base de datos. A continuación buscaremos la información relevante de las Serine proteasas que hemos trabajado.

3 En la pestaña titulada “Retrieve” ingrese los códigos PDB de las proteínas descritas en la tabla 3. Luego presione el botón “Retrieve” y siga el link que dice: “UniProtKB” .

Esto le dará acceso directo y rápido a las páginas en UniProt de cada una de las 6 proteínas que estamos estudiando (figura 4).

P.6 Describa cada una de sus proteínas. Utilice la tabla de descripción utilizada anteriormente.

P.7 Compruebe los datos de la tabla 3.

Además, la página de resultados le permite realizar estudio de las secuencias seleccionadas a través del uso de alineamientos múltiples

4 Utilice el botón “Align” (figura 4, botón 1) para generar un alineamiento múltiple con las secuencias de las proteínas seleccionadas.

En la columna “Highlights”, se indican otras capas de información que podemos utilizar para colorear el alineamiento.

P.8 Utilice los campos, Similarity, identity y beta strands para observar como estos parámetros se conservan a través de las secuencias.

A simple vista y de acuerdo con lo observado en la figura 1, en este alineamiento realizado en UNIPROT se observa que hay secuencias que no conservan residuos entre ellas, por ejemplo, Subtilisina. Sin embargo, todas estas proteínas se encuentran descritas por la Enzyme commission como Serine Proteasas.

5 Realice el alineamiento sin considerar a Subtilisina, elimine otras secuencias hasta mejorar el alineamiento

P.9 ¿Por que cambia el alineamiento cada vez que elimina una secuencia, se obtiene un resultado similar al eliminar una secuencia diferente de subtilisina?

P.10 A partir de lo aprendido e incluyendo la información recopilada de las proteínas estudiadas discuta sobre el proceso de evolución de las serine proteasas.

2.1. Caracterización basado en Secuencia

P.11 Con lo aprendido, discuta la siguiente frase “identificar la familia a la cual pertenece una secuencia a menudo permite inferir su funcionalidad”

A continuación utilizaremos diferentes recursos bioinformáticos para caracterizar la familia a la cual pertenece una proteína, para luego comparar estos resultados con los obtenidos anteriormente, con el fin de entender la relación entre secuencia, estructura y función.

PROSITE

PROSITE es una base de datos de familias de proteínas y dominios que utiliza modelos ocultos de Markovs (HMM, cuadro Perfiles y modelos ocultos de Markov) para encontrar proteínas homólogas en bases de datos de secuencias aminoacídicas. Su clasificación se basa en que a pesar que existe un gran número de proteínas, estas se agrupan en base a su similitud de secuencias en un número pequeño de grupos, de los cuales se pueden extraer patrones característicos para una familia de proteínas. De lo anterior se desprende que las proteínas o dominios que son caracterizados como un miembro de un grupo en particular, generalmente comparten atributos funcionales y derivan desde un ancestro en común.

1 Ingrese a Prosit¹⁵.

P.12 Describa que información que le entrega esta página

A partir de los códigos UNIPROT de las proteínas elegidas (Pista: Se puede ingresar un máximo de 10 códigos UniProt por búsqueda).

P.13 Incorpore a su lista las características de las siguientes proteínas (códigos PDB) representantes de las familias

- Glutathion-s transferase: 1AGS (Cadena A)
- G-proteins: 1AGP (Cadena A)

P.14 Registre el/los nombre(s) de los dominios y patrones encontrados para cada código UniProt.

P.15 ¿Cuántos patrones únicos diferentes identificó entre las secuencias? informe el nombre del patrón y la secuencia consenso para cada uno de ellos.

Pfam

Pfam es una base de datos de familias de dominios proteicos. Contiene información de grupos (familias) de proteínas homologas alineadas, sus anotaciones funcionales, y perfiles HMM extraídos de alineamientos de secuencias, los que pueden ser usados para clasificar proteínas en familias. Cada familia de Pfam consiste en un alineamiento revisado obtenido a partir de un set pequeño de secuencias, el que luego es extendido utilizando perfiles HMM en un nuevo alineamiento que contiene todas las secuencias de proteicas reconocibles a partir de una base de datos de secuencias primarias.

Cada entrada Pfam esta clasificada en seis grupos generales

- Familia: Colección de regiones de proteínas conservadas.
- Dominio: Una unidad estructural.

¹⁵<http://prosite.expasy.org/>

- Repetido: Una unidad de largo pequeño que es inestable cuando se encuentra aislado, pero forma una estructura estable cuando existe múltiples copas de éste.
- Motivo: Una unidad encontrada fuera de dominios globulares.
- Coiled-Coil: Regiones que contienen motivos coiled-coil motifs (α -helices que se condensan en paquetes de entre 2 a 7 unidades.
- Desordenada: Regiones conservadas que muestran o están predichas de contener secuencias que generan regiones intrínsecamente desordenadas.

Las entradas Pfam se agrupan en clanes, cuya relación esta definida a través de la similitud de secuencia, estructura o perfil-HMM.

2 Ingrese a Pfam ¹⁶ y en la sección titulada “search” busque cada una de las de las proteínas indicadas anteriormente. Para esto necesitará el código UNIPROT.

P.16 ¿Que información le entrega esta página?

P.17 ¿Como se clasifican las proteínas en ella?

P.18 ¿A qué familia(s) corresponde cada secuencia.

P.19 informe el nombre de la familia, dominios y al menos una porción del logo que define cada una de las familias

InterPro

Interpro provee herramientas de análisis de secuencias de proteínas, permitiendo su clasificación en familias y prediciendo la presencia de dominios y sitios importantes.

3 Vaya a InterPro ¹⁷ y busque nuevamente las secuencias indicadas.

P.20 Para cada proteína, vaya a la sección “Detailed signature matches”. Registre los patrones/familias identificadas por InterPro para cada una de las secuencias. (PISTA: Si pasa el mouse sobre cada fila con información en esta sección, se expandirá una ventana con el nombre completo del motivo/familia y la base de datos de donde proviene.

P.21 La información obtenida para Pfam y Prosite desde InterPro, ¿corresponde con aquella obtenida por usted desde esas bases de datos directamente en los puntos anteriores?

P.22 ¿Qué otras bases de datos, además de Pfam y Prosite, están contenidas en InterPro y que información aportan sobre las proteínas que usted investigó? Nombre las bases de datos.

HMMER

HMMER permite buscar secuencias homologas en bases de datos en secuencias homologas, además de realizar los alineamientos necesarios utilizando perfiles HMM. Habitualmente se utiliza en conjunto con bases de datos de perfiles como Pfam y aquellas que participan de Interpro. Su principal diferencia es que no es solo para obtener perfiles, sino que permite obtener información de secuencias problemas, al igual que BLAST. Entre sus herramientas para busqueda de muestras problemas están Phmmer o la herramienta interactiva jackhmmmer.

¹⁶<http://pfam.xfam.org/>

¹⁷www.ebi.ac.uk/interpro

4 Vaya a HMMER ¹⁸ y busque nuevamente las secuencias indicadas.

P.23 La información obtenida para Pfam y Prosite, InterPro y HMMER, ¿corresponde con aquella obtenida por usted desde esas bases de datos directamente en los puntos anteriores?

P.24 ¿Qué otras bases de datos, además de Pfam y Prosite, están contenidas en InterPro y aportan información para las proteínas que usted investigó? Nombre las bases de datos.



Perfiles y modelos ocultos de Markov. El método estadístico de modelado de datos de modelos ocultos de Markov ha sido utilizado en diferentes campos y, actualmente, permite producir perfiles de secuencia de mejor calidad que los métodos habituales. Los modelos de perfiles ocultos de Markov (HMM) tienen una base probabilística y cumplen con fundamentos que los relacionan de manera consistente con los puntajes asignables a variaciones, inserciones y deleciones. Estos modelos estiman la frecuencia verdadera de encontrar un residuo en una posición dada a partir de la frecuencia observada en el alineamiento, mientras que los métodos convencionales lo hacen solo a partir de la frecuencia observada. Esto significa que un perfil generado a partir de un alineamiento de 10 a 20 secuencias utilizando HMM, es equivalente al perfil obtenido de un alineamiento clásico utilizando 40 o más secuencias. Lo cual permite inferir relaciones evolutivas a partir de un conjunto de datos más reducido o de menor cercanía evolutiva.

A continuación utilizaremos alineamientos múltiples para identificar patrones en secuencias problemáticas. El objetivo es asignar una familia a cada secuencia, a partir de la cual podemos inferir un posible plegamiento y función.

5 Realice la búsqueda utilizando la secuencia de la Chymotripsina-C humana, tripsina ácida, subtilisina, una Glutathion-S transferase y una proteína-G en UNIPROT, PROSITE, pfam e InterPro.

P.25 ¿Son los resultados entregados equivalentes al buscar por código que por secuencia?

P.26 ¿Que puede inferir de las relaciones entre estas proteínas con los resultados entregados?

En el caso que Ud. deba que caracterizar una muestra problema,

P.27 Genere una hipótesis que pueda resolver utilizando la información contenida en los diferentes sitios observados

P.28 ¿Cuál programa o base de datos utilizaría?

Tasas de mutaciones y caracterización funcional

Para describir una secuencia, habitualmente se buscan patrones conservados a través de alineamientos, de los cuales inferimos la función ya que hemos descrito que ambos están relacionados. Esta relación se debe a que, en general, la tasa de mutación en organismos unicelulares eucariotas y bacterianos es cercano a 0.003 mutaciones por genoma por generación celular, lo que sugiere que el humano acumula cerca de 64 nuevas mutaciones por generación, ya que en cada nueva generación

¹⁸<https://www.ebi.ac.uk/Tools/hmmer/>

implica una nueva división de gametos.¹⁹ Cuando estas mutaciones permanecen, las secuencias van siendo modificadas paulatinamente, disminuyendo la similitud de las secuencias y posiblemente la función proteíca.

Utilizaremos las herramientas bioinformáticas ya estudiadas para simular el proceso de evolución de las Serine Proteasas e inferir cual es la relación entre similitud de secuencia y la capacidad de detección de la función proteíca.

6 Vaya a Seqvolver ²⁰

7 Ingrese en el campo de búsqueda la secuencia de una Serine Proteasa en formato FASTA

Para simular los cambios evolutivos haremos uso de la matriz de puntuación PAM, la cual permite describir la evolución en términos de unidades llamadas PAM, que representan una cantidad de cambios evolutivos como un promedio de cambios de residuos por generación. De esta forma, la unidad 1-PAM simboliza un proceso de evolución en el cual el 1 % de los residuos han cambiado, 2-PAM el 2 % de residuos y así sucesivamente. Cabe destacar que un mayor número de unidades PAM no necesariamente promueve un cambio mayor en la identidad de secuencia, ya que los cambios anteriores pueden ser revertidos por los cambios subsiguientes.

8 Ingrese en los campos “rate” y “Time” valores que su multiplicación de como resultado el porcentaje de cambio a simular. Por ejemplo, para calcular un cambio de 40-PAM, se deben colocar valores que multiplén 40. Ya sea 40 cambios por generación y un tiempo 1 (una generación) o 1 cambio por generación en 40 generaciones.

9 Una vez seleccionada la distancia evolutiva a analizar, seleccione el botón “evolve”

Se abrirá una nueva ventana con la secuencia con la distancia evolutiva, en unidades PAM, solicitada.

Si Ud. utiliza una tasa de cambio de 1 o 10 unidades PAM, esta secuencia debe ser clasificada como un nuevo miembro de la familia de las Serine proteasas, la cual corresponderá a una secuencia homóloga cercana de su proteína.

P.29 Utilizando las herramientas de este práctico compruebe que proteínas de baja distancia evolutiva entre ellas son reconocibles como miembros de una misma familia.

P.30 Que sucederá al aumentar la distancia evolutiva entre estas proteínas, por ejemplo utilizando 20,30,40,50,60,70,80 y 90 unidades PAM, cambiará el porcentaje de identidad? el porcentaje de similitud?

P.31 ¿Que cambios deben ocurrir para que la distancia evolutiva no permita reconocer los homólogos cercanos?

2.2. Relación secuencia-función

Los análisis realizados muestran que un aumento en la distancia evolutiva entre dos proteínas puede significar la modificación o pérdida de los patrones o motivos en las secuencias. Si consideramos que estos son utilizados por los diferentes programas para clasificar una secuencia como miembro de una familia – y subsecuentemente su función –, es inherente generar la siguiente hipótesis:

“Un bajo porcentaje de identidad de secuencia aminoacídica no permite inferir la función”

Para finalizar,

¹⁹Rates of spontaneous mutation". Genetics. 148 (4): 1667–86

²⁰<http://fasta.bioch.virginia.edu/seqevolver/>

1 Diseñe un experimento que le permita, con las herramientas de este práctico, corroborar o refutar la hipótesis planteada.

P.32 En base a sus resultados, ¿qué relación debe existir entre la secuencia, estructura y la función?. Genere una hipótesis de trabajo.