Revista
Torreón Universitario
www.faremcarazo.unan.edu.ni - ISSN electrónico 2313-7215
Año 4 | No. 8 | pág. 18 - 25 | Octubre 2014 - Enero 2015

Análisis de algoritmos basados en técnicas de conglomerado aplicados en el alineamiento y comparación de secuencias de proteínas

MSc. CONCEPCIÓN MENDIETA B.
Universidad Nacional Autónoma de Nicaragua, Managua
Facultad Regional Multidisciplinaria de Carazo
connymendieta@yahoo.com

Palabras clave: *Bioinformática, Minería de Datos, estrategias de Minería de Datos, Clasificación, Conglomerado, Proteómica, secuencias de proteínas*

Resumen

La Bioinformática tiene como objetivo el desarrollo y uso de técnicas matemáticas y computacionales para ayudar a resolver problemas referentes a la Biología.

En la actualidad existen muchas técnicas de *Minería de Datos* que han posibilitado el desarrollo de ésta, entre las que sobresalen la *Clasificación* y el *Conglomerado* con la finalidad de construir herramientas de análisis más eficientes. No obstante, dada la complejidad que involucra la búsqueda de información interesante en las bases de datos biológicas, desde una perspectiva proteínica, una necesidad en la ciencia actual recae en demandar mayor capacidad de almacenamiento y tratamiento de los datos recopilados a través de los años en los distintos experimentos científicos de orden biológico. Esta necesidad, por tanto, ha implicado la afloración de muchos algoritmos afines al problema de estudio. Sin embargo, la calidad de resultados varía considerablemente al aplicar diversos algoritmos a un mismo conjunto de datos proteínicos.

En este documento se presenta un análisis de algunos algoritmos de *Conglomerados* aplicados en áreas específicas de la Bioinformática, la *Proteómica*, desde el punto de vista de alineamiento y comparación de secuencias de proteínas. Para tal fin, se examinaron tres algoritmos muy populares por su amplio uso, siendo estos: ClustalW, Muscle y T-Coffee.

Dado los resultados experimentales se determinó que el mejor algoritmo, desde el punto de vista de tiempo de ejecución fue Muscle, pero T-Coffee presentó mayor calidad y claridad de los alineamientos resultantes.

Introducción

El alineamiento de secuencias de proteínas tiene como objetivo alinear dos o más secuencias de un mismo organismo o de organismos diferentes con el fin de destacar las regiones similares entre las secuencias sujetas a comparación [Corp88]. De esta manera, se permite inferir si una secuencia desconocida es similar, parcial o totalmente a alguna secuencia ya sea desde el punto de vista de su estructura o función.

Los alineamientos de secuencias pueden usarse desde dos perspectivas: *el alineamiento local* y *el*

alineamiento global [Carr88]. Mediante un alineamiento global, todas las bases se alinean con otra base o con un gap ("-"), se pueden identificar repeticiones internas entre las secuencias de proteínas sujetas a investigación, así como encontrar las partes de las secuencias que se conservan entre las especies. No obstante, en el alineamiento local sólo se comparan trozos de las secuencias sujetas a experimentación. Por tanto, entre las aplicaciones rescatables denota la predicción de funciones de una proteína desconocida mediante la búsqueda de *dominios funcionales comunes*. Este proceso es posible tanto en los alineamientos locales entre dos secuencias como en el alineamiento múltiple entre conjuntos de secuencias.

Las circunstancias anteriores han convertido el alineamiento de secuencias de proteínas en una de las áreas de Bioinformática más estudiadas. En consecuencia, existe una gran cantidad de algoritmos concebidos a través de los años con fines de determinar las regiones conservadas (regiones que se mantienen idénticas entre las secuencias sujetas a comparación), por tanto, la necesidad hoy en día de contar con información confiable desde esta perspectiva es fundamental para poder, así, determinar nuevos fármacos que ayuden al tratamiento de enfermedades que hasta ahora no tienen cura alguna. Otros de los aspectos rescatables es la obtención de mayor información acerca de nuestra evolución como organismos vivientes interrelacionados. No obstante la información obtenida, referente al análisis y alineamiento de proteínas, sujetas a comparación, varía considerablemente al utilizar diferentes algoritmos a un mismo conjunto de datos. Por tanto, la pregunta que orientó la investigación en cuestión fue: *¿Qué algoritmo, sujeto a estudio, resulta más confiable para el alineamiento de proteínas?*, pues se debe recordar que la información resultante será utilizada para la inferencia de hipótesis que posteriormente conlleve a nuevos modelos computacionales y por ende, al establecimiento de nuevo conocimiento que potencialmente mejoren la calidad de vida humana.

Sin embargo, en la actualidad existen muchos algoritmos disponibles en la Web para el alineamiento y comparación de secuencias. Esta diversidad implica que el investigador debe hacer una adecuada selección de los parámetros de búsqueda, pues una incorrecta selección conlleva a la no detección de similitudes relevantes entre secuencias de proteínas.

A continuación se hace una reseña acerca de los métodos más importantes.

I. Algoritmos usados para el alineamiento de secuencias de proteínas

Los algoritmos tradicionales tratan de encontrar el alineamiento óptimo y se basan en programación dinámica, métodos concebidos por Needleman y Wunsch en 1970, así como por Smith y Waterman en 1981 [Need70]. Bajo este enfoque, el alineamiento óptimo se obtiene utilizando una valoración alcanzada a partir de una matriz de puntuación para los diferentes pares de aminoácidos o nucleótidos que puedan quedar enfrentados en dicho alineamiento y las penalizaciones por la introducción de gap en las zonas donde las secuencias no concuerdan.

Dado que los algoritmos basados en *programación dinámica* requieren un tiempo computacional de $O(N^2)$, es decir que su tiempo de cálculo es proporcional al cuadrado del tamaño de las secuencias que se comparan, se han convertido en técnicas muy caras y no óptimos en el alineamiento múltiple de secuencias de proteínas. Esta situación ha conllevado al desarrollo de nuevas técnicas de tipo heurísticos (métodos basados en un comportamiento común), que buscan velocidad a cambio de sensibilidad y selectividad.

Un ejemplo de estas técnicas es la serie de programas FASTA [Pear88], que basan su estrategia en identificar diagonales con mayor número de identidades que luego analiza usando el esquema de puntuación. En consecuencia, se realiza una revisión completa de las dos secuencias en un tiempo proporcional a O(N+M) es decir, proporcional a la suma de los tamaños de las secuencias comparadas.

No obstante, diversos objetivos en el análisis de secuencias de ADN y proteínas implican un alineamiento múltiple a fin de encontrar la homología entre éstas.

Dado que el alineamiento múltiple de secuencias se concibe como una generalización del alineamiento de pares de secuencias, cuya complejidad crece exponencialmente con el número de secuencias que interviene, hace que los métodos anteriores queden limitados a un alineamiento sencillo, poco confiable y no óptimo para satisfacer las necesidades actuales en lo referente al alineamiento múltiple de proteínas.

Esta situación implicó la búsqueda de otras técnicas desde ópticas totalmente distintas. Es así que en [Corp88] se estableció una solución basada en las técnicas de conglomerados**,** pues sostenía que estos conglomerados se resolvían progresivamente. Para tal efecto, se necesitaba una medida de similitud, a fin de medir la distancia entre cada par de secuencias del conjunto en estudio, luego se elegirían el par de secuencias con el valor más alto y se alinearían y agruparían entre sí hasta formar un sólo grupo de secuencias y a partir de este hecho el conjunto de secuencias es tratado como una sola. El proceso se repite hasta tener una secuencia consenso.

En consecuencia, se definió experimentar con tres algoritmos basados en técnicas de conglomerados y es que cuantiosos estudios han demostrado la eficiencia de los algoritmos basados en esta técnicas [Braz00], [Hhei98], [Eise98], etc**.** Una de las razones para esto es la filosofía que adoptan los algoritmos al procesar los datos: "el *conglomerado entre objetos semejantes*". Basado en este principio, se decidió estudiar algunos de los métodos concebidos bajo el enfoque de conglomerados.

Los argumentos que justifican esta elección fueron:

- Porque la calidad del conglomerado resultante, una vez que los datos son analizados, muestra una buena representación gráfica lo que contribuye con la tarea de análisis e interpretación del investigador.
- Porque usan medidas de distancias sencillas, que son de gran aplicación en la mayoría de los algoritmos de conglomerados por su eficacia, siendo estas: *la distancia Euclidiana y el coeficiente de correlación de Pearson*.
- Porque son algoritmos adecuados para tratar los tipos de datos en estudio.

De esta manera, lo fundamental es determinar qué algoritmo resulta ser más eficiente en el análisis de datos biológicos concernientes a las áreas de investigación.

A continuación se aborda cada uno de los algoritmos en estudio.

1.2 Técnicas de conglomerado sujetas al análisis y experimentación

1.2.1 ClustalW

Propuesto por Thompson et al, en el año de 1994. Este algoritmo es una mejora de *Clustalv* propuesto por Higgins et al [Higg92], la *W* hace referencia a "weighting". Según sus autores, ClustalW mejora la sensitividad del alineamiento de múltiples secuencias a través de la asignación de pesos a las secuencias, posición específica de gap y pesos de la matriz escogida.

Para el alineamiento inicial se puede elegir entre un algoritmo de Programación Dinámica y uno heurístico de tipo FASTA, por ejemplo.

La distancia genética es el número de posiciones "mismatched" dividido entre las "matched (no se cuen-

tan las enfrentadas a gaps) y los pesos se relacionan con la distancia a la raíz del árbol.

Un análisis más detallado se puede encontrar en [Thom94].

1.2.2 Muscle

Este algoritmo fue propuesto por Robert C, en el 2004 a fin de mejorar el tiempo de ejecución y la calidad del alineamiento de los algoritmos propuestos bajo conglomerados. Muscle mejora, según sus autores, dos requerimientos básicos de los métodos concebidos para el alineamiento de secuencias múltiples: la exactitud biológica de la información resultante del proceso y la complejidad computacional requerida en dicho proceso es decir, el tiempo y los requerimientos de memoria.

Un análisis más detallado se puede encontrar en [Robe04].

1.2.3 Coffee

Este algoritmo fue propuesto por Notredame et al, en el 2000. Mediante este algoritmo es posible combinar resultados obtenidos de otros métodos de alineamiento. Este algoritmo separa de dos en dos todas las secuencias sujetas a la comparación, hasta producir un alineamiento global y una serie de alineamientos locales y luego combina estos alineamientos hasta obtener, mediante programación dinámica, una única y consensuada secuencia.

Un análisis más detallado se puede encontrar en [Notr00].

II. Aspectos relevantes sujetos a investigación

Mediante los experimentos realizados se pretendió comprobar las afirmaciones positivas y negativas que se atribuyen a cada uno de los algoritmos en estudio y así determinar el algoritmo con mejor desempeño computacional y biológico.

Haciendo una retrospectiva acerca del funcionamiento de dichos algoritmos, se afirma que en el caso de *Clustalw* mejora en gran manera la sensitividad en el alineamiento múltiple de secuencias al proporcionar pesos a cada subparte de las secuencias sujetas a comparación, además que posee buen tiempo de ejecución. Aunque la contraparte afirma que *Clustalw* es muy sensible a cometer errores porque presenta alto grado de degradación durante el proceso y que por tanto no es recomendable para ser aplicado al área en estudio.

En el caso de Muscle, se afirma que es superior a T-Coffee y Clustalw porque cumple fielmente con los dos requerimientos básicos y deseables que debe tener un algoritmo para llevar a cabo correctamente el alineamiento de secuencias y es que sus autores sostienen que Muscle alcanza mejor promedio de extensión y velocidad en su ejecución, de manera que ahorra más recursos de memoria.

Los autores de T-Coffee afirman que es el mejor entre los algoritmos afines al área de estudio porque combina tanto la información local como global, algo que no son capaces de realizar los métodos anteriores. De manera que los usuarios finales se benefician de la simplicidad del algoritmo al no proveer ningún parámetro extra. No obstante, la contraparte afirma que es bastante lento, por tanto, no es adecuado para el alineamiento múltiple de secuencias.

A fin de comprobar o refutar las afirmaciones anteriores, se definieron tres parámetros con el objetivo de analizar la calidad de ejecución de los algoritmos. Estos parámetros fueron:

- Tiempo de ejecución
- Calidad de los resultados

- Claridad en la estructura en que se presentan los resultados del análisis para los usuarios finales.

III. Método

A. Herramientas de software usadas

Las herramientas seleccionadas para los experimentos fueron: SRS (*Sequence Retrieval System*), que contiene información de una diversidad de base de datos diferentes y que además incluyen bibliografía (MEDLINE), secuencias, motivos, señales, información adicional (taxonomía, código geneático, etc.,) y puede ser accedido en EBI (http://srs.ebi.ac.uk/). Entre los recursos usados destaca el NCBI (http://www.ncbi.nlm.nih.gov/). Cabe señalar que tanto EBI como NCBI intercambian información relevante a secuencias de ADN y proteínas. Dado su carácter público y sobre todo por tener información de gran calidad, que es actualizada constantemente, es uno de los recursos más usados por los investigadores en el nivel mundial referente al análisis de genomas y proteínas.

B. Fuente y selección de datos

Para este caso se recurrió a CluSTr (http://www.ebi.uniprot.org/uniprot-srv/uniProtClustrSearch.do), que contiene un cluster de las mejores bases de datos relacionadas al área en estudio. Desde este recurso se accedieron las bases de datos *UniProtKB-Swiss-Prot* (http://www.ebi.uniprot.org/uniprot- srv/uniProtInterproSearch.do**)**, de donde se obtuvieron las secuencias y demás información de las proteínas seleccionadas.

C. Características de los datos sujetos al análisis

La proteína seleccionada para el primer experimento, fue de tipo FosB obtenida de cinco organismos distintos pertenecientes al reino animal. Para el segundo experimento se seleccionaron cinco dehidrogenasas distintas para un mismo organismo, en este caso para el humano.

En la Tabla 1. se detallan las características esenciales de las secuencias de proteínas seleccionadas para el proceso de experimentación.

Tabla 1: Datos de referencia de las secuencias de proteínas usadas en los experimentos

Tipo de proteína	Nombre de la secuencia	Longitud de la secuencia (cantidad de aminoácidos, aa)	Tipos de Datos
FOSB	FOSB_Humano	338	Secuencial
FOSB	FOSB_Raton	338	Secuencial
FOSB	FOSB_Perro	338	Secuencial
FOSB	FOSB_Chimpanse	338	Secuencial
FOSB	FOSB_Gato	380	Secuencial
dehidrogenasa	NP_626552	333	Secuencial
dehidrogenasa	YP_243655.2	297	Secuencial
dehidrogenasa	YP_241425.1	291	Secuencial
dehidrogenasa	AAY49635.1	332	Secuencial
dehidrogenasa	NP_765672.1	289	Secuencial

D. Preprocesamiento de las secuencias

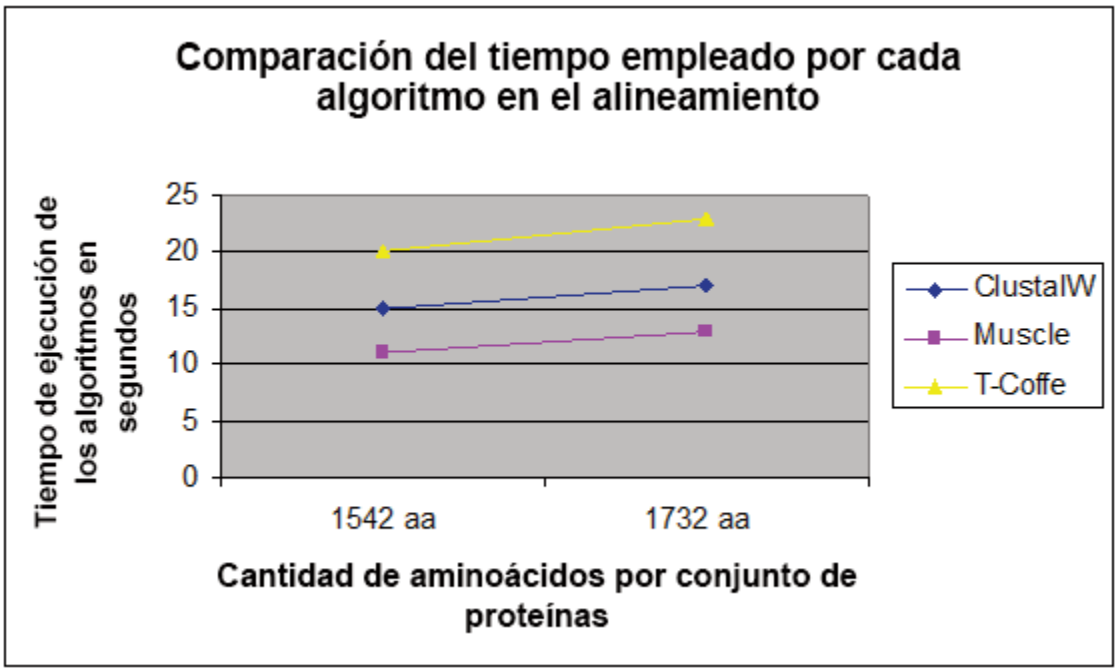
El proceso de conglomerado no se puede realizar sin antes trasladar las secuencias de proteínas a un formato aceptado por la comunidad científica y por consiguiente estándar para los tres algoritmos en estudio. Para fines de este estudio se recurrió a trasladar las secuencias anteriores a un formato tipo FASTA.

IV. Resultados experimentales

Una vez que se obtuvieron las secuencias en un formato aceptable para su respectiva ejecución, se prosiguió con el paso de conglomerado. Cabe señalar que no se cambiaron los valores predeterminados para el caso de ClustalW, pues estos cambios quedan a opción del investigador y en muchas ocasiones el cambio de los parámetros puede ser causa de errores graves en el alineamiento final, por consiguiente, las hipótesis que puedan surgir de dichos resultados serán también irrelevantes. De esta manera, el tiempo de ejecución empleado por cada algoritmo, al procesar los conjuntos de secuencia de proteínas en estudio, queda detallado en la siguiente tabla:

Tabla 2: Tiempo de ejecución de los algoritmos sujetos a análisis

Cantidad total de aa por tipo de proteína	Tiempo empleado por cada algoritmo en segundos		
	ClustalW	Muscle	T-Coffee
Secuencia de proteínas tipo FosB, 1732 aa	17	13	23
Secuencia de proteínas tipo dehigrogenasa, 1542 aa	15	11	20
Total de Tiempo E.	32	24	43



Esquema I: Representación gráfica del tiempo de ejecución empleado por cada algoritmo para cada conjunto de secuencias de proteínas

Con base en los resultados anteriores se estableció el siguiente cuadro comparativo:

Tabla 3: Tabla comparativa de los algoritmos estudiados

Aspectos Evaluados en el análisis	***ClustalW***	***Muscle***	***T-Coffee***
Nivel de Comprensión del alineamiento resultante	Alta, para Conjuntos de secuencias no mayores a 3. Baja si el conjunto de secuencias es mayor a 3.	Alta si el investigador es un experto en el área de estudio. En caso contrario baja.	Alta tanto para conjuntos pequeños de secuencias como para conjuntos grandes.
Capacidad de trabajar con conjuntos de secuencias grandes.	Depende de los parámetros introducidos por el investigador y de la longitud de las secuencias involucradas, por tanto es un sí condicional.	Sí, pero el investigador debe tener gran dominio en el área de investigación, es decir en el alineamiento múltiple de genes y proteínas, de lo contrario se comentarán errores en la interpretación de la formación resultante.	Sí, aunque esto implica mayor tiempo de ejecución
Capacidad de trabajar con conjunto de secuencias con errores en los datos de entrada	No, pues los errores cometidos en las primeras etapas del alineamiento por errores en los datos de entrada, serán reflejados en el alineamiento final	No	Sí
Recomendable para el caso del alineamiento múltiple de secuencias de proteínas.	Solo si el conjunto no excede de 3 secuencias	Solo para investigadores expertos.	Se recomienda aunque para el caso implique mayor tiempo de ejecución.

V. Discusiones

En el esquema No. 1 se observa el comportamiento de cada algoritmo en estudio al procesar las secuencias de proteínas. El algoritmo que duró menos fue Muscle, con 24 segundos aproximadamente. En cuanto a ClustalW, le correspondió el segundo lugar, con un total de 32 segundos, aunque hay que recalcar que el tiempo de ejecución de este algoritmo varía dependiendo de los parámetros introducidos por el investigador y del número de secuencias que se alinean. Por otro lado T-Coffee fue el algoritmo que tuvo el tiempo de ejecución mayor, 43 segundos aproximadamente, no obstante, los alineamientos resultantes presentaron mayor calidad y facilidad de interpretación.

Por tanto, dado los resultados en la etapa de experimentación, se determinó que existen diferencias significativas, en cuanto a sensibilidad y calidad biológica entre un algoritmo y otro, esto quiere decir que existe un elemento fundamental que, definitivamente, infiere en la calidad de información obtenida

a través del proceso experimental, ese elemento al que me refiero es el nivel de conocimiento que el usuario tiene acerca de su problema sujeto a investigación, por supuesto desde una perspectiva biológica y que involucra un área específica de la bioinformática, el tipo de dato, la fuente de esos datos y por supuesto el tamaño de los datos sujetos a experimentación.

VI. Conclusiones

- Para el alineamiento múltiple de proteínas se recomienda emplear T-Coffee, pues aunque tuvo el mayor tiempo de ejecución, presentó los resultados de forma más detallada y confiable desde el punto de vista biológico.
- ClustalW queda restringido para un alineamiento no mayor a las 3 secuencias. Esta restricción se da con el fin de prevenir errores en el alineamiento final.
- Muscle demostró un buen desempeño y por tanto es óptimo para el alineamiento múltiple de secuencias así como para el alineamiento de dos secuencias. No obstante, su uso queda limitado para expertos en el área.
- Se comprobó que la aplicación entre uno y otro algoritmo, en este caso basado en técnicas de conglomerados, incide definitivamente en los resultados obtenidos y que estos resultados facilitan o dificultan la interpretación de la información por parte del investigador y en consecuencia el desarrollo de la Bioinformática en cuanto a alineamiento y comparación de proteínas.

Referencias Bibliográficas

Apuntes de los cursos de Bases de Datos Avanzadas y Minería de Datos Avanzadas impartidas por el Dr. Carlos González Alvarado en el Instituto Tecnológico de Costa Rica durante el II semestre del 2005 y I semestre del 2006.

Curso de *Análisis de secuencias de proteínas y genes* del 9 al 13 de octubre del 2006, ITCR.

[Alts91] Altschul, S., et al, *" Aminoacid substitution matrices from an information theoretic perspective"*. J. Mol. Biol, 219:555565. 1991.

[Alts97] Altschul, S., et al, *"Gapped BLAST and PSI-BLAST a new generation of protein DB search programs"*. Nucleic Acids Res. 25:3389-3402. 1997.

[Carr88] Carrillo, H., et al, *"The multiple sequence alignment problems in Biology"*. SIAM J.Appl.Math, 48:10731082. 1988.

[Corp88]Corpet, F., et al, *"Multiple sequence alignments with hierarchical clustering"*. Nucleic Acids Res. 16:108810890. 1988.

[Day05] Day. A Roberta, *"Cómo escribir y Publicar una Tesis"*. 3ª.ed. Washington, D.C.: OPS, © 2005. (publicación científica y técnica No. 598).

[Hhei98] Sheikholeslami, G., et al, *"WaveCluster:A Multi-Rsolotion Clustering Aproach for Very Large Spacial Databases"*.Computer Science Dep. SUNY at Buffalo, NY, 1998.

[Higg92] Higgins, D., et al, *"Fast and sensitive multiple sequence alignments on a microcomputer"*. CABIOS, 5:151-153. 1992.

[Need70] Needleman, S., et al, *"A general method applicable to the search for similarities in the amino acid sequences of two proteins"*. J. Mol. Biol. 48:444-453. 1970.

[Notr00] Notredame, C., et al, *"T-Coffee:A novel method for multiple sequence alignments"*. Journal of Molecular Biology, Vol 302, pp205-217, 2000

[Pear88] Pearson, W., et al, *"Improved tools for biological sequence comparison*". Proc. Natl, Acad. Sci. USA 85:24442448. 1988.

[Robe04] Robert, Edgar., et al. *"MUSCLE: a multiple sequence alignment method with reduced time and space complexity."* BCM Bioinformatics. 2004

[Thom94] Thompson, J., et al. *"CLUSTAL W: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice"*. European Molecular Biology Laboratory, Heidelberg, Germany 94.

Sitios Web de interés

http://genomics.stanford.edu

http://bioinfo.cnio.es/sotaray/

http://www.ebi.ac.uk./microarray/

http://www.ebi.uniprot.org/uniprot-srv/uniProtClustrSearch.do

http://www.ebi.uniprot.org/uniprot- srv/uniProtInterproSearch.do