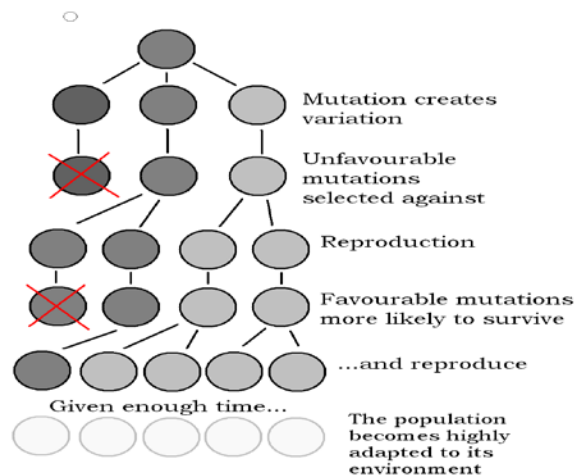


Evolución Molecular



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 de 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 está 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 está expuesto a solvente, este interrumpe la red de enlaces de hidrógeno y restringe 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 está 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 plantea 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

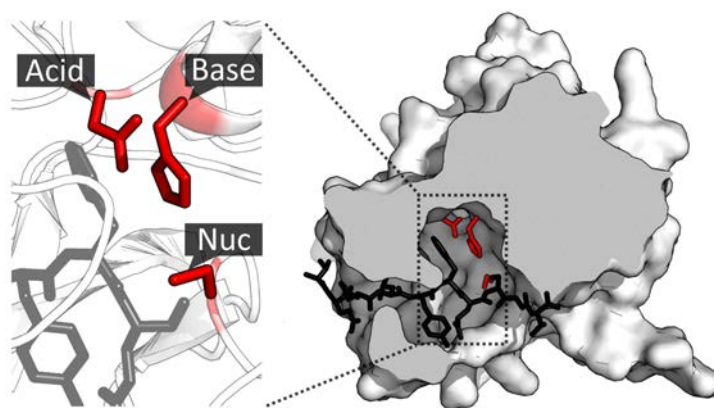


Figura 1. Sitio activo y triada catalítica de la enzima Serine Proteasa TEV (PDB: 1lvm). Los residuos aspartato (Acido), histidina (base) y serina (nucleófilo) están destacados en rojo, mientras que el sustrato se encuentra en negro. Adaptado de Shafee (2014).

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

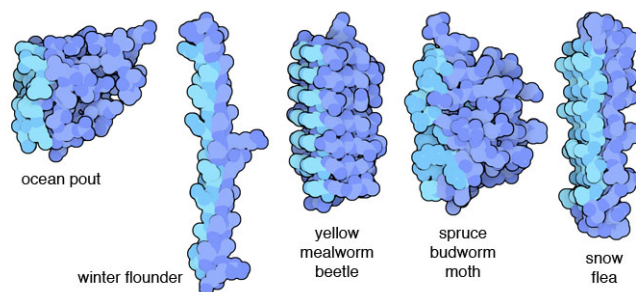


Figura 2. Proteínas anticongelantes (AFP). InterPro:IPR003460,SUPERFAMILY 1ezg,Pfam PF02420SCOP 1ezg, Pfam structures.

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

- | | |
|----------------|----------------------|
| ■ Chymotrypsin | ■ Plasmin |
| ■ Trypsin | ■ Thrombin |
| ■ Elastase | ■ Acrosomal protease |

¹PNAS 1997 94 (8) 3817-3822

²se han conservado sus nombres en inglés

- 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.



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 nucleotido 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).

³lo cual aun no hemos definido

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 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.

Programas Requeridos

Los siguientes programas serán requeridos en este práctico:

- **VMD**⁴ (para todas las plataformas)

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

- NAMD⁵
- Jalview⁶
- **Programa de graficación matemática:** Será necesario utilizar programas para gráficar 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.

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

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

⁶<http://www.jalview.org/>

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/>

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

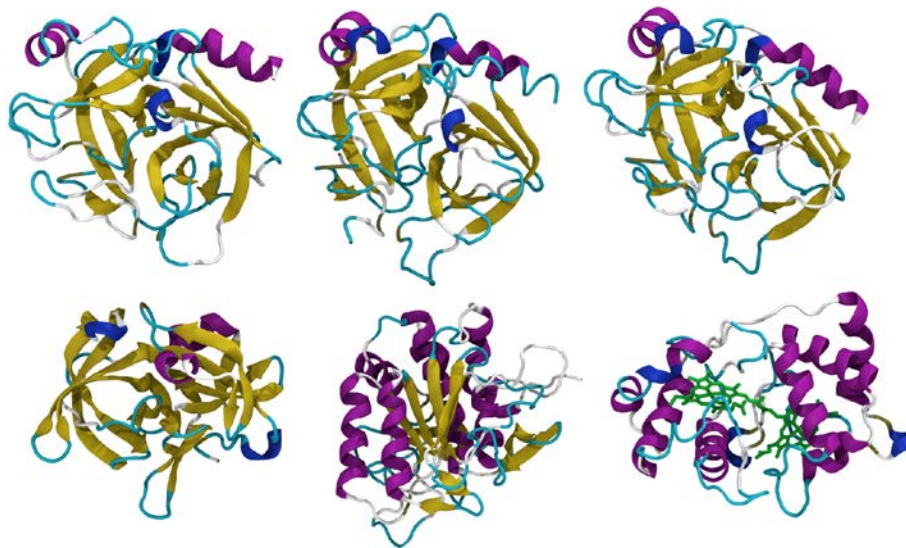


Figura 3. Estructura de cuatro Serine Proteasas y dos proteínas no relacionadas. (A) Tripsina Bovina (pdb 3OTJ) (B) Chymotripsina bovina (pdb 1YPH) (C) Tripsina de *S. griseous* (pdb 1SGT) (D) Proteasa A *S. griseous* (E) Subtilisina (pdb 1SBT) (F) Citocromo c4 (pdb 1ETP).

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).



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.

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.

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

Tabla 1. Diferentes Serine proteasa y su triada catalítica.

	60	70	80	90	120
Chymotrypsin	G..GTLIASNFV	LTAAHCIS..N	TRTYRVA	VGKNN	ALLLRNDIALIKLA
Cationic	G..GSLINSQW	VVSAAHCL...	YKSGIQVRL	GED	SNTLNNDIMLIKLK
Chymotrypsinogen	G..GSLINENW	VVTAAHCL...	GVTTSDEVV	VAGEF	SLTINNDITLLKLS
Trypsin	G..GALYAQDI	LTAAHCVSGSG	NNISITATG	GG..	G..TGKDWALIKLA
Streptogrisin-A	..YTEASTGKI	VLTAADSTVSK	AELAKVSNAL	LAGSK	SLGFNVSVNGVAHA
Subtilisina	GTVAALNNSIG	VLGVAPS....	ASLYAVK	VLGA.NMDVINMSLG

	130	200	210	220
Chymotrypsin	EHVELSD	MVCAG...GD	GV I.....SACNGDSGGPLNC
Cationic	SAASLNS	MFCAGYL.EG	GK.....DSCQGDSGGPVVC
Chymotrypsinogen	TAASFSQ	MCAG...AS	GV.....SSCMGDSGGPLVC
TrypsinQ	EICAGYPDTG	GV.....DTCQGDSGGPMFR
Streptogrisin-A	LTAGHCT	TVNYG...SS	GLVYGMQITNV	CAE..PGDSGGSLFA
SubtilisinaG	VIAGV.....	AV.....DSSNQ	RASFSSVVGPELD

Tabla 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

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.

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

¹⁴<http://www.uniprot.org/>

Tabla 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

Entry	Entry name	Protein name	Gene names	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
1SBT	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.

proporciona 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 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 3ref: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, busque y guarde los siguientes descriptores para las proteínas 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

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 Markov (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 Prosite ¹⁵.

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.
- 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: Regions conservadas que muestran o están predichas de contener secuencias que generan regiones intrinsecamente desordenadas.

¹⁵<http://prosite.expasy.org/>

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 Pfam¹⁶ 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 homólogas 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 jackhmmer.

- 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.

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

¹⁶<http://pfam.xfam.org/>

¹⁷www.ebi.ac.uk/interpro

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

- 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 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 proteica.

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 proteica.

6 Vaya a Seqolver ²⁰

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



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.

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

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

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 multipliquen 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 comparada con la secuencia original?

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,

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, ¿que relación debe existir entre la secuencia, estructura y la función?. Genere una hipótesis de trabajo.

2.3. Alineamientos múltiples

La mayor parte de los programas utilizados en este práctico usan alineamientos de múltiples secuencias para asignar un grupo o característica a una muestra problema. Los alineamientos múltiples permiten inferir las relaciones evolutivas en base a los patrones de secuencias y motivos conservados que existan entre 3 o más secuencias biológicas. En la página web de EMBL-EBI ²¹ podemos encontrar varios servicios que permiten realizar alineamiento múltiple de secuencias, entre los cuales están ClustalO, MAFFT, MUSCLE, T-coffe, entre otros, hoy en día.

Los alineamientos múltiples de secuencias (MSA, por sus siglas en inglés) se pueden usar para **identificar** secuencias relacionadas, mediante la construcción de modelos estadísticos como son las matrices de puntuación de posiciones específicas (PSSMs), los perfiles y los modelos ocultos de Markov (HMMs, ver cuadro Perfiles y modelos ocultos de Markov).

Dentro de este ámbito, el análisis de perfiles busca de describir consensos entre múltiples secuencias, asignando un puntaje específico para cada posición, lo cual permite capturar la información de conservación de cada residuo y establecer relaciones incluso en secuencias que a simple vista no están relacionadas. Esto los convierte en un método más sensible y específico que otros métodos basados en alineamientos como Blast y Fasta, que usan funciones de jerarquización basados en propiedades generales de las secuencias y que no están basados en la posición que tiene cada residuo en la secuencia (ver cuadro Perfiles de secuencia).



Perfiles de secuencia. A menudo los alineamientos múltiples de secuencias tienen huecos o gaps de longitud variable. Cuando las penalizaciones por abrir y extender gaps son incluidas en la construcción de las matrices de PSSM entonces se está construyendo un perfil. En otras palabras, un perfil es una PSSM con información de penalización de inserciones y eliminaciones específico para las proteínas analizadas, que habitualmente pertenecen a una familia de secuencias. Los perfiles pueden ser usados en las búsquedas en bases de datos para encontrar secuencias homólogas remotas, o dicho de otra forma, para detectar relaciones distantes entre secuencias proteicas. Estas relaciones distantes sirven para plantear hipótesis biológicas, sobre la base de la relación entre secuencia, estructura y función, ya que las secuencias biológicas funcionales típicamente vienen en “familias” que a lo largo de la evolución han mantenido la misma función o una muy relacionada.

Cuando realizamos búsquedas de secuencia utilizando BLAST a menudo sucede que no se encuentran secuencias con un alto grado de similitud, sin embargo, esto no significa que una parte o sección de la secuencia no este conservada entre los miembros de una misma familia. De esta forma, un MSA puede revelar la existencia de patrones comunes a un grupo de secuencias determinado. Al estudiar las secuencias de diferentes familias de proteínas, como las Serine Proteasas, es evidente que existen regiones que presentan mayor grado de conservación que otras. Estas regiones y residuos son generalmente importantes ya que cumplen un rol en la mantención de la función, ya sea a través de mantener las propiedades de superficie necesarias para unir un sustrato o interactuar con otras proteínas, o directamente por ser relevantes para la estabilización de la estructura tridimensional necesaria para posicionar los residuos del sitio activo.

Esta conservación de pequeños motivos suele tomarse como indicativo de que existe homología remota, ya que estos patrones actúan como firmas funcionales de las proteínas (Cabe destacar que una proteína puede tener más de un motivo presente en su secuencia) y su existencia permite establecer, a través de una búsqueda, si una secuencia pertenece o no a una familia caracterizada

²¹<https://www.ebi.ac.uk>

por presentar el mismo patrón. De esta forma estos patrones caracterizan una “huella dactilar” de las proteínas y pueden ser utilizados para asignar una familia a una secuencia de una proteína desconocida y formular hipótesis acerca de ella.

Los modelos HMM (ver cuadro Perfiles y modelos ocultos de Markov) reflejan información acerca de la frecuencia de los residuos de aminoácidos y nucleótidos en un alineamiento múltiple. El modelo no solamente captura las frecuencias observadas de los residuos, sino que también predice las frecuencias de caracteres o residuos no observados. Otro propósito de estos modelos es permitir alinear o emparejar parcialmente secuencias, lo cual permite realizar búsqueda de secuencias relacionadas en bases de datos para detectar con alta sensibilidad miembros distantes de una misma familia de secuencias. Un modelo de Markov, también conocido como cadena de Markov, describe una secuencia de eventos que ocurren uno tras otro en secuencia. Cada evento determina la probabilidad del siguiente evento. Una cadena de Markov puede verse como un proceso que se mueve en una dirección de un estado al siguiente con una cierta probabilidad, la cual es conocida como la probabilidad de transición. Un ejemplo de un modelo de Markov es el cambio de la señal en un semáforo, ya que el estado de la señal actual depende del estado de la previa.

Al igual que los perfiles, los modelos ocultos de Markov (HMM) son utilizados para modelar características estadísticas de toda una familia de secuencias. El modelo estadístico dado por el HMM es usado para buscar en las bases de datos secuencias relacionadas con la familia modelada.

- 1 Vaya a Uniprot ²², y en la pestaña titulada “Retrieve” ingrese los códigos Uniprot de TODAS las proteínas de interés. Luego presione el botón “Retrieve” y baje el archivo de extensión fasta. Puede omitir este paso si ya tiene las secuencias de sus proteínas.

Para realizar el primer alineamiento múltiple utilizaremos un algoritmo progresivo.

- 2 En la misma página donde se encuentra dentro de UniProt, presione la pestaña titulada “Align”. Ingrese las secuencias que bajo en formato FASTA y luego presione “Align”.

Aquí usted está ejecutando el programa Clustal Omega de alineamiento múltiple progresivo.

- P.33** ¿Qué tan largo es el alineamiento (pista: cada bloque alineado consta de 60 residuos/gaps)?
Describa el alineamiento obtenido.

- 3 Baje el alineamiento en formato FASTA.

Ahora utilizaremos un algoritmo iterativo de alineamiento múltiple.

- 4 Vaya a Genome ²³. Cambie el “Output Format” a CLUSTAL. Luego ingrese las secuencias que bajo en formato FASTA desde UniProt y ejecute el alineamiento.

- P.34** ¿Qué tan largo es el alineamiento (pista: cada bloque alineado consta de 60 residuos/gaps)?
Describa el alineamiento obtenido.

- P.35** A continuación utilizaremos algoritmo basado en bloques de alineamiento múltiple. Vaya a BiBiServ ²⁴ [submission.html](http://bibiserv.techfak.uni-bielefeld.de/). Ingrese las secuencias en formato FASTA y presione “Submit”. Una vez que el resultado del cálculo esté listo, siga el link para ver “original dialign output”.

- P.36** ¿Qué tan largo es el alineamiento (pista: cada bloque alineado consta de 50 residuos/gaps)?

Para comparar los alineamientos podemos utilizar diferentes métricas, por ejemplo, podemos utilizar como criterio que el mejor alineamiento es el que posee una menor cantidad de gaps.

- P.37** Ordene los alineamientos de peor a mejor. Por ejemplo, basado en bloques <iterativo <progresivo.

²²<http://www.uniprot.org/>

²³<http://www.genome.jp/tools/prn/>

²⁴<http://bibiserv.techfak.uni-bielefeld.de/dialign/>

Ejecutaremos el programa PSI-BLAST para crear un perfil para cada una de las proteínas estudiadas, y encontrar el mayor número posible de secuencias en la base de datos no redundante.

Para cada una de las 5 proteínas haga lo siguiente.

2 Ingrese a la pagina de BLAST ²⁶ y seleccione blastp.

Ingrese el código UniProt de la proteína de interés. Donde dice “Program Selection” seleccione como “Algorithm” PSI-BLAST. Extienda la sección que dice “Algorithm parameters” y en “Max target sequences” seleccione 20000, y en “Expect threshold” escriba 0.00001. Finalmente en la sección que hace referencia a PSI BLAST al final de la página, en “PSI-BLAST Threshold” escriba 0.00001. Verifique que como “Database” este utilizando “Non-redundant protein sequences (nr)”.

P.40 Describa cada uno de estos parámetros.

3 Una vez ingresados los parámetros, presione el botón BLAST. Una vez que aparezcan los resultados, vaya a la sección que dice “Sequences producing significant alignments” Presione el boton que dice: Select All. ¿Cuántas secuencias se obtuvieron a partir de la primera búsqueda? Vaya a la sección que dice “Run PSI-Blast iteration 2” y presione el botón “Go”. ¿Cuántas secuencias se obtuvieron a partir de la segunda iteración? Repita una tercera, cuarta y quinta iteración, siempre anotando el número de secuencias obtenidas. (Nota: No todas las proteínas le van a permitir hacer 5 iteraciones.) ¿Cómo se compara el número de secuencias obtenidas usando PSI-BLAST, a aquéllas reportadas para las respectivas familias de cada proteína en Pfam?

c. Búsqueda en bases de datos utilizando HMMs . Ahora realizaremos una búsqueda más compleja. Usando el mejor alineamiento múltiple encontrado en la pregunta 1 (pista: bájelo en formato FASTA), vaya a la pagina de hmmer ²⁷ y ejecute una búsqueda utilizando HMMs , manteniendo todos los parámetros predeterminados. (Note que está haciendo una búsqueda contra la base de datos NR: non redundant, la misma que utilizó en la pregunta anterior). ¿Cuántas secuencias se obtuvieron a partir de la búsqueda? ¿Cuántas secuencias hubiera esperado usted obtener de la búsqueda (pista: compare con los resultados obtenidos en 2a y 2b)?

2.5. Clasificación estructural

Los dominios estructurales de una proteína entregan mayor información que una descripción general de proteína. De esta forma, una clasificación basada en dominios entrega una mayor precisión y permite distinguir las relaciones evolutivas.

La primera clasificación detallada de las estructuras de proteínas y los dominios que las componen fue realizada por Jane Richardson en 1981 ²⁸. Actualmente existen múltiples clasificaciones, entre ellas las más importantes son ²⁹:

- CATH: cuyo acrónimo viene dado por los niveles de organización estructural de las proteínas: Clase, arquitectura, topología y homología. Su proceso de agrupamiento tiene pasos automatizados y revisados manualmente. Define tres clases principales todo α , todo β y α/β . En el nivel arquitectura, los dominios se agrupan a partir de propiedades generales con respecto a la forma de plegamiento, pero sin considera la forma en que están conectados entre ellos. La topología es el nivel analogo al plegamiento en SCOP, agrupa estructuras que tienen un número similar y arreglo tridimensional y conectividad de los elementos de estructura secundaria. Su último nivel es corresponde a la superfamilia de homologos, con rangos mayores a un $>30\%$ de identidad de secuencia y domios de alta similitud estructural y funcional, lo cual sugiere una evolución de estos partir de un ancestro en común.

²⁶<http://www.ncbi.nlm.nih.gov/BLAST/>

²⁷<http://hmmer.janelia.org/search/hmmsearch>

²⁸<http://kinemage.biochem.duke.edu/teaching/>

anatax/

²⁹BMC Structural Biology 2009, 9:23

- SCOP: Su nombre viene de Clasificación estructural de proteínas (en ingles). Que organiza las proteínas en: Clase, Plegamiento (Fold), Superfamilia, Familia y dominio. Contiene información de revisada. Define cuatro clases principales todo α , todo β , $\alpha + \beta$ y α/β . En ella, un dominio en un plegamiento (Fold) común tiene elementos de estructura secundaria en un arreglo similar y con igual conexión topológica. En una misma superfamilia, los dominios comparte una baja identidad de secuencia, pero sus estructuras – y a veces función – sugieren una relación evolutiva, mientras que en una familia, los dominios tienen una alta probabilidad de tener un ancestro evolutivo común, observado en un mayor valor de similitud de secuencia (>30 %) y función.

Cabe destacar que las proteínas pueden ser clasificadas en más de una categoría, ya que ambas clasificaciones se basan en protocolos de clasificación. Por ejemplo, en SCOP cerca del 80 % pueden ser clasificadas en una sola superfamilia, mientras que un 15 % es clasificada en dos. Ejemplo de esto es la estructura de la tylosina (pdb 1k9m), la cual esta clasificada en 23 superfamilias.

En resumen, a partir de la clasificación SCOP uno puede inferir las siguientes relaciones evolutivas.

- Familia: Una clara relación evolutiva. (>30 % identidad entre pares de secuencia aminoacídica)
- Superfamilia: Poseen un probable ancestro en común. Baja identidad de secuencia, pero estructura y función similar.
- Plegamiento (Fold): Similitud estructural relevante. Igual plegamiento implica un arreglo, topología y conectividad de elementos de estructura secundaria.



Dominio proteico. Se refiere a una región de la proteína que puede plegarse en una estructura tridimensional de manera independiente del resto de la proteína. Esta estructura puede mantener la función específica asociada al dominio, por ejemplo Dominio de unión NAD⁺, crear a un motivo de unión para otra molécula o proveer de las propiedades necesarias para que las proteínas sobrevivan a ciertas condiciones. Los dominios proteicos son habitualmente conservados evolutivamente, destacando como regiones de alta similitud de secuencia tanto en superfamilias o en proteínas que cumplen una función similar. El término superfamilia utilizado aquí se refiere a proteínas que tienen una relación evolutiva que no es evidente a partir de estudios de su secuencia nucleotídica, pero que si lo es a partir del estudio de su estructura o de patrones consensos únicos que permiten describir sus aminoácidos críticos.

Clasificación estructural de Serine Proteasas

P.41 A partir de lo aprendido y ayudado por la figura 3 infiera una clasificación SCOP hasta el nivel de superfamilia para Tripsina y Subtilisina. Justifique

1 Ingrese a la categoría principal de SCOP ³⁰

P.42 ¿Que clase es la que tiene la mayor cantidad de miembros? ¿En base a sus conocimientos biológicos explique esta observación?

³⁰<http://scop.mrc-lmb.cam.ac.uk/scop/>

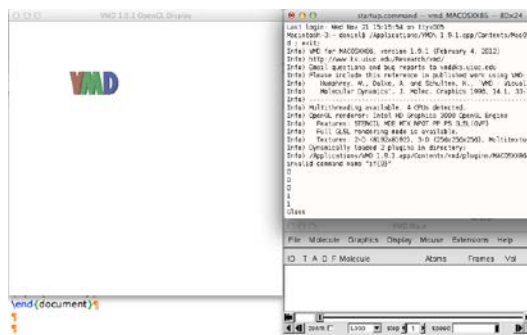


Figura 5. Ventanas abiertas al ejecutar VMD.

2 Encuentre y anote la jerarquía SCOP de la Tripsina humana

3 Encuentre y anote la jerarquía SCOP de Subtilisinas

P.43 Explique las diferencias a nivel de supefamilia entre las Tripsinas y las Subtilisinas

P.44 Utilizando la información disponible. ¿Describa el proceso evolutivo de las Serine Proteasas?

2.6. Uso Básico de VMD

2.6.1. Inicio

En esta unidad utilizaremos VMD para visualizar y analizar la estructura de una serine proteasa, el cuál servirá en el resto de este practico. Solo se abordarán aspectos básicos de uso, por lo que se recomienda leer el manual de usuario disponible en ³¹.

- 1 Para iniciar VMD escriba `vmd` en una ventana de terminal a Unix, doble-click en el icono de la aplicación de VMD en la carpeta **Applications** en Mac OS X, o haga click en el menu Inicio → Programas → VMD en Windows. En Unix escriba `vmd` en una ventana de terminal.



Webpdb. VMD puede descargar estructuras desde en formato pdb desde el Protein Data Bank. Si hay conexión a internet escriba el código de cuatro letras del cristal seleccionado, por ejemplo 1STP.

³¹<http://www.ks.uiuc.edu/Research/vmd/current/ug/>

2.6.2. Cargando el sistema

Para iniciar se deben cargar las coordenadas de los distintos átomos del sistema, las cuales se pueden obtener desde distintas fuentes como el *protein data bank* en el caso de proteínas o directamente con VMD. El archivo `30TJ.pdb` contiene las coordenadas atómicas de la cadena E de la proteína Serine Protease de origen bovino.

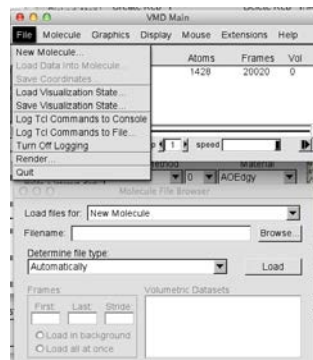


Figura 6. Cargando una molécula.

- 2 Elija el menú `File` → `New Molecule...` Fig. 6 en la ventana `VMD Main`. Aparecerá otra ventana `Molecule File Browser`, la que permite elegir los archivos dentro de los directorios disponibles.
- 3 Use el botón `Browse...` para localizar el archivo `30TJ.pdb` en `vmd-tutorial-files`. en el directorio `practico-evolucion`. Note que luego de seleccionar este archivo se encuentra nuevamente en la ventana `Molecule File Browser`. Para cargar el archivo, presione el botón `Load`. Recuerde revisar que esta seleccionado `New Molecule` en la sección `Load Files for`. Si lo desea, puede cerrar la ventana `Molecule File Browser`.
- 4 El sistema se visualiza en la ventana `OpenGL Display`. Utilice el ratón para rotar el sistema, para escalar la representación utilice el tercer botón del ratón o haga click la tecla “S” en el teclado y luego mueva el cursor en la ventana `OpenGL Display`. Para volver a rotar haga click en la tecla “R” en el teclado. Para trasladar el sistema, utilice la tecla “T”.
- 5 Para volver a la visualización inicial presione `Display` → `Reset View`.
- 6 En el menú `Menu` → `Graphics` elija `Representations`. Se abrirá la ventana `Graphical Representations`, en verde se destaca la representación que se está utilizando para mostrar su molécula.
- 7 En la ventana `Graphical Representations` elija la viñeta `Draw Style`, esta permite cambiar el estilo de visualización de cada representación creada a través del menú `Drawing Method`. El método o estilo predeterminado es *Lines*, que muestra en forma de líneas los enlaces entre átomos.
- 8 Cada método de dibujo tiene sus propios parámetros. Por ejemplo, modifique el campo `Thickness` para aumentar el ancho de la líneas que representan los enlaces atómicos.
- 9 Cambie el método de dibujo a `VDW` (van der Waals). Cada átomo es ahora representado como una esfera, esto permite visualizar la forma y el volumen de la proteína. Elija ahora el método de dibujo `New Cartoon`.

El método `New Cartoon` permite visualizar de forma simplificada la estructura secundaria de su proteína. Las hélices se visualizan en forma de espiral, láminas β en forma de flechas y las otras estructuras en forma de un tubo delgado. Este método es el de mayor uso en representaciones de sistemas proteicos

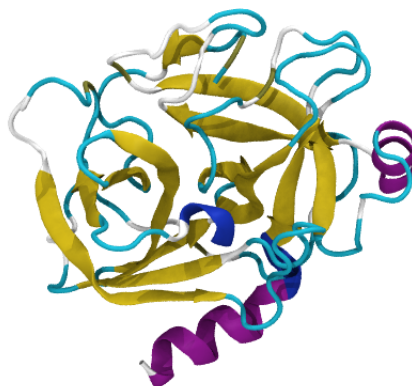


Figura 7. Serine Proteasa Bovina, pdb 3OTJ cadena E

- 10** En el menú **Coloring Method** es posible elegir el color de la representación. Diferentes estilos se encuentran disponibles, por ejemplo: Nombre, tipo de átomo, nombre del residuo, estructura, etc. Coloree la representación usando la información de estructura secundaria.



Otros métodos de dibujo. CPK y Licorice son métodos de dibujo que permiten observar estructuras con facilidad. La primera emula los sistemas de bolas de radio variable y tubos utilizados en química para visualizar moléculas, mientras que la segunda representa los átomos como esferas de radio fijo que no puede ser modificado.

11 La entrada de texto “Selected Atoms” de la ventana Graphical Representations permite acotar los átomos que se están visualizando. Reemplace la palabra *all* por *protein* y presione Apply (o presione Enter en el teclado), lo cual debe realizar cada vez que modifica un campo en la ventana Graphical Representations.

12 En la ventana Graphical Representations elija la viñeta Selections. En la sección Singlewords se encuentran los diferentes campos por los cuales se pueden realizar selecciones rápidas. Seleccione “beta” en el campo de texto para seleccionar solo los átomos de la proteína que se encuentran formando láminas β .

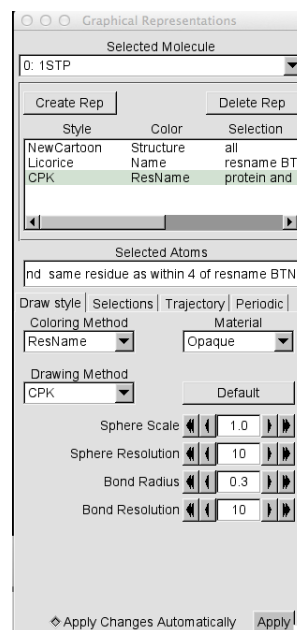


Figura 8. Ventana Graphical Representations

13 Para visualizar la proteína sin helices y láminas β escriba en Selected Atoms: *(not helix) and (not betasheet)*.

Con el ratón se pueden crear y seleccionar diferentes representaciones. Además, estas pueden activar o desactivar haciendo doble-click sobre ellas o ser borradas utilizando el botón Delete Rep button, etc.

14 Modifique la selección de átomos para representar toda la proteína, luego desactívela haciendo doble click sobre la representación.

15 Cree una nueva representación, sin borrar la existente, para esto utilice el botón Create Rep.



Selecciones geométricas complejas. Distintos tipos de geométricas se pueden lograr utilizando selecciones complejas en relación al espacio cartesiano. Cuál es la selección apropiada para obtener una esfera a partir de las moléculas presentes. Recuerde que una molécula no puede quedar incompleta!

En la viñeta Selections en el campo Keywords es posible ver distintas palabras claves que sirven para seleccionar átomos. Inspeccione en Keywords los nombres de los residuos que componen el sistema. Para este caso, se muestran los nombres de los residuos aminoacídicos de la proteína.

16 Cree una nueva representación, para esto utilice el botón **Create Rep**

17 Al crear una nueva representación se copian los valores de la selección anterior. Para seleccionar un residuo de nombre conocido, por ejemplo el residuo de nombre HIS, utilizaremos la palabra clave “*resname*” seguida del nombre del residuo como aparece en el archivo PDB, para ello escriba “*resname HIS*” en el campo “*Selected Atoms*”. Para una mejor representación cambie el método de dibujo por *licorice* y coloree la representación por nombre de átomo.

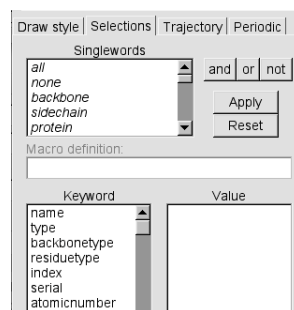


Figura 9. Campos disponibles para seleccionar

En esta nueva representación se ha visualizado los residuos de histidina que están presentes en esta proteína.

Diferentes campos pueden utilizarse para crear selecciones de mayor complejidad. El cuadro 5 muestra algunas selecciones posibles.

Selección	Acción
protein	muestra una proteína si esta presente.
protein and chain E	muestra la cadena E si esta es proteína y si esta presente.
resid 1	el primer residuo
resid 1 2 12	el primer residuo, el segundo y el decimosegundo residuo
x >0	átomos que tienen la coordenada <i>x</i> positiva
same residue as x>0	átomos de residuos que tienen un átomo cuya coordenada <i>x</i> es positiva
name CA and x>0	átomos de nombre CA (carbonos alfa) y coordenada <i>x</i> es positiva
within 5 of resid 1	átomos a 5 Angstrom del residuo 1
backbone	átomos del backbone

Tabla 5. Ejemplo de selecciones

18 Para observar los residuos de la proteína que se encuentran cercanos al residuo *HIS* se utiliza una selección de mayor complejidad. Cree una nueva representación y escriba la siguiente selección para los átomos que están a 4 Å de distancia del residuo 63: **protein and within 4 of resname HIS and resid 63**. Utilice el método de dibujo CPK y coloree por nombre del residuo.

En la ventana OpenGL se observan átomos de distintos colores, sin embargo, es de mayor utilidad seleccionar los residuos que tienen átomos que cumplen una condición.

- 19 Cree representaciones para los siguientes residuos del sitio activo 63 75 77 80 85 107 200, los cuales están descritos en UNIPROT para esta proteína³². El resultado debe asemejar a la figura 10 en la cual hemos agregado una representación de la superficie de la proteína utilizando el método QuickSurf

<http://www.uniprot.org/uniprot/P00760>.

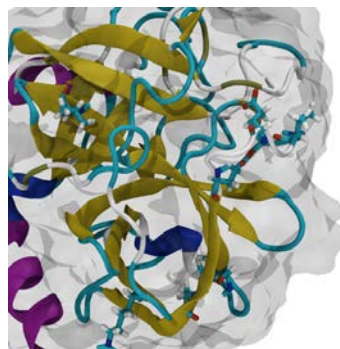


Figura 10. Representación Final

- 20 Utilice el ratón para orientar la visualización. Para volver a la visualización inicial presione Display → Reset View.



Enlaces de Hidrógeno. La diferencia de electronegatividad entre los átomos de hidrógeno y oxígeno permite que las moléculas de agua puedan compartir H entre ellas, formando lo que se denomina puente o enlace de hidrógeno. Estos se pueden visualizar utilizando el método Hbond en la ventana de representaciones. Cuales son los parámetros necesarios para que se formen? tiene alguna relación la formación de puentes de hidrógeno con la estructuras observadas?.

- 21 Para reconocer un residuo específico es posible utilizar el ratón y la ventana OpenGL para conocer su identificador. Para ello en la ventana VMD Main, elija el menú Mouse → Query y haga click sobre un átomo de interés. Al hacerlo, aparecerá en pantalla información básica del átomo seleccionado. Para obtener mayor información en la ventana VMD Main elija el menú Graphics → Label, que abrirá una ventana con la descripción del átomo seleccionado.

Otra forma de seleccionar un átomo es utilizando el modulo **Sequence Viewer**, el cual se encuentra en la Vineta **Extension/Análisis** de la pantalla **Main**

- 22 Utilice **Sequence Viewer** para verificar que los residuos que seleccione utilizando su número de residuo coinciden con los esperados a partir de la numeración en la secuencia encontrada en UNIPROT.

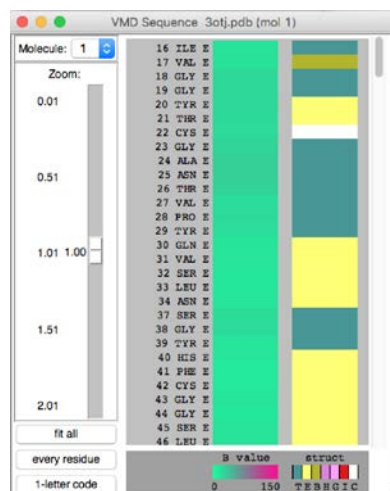


Figura 11. Sequence Viewer

VMD es una herramienta útil para describir y no hemos explorado todas sus capacidades. Por ejemplo, seleccionando en la ventana VMD Main el menú **Mouse** → **Labels** → **Bond** es posible conocer la distancia entre dos átomos de interés. Al hacerlo, aparecerá en la ventana **Label** bajo la etiqueta **Bonds** → **value** la distancia entre los átomos seleccionados en la configuración actual.

2.6.3. Guardando su trabajo

La imagen creada puede ser almacenada en archivos utilizando VMD. Además, las representaciones creadas pueden ser almacenadas utilizando un estado de VMD. El archivo de estado contiene la información necesaria para reiniciar las sesiones de VMD, sin perder el trabajo realizado.

23 En la ventana VMD Main, elija el menú **File** → **Save State**. Escriba un nombre apropiado para el archivo (i.e, 30TJ.vmd) y guárdelo en el directorio de trabajo.

24 cierre y reinicie VMD para cargar el estado guardado

El menú **File** → **Load State** permite cargar el estado guardado previamente. Para guardar en imágenes las representaciones realizadas, VMD puede generar un archivo imagen que puede ser utilizado en otros programas.

25 Usando lo aprendido, genere una imagen de calidad, que muestre alguna característica de interés del sistema (por ejemplo la interacción entre dos residuos de interés o un metal). Ponga cuidado en la resolución de cada representación.



Operadores lógicos. Habitualmente utilizamos los operadores booleanos “Y”, “O” y “NO”. El primer es un operador “incluyente”, el segundo de “exclusión”, mientras que el tercer operador es considerado de “reunión” o suma lógica. Podemos ver la diferencia entre ellos observando la cantidad de átomos de selecciones que las incluyan como: “protein or x>0”], “protein and x>0”] y “protein and not x>0”].

26 Cambie el color del fondo seleccionando el menú **Graphics** → **Colors**. Elija la categoría **Display** → **Background** → **8 white**. El fondo debe estar en blanco ahora.

27 Elija el menú **File** → **Render....** Se abrirá una ventana llamada **File Render Controls**.

28 Puede utilizar distintos métodos para generar la imagen, seleccione **Snapshot** o **Tachyon**.



Troubleshooting. Si el directorio donde guardara los archivos contienen espacios, es necesario que antes de utilizar **File Render Controls**, seleccionar el directorio de trabajo utilizando la **Console Tk**. Para esto seleccione el menú **Extensions** → **Tk Console**. Use cremillas para navegar a la dirección, i.e., cd “C:\ Documents and Settings”.

29 Escriba el nombre del archivo a generar en el campo de texto **Filename**, i.e., **imagen.bmp**.

30 Presione el botón **Start Rendering** y se generará el archivo con la imagen que aparece en la ventana **OpenGL**. Esto puede tardar algún tiempo dependiendo de las capacidades del computador utilizado. Al finalizar, el archivo se encuentra en el directorio seleccionado o en el directorio de trabajo.

- 31 Antes de comenzar con la siguiente sección explore VMD libremente. Al terminar, reinicie VMD.



Archivos de Coordenadas y de Conectividad. Los programas de visualización como VMD, infieren los distintos enlaces a partir de las distancias entre átomos, por lo que habitualmente se pueden encontrar átomos con una mayor cantidad de enlaces que los permitidos, ie. carbonos con 5 átomos asociados, moléculas de agua con 4 o más hidrógenos. Para solucionar esto se hace uso de un archivo extra que contiene la conectividad de cada átomo, además de otros parámetros necesarios para realizar una simulación, por ejemplo las cargas atómicas.

3. Evolución y estructura

Históricamente, las proteínas han sido clasificadas usando la similitud de secuencia. Las familias de proteínas son combinadas en superfamilias basados en su actividad catalítica, motivos de secuencias u otras cualidades conservadas. Dado lo anterior, se infiere que proteínas dentro de la misma superfamilia provienen de un ancestros en común, aún cuando puedan tener actividades enzimáticas diferentes o desconocidas.

La evolución divergente puede ser trazada a través de similaridades de secuencias y/o plegamiento. Las familias divergentes típicamente pertenecen a la misma superfamilia, comparten motivos estructurales y funcionales, y exhiben identidad de secuencia detectable.

En cambio la evolución convergente implica la aparición de propiedades estructurales o funcionales similares dentro de familias de proteínas que no divergieron de un ancestros en común. Típicamente, estas familias pertenecen a distintas superfamilias y no exhiben identidad de secuencia.

4. Descripción del práctico

El objetivo del práctico de hoy, es determinar si existe **evolución divergente o convergente** en cada uno de los sets otorgado. Para ello, utilizaremos cuatro sets de archivos PDB. **Es su misión averiguar cual es la función de cada grupo.*

- **Glutation:** 1AGS (Cadena A), 1EEM (Cadena A), 1FHE (Cadena A), 1GSE (Cadena A) y 1K0D (Cadena A)
- **G-proteínas:** 1AGP (Cadena A), 1BYU (Cadena A), 1EGA (Cadena A), 1EGA (Cadena A) y 1NF3 (Cadena A)
- **Serine:** 1EP5 (Cadena A), 1HAX (Cadena A), 1LMW (Cadena A), 1OP0 (Cadena A) y 1RTF (Cadena A)
- **Subtilisinas:** 1DUI (Cadena A), 1NDU (Cadena A), 1SBI (Cadena A), 1T1E (Cadena A) y 2ID4 (Cadena A)

Puede descargarlos desde la página de Protein Data Bank, o acceder a la carpeta otorgada junto con este práctico.

5. Alineamiento estructural de estructuras

Para realizar esta parte debemos cargar el grupo de estructuras a evaluar en VMD (Si hay dudas en este paso, recurrir a tutoriales anteriores).

Una vez cargadas, iremos a la selección del menú principal: Extensiones, Analysis ->Multiseq. Si estamos abriendo por primera vez Multiseq, el programa nos pedirá instalar o actualizar bases de datos. Si este es el caso, debemos asignar "Yes" esperar que se carguen todas las librerías.

En el menú de Multiseq (Figura 12) se verán las secuencias de las estructuras cargadas en el menú principal de VMD. Si la estructura tiene varias cadenas se cargaran cada una por separado o bien otro tipo de moléculas (Agua, iones, etc), solo se verán signos de interrogación (?). Estos últimos deben ser descartados. Adicionalmente en el la parte File, se pueden agregar mas secuencias provenientes de estructuras o en otros formatos.

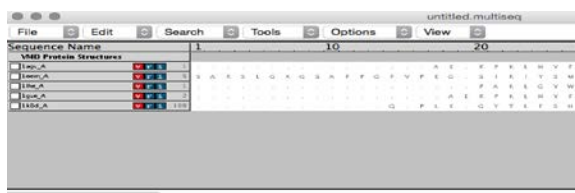
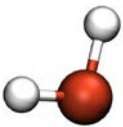


Figura 12. Menú principal de Multiseq

A continuación utilizaremos el programa STAMP para alinear nuestras estructuras.

Dirigirse a Tools->Stamp Structural Alignment Tool. La ventana mostrará los parametros requeridos. Por ahora usaremos el cuadro por defecto.



STAMP. El programa STAMP (Structural Alignment of Multiple Proteins) es una herramienta que alinea secuencias de estructuras de proteínas utilizando la estructura tridimensional. El algoritmo de STAMP busca minimizar la distancia entre C α de residuos aplicando rotaciones de cuerpo rígido y traslaciones. STAMP solo puede alinear proteínas que tengan secciones comunes. Información adicional: <http://www.rfcgr.mrc.ac.uk/Registered/Help/stamp/stamp.html>

Una vez aplicado, en la pantalla principal, veremos nuestras proteínas alineadas. En la ventana de Multiseq, veremos que las secuencias se alinearon en relación a las estructuras. En esta etapa puede colorear las estructuras bajo distintos parámetros: Conservación de secuencia, identidad, etc. Buscar esas opciones en la sección View->Coloring (Figure 13)

De resultado, podremos ver las secuencias y las estructuras coloreadas bajo el parámetro deseado.

5.1. Árbol filogenético

La opción de Arbol Filogenético en el programa Multiseq puede mostrarnos relaciones entre las proteínas consideradas.

Los árboles filogenéticos basados en estructuras se construyen a partir de valores de RMSD o Q value entre las moléculas evaluadas.

Para usar esto debemos

- Alinear las estructuras usando STAMP
- En el programa Multiseq elegir Tools->Phylogenetic Tree. Aquí se puede elegir los parámetros requeridos, entre ellos valor Q_H o RMSD. ¿Qué es el valor Q_H ?
- Seleccionar Árbol estructural usando Q_H y presiona OK. (Figura 15)

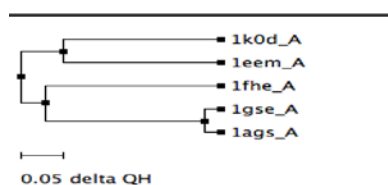
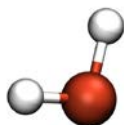


Figura 15. Imagen de ejemplo de árbol estructural



Valor Q_H . El valor Q_H es una métrica que señala homología estructural similar al valor Q . Se compone de dos partes: Q_{align} (parte alineada) and Q_{gap} (parte con gaps).

Una vez terminada esta parte, podemos repetir la experiencia con los otros set de estructuras.