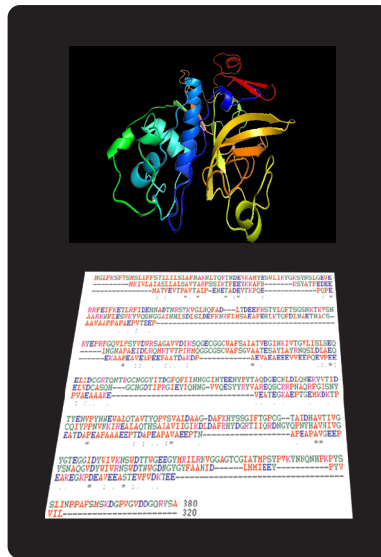


Alineamiento de Secuencias De Aminoácidos

3 Capítulo



Los alineamientos de secuencias de aminoácidos proporcionan una herramienta poderosa para comparar secuencias relacionadas, permitiendo detectar orígenes evolutivos similares y representar una estructura común y/o un rol catalítico.

Las inserciones y sustituciones de residuos singulares están generalmente enfatizadas en los alineamientos. Las inserciones aparecen representadas por caracteres nulos añadidos a una de las secuencias, las cuales pueden ser alineadas con letras en las otras (Rehm 2001). Existen dos tipos de alineamiento según el número de secuencias en estudio: El alineamiento por pares y el múltiple (ASM).

La búsqueda en las bases de datos con el objeto de extraer secuencias homólogas es el fundamento para el análisis de secuencias. Para cumplir este propósito una variedad de métodos han sido desarrollados y aplicados en amplios paquetes de pro-



UNIVERSIDAD DE CARTAGENA

K	H
S	Y
M	T
S	S
F	F
A	A
N	N
I	I
A	L
F	M
S	T
Y	Y
P	P
K	K
P	P
H	H
Q	D
N	N

N	N
D	Q
H	H
P	P
K	K
P	P
Y	Y
T	S
M	F
L	A
I	I
N	N
A	A
F	F
S	S
T	M
Y	S
H	K

gramas y servidores de Internet. Los programas de búsqueda en bases de datos difieren en la manera de cómo están diseñados los algoritmos que usan. Lo anterior tiene influencia en el tiempo de ejecución (velocidad) y la sensibilidad a la hora de realizar los alineamientos.

Los algoritmos de alta velocidad usan principios simplificados para establecer la similitud entre secuencias, en donde el tiempo que esta tarda en llevarse a cabo depende de la sensibilidad del algoritmo, estando fuertemente influenciado por la longitud de la secuencia y el tamaño de la base de datos. Por su parte, los algoritmos de Smith-Waterman (1981) están basados principalmente en métodos de programación dinámica, buscando óptimos en alineamientos locales de pares de secuencias. De esta forma el tiempo de cálculo es proporcional al cuadrado del tamaño de las secuencias comparadas, por lo tanto su velocidad es lenta para realizar búsquedas en grandes bases de datos. De otra parte, los programas que utilizan algoritmos FASTA (Pearson y Lipman, 1988) y BLAST (Alschul *et al.*, 1990) fueron desarrollados con el objeto de ser de alta velocidad y baja sensibilidad, respectivamente, ya que estos últimos están basados en estrategias heurísticas que concentran sus esfuerzos en las regiones de la secuencia más probablemente relacionadas (posición de mayor coincidencia entre la secuencias) en un tiempo de ejecución corto, ofreciendo buenos resultados. Las consultas a través de bases de datos públicas en el internet constituyen un recurso invaluable para investigadores que están trabajando en el campo de la biología molecular, química de proteínas, y diagnóstico molecular.

El inmenso número de secuencias de proteínas que pueden ser consultadas a través de bases de datos públicas en el internet es un recurso invaluable para investigadores que están trabajando en el campo de la biología molecular, química de proteínas diagnóstico clínico. Para optimizar el proceso de alineamiento, estos servidores permiten a los investigadores introducir sus secuencias y escoger varios parámetros, tales como valores de penalidad asociados con la inserción de gaps (espacios) y el tipo de matriz (blosum, pam, entre otras) (Gaskell, 2000).



La mayoría de métodos de alineamiento de secuencias busca optimizar el criterio de similitud. Hay dos modos de evaluarla, local y global. Los métodos locales intentan determinar si subsegmentos de secuencia (A) están presentes en otra (B).

Estos métodos tienen su máxima aplicabilidad en la recuperación y búsqueda en bases de datos (e.g Blast, Atschul *et al.*, 1990). A través de ellos es posible detectar secuencias con cierto grado de similitud que pueden o no ser homólogos. Los métodos globales hacen comparaciones alrededor de la longitud total de la secuencia.

El alineamiento de secuencias por pares y múltiple continúa siendo una de las áreas más activas de los recursos bioinformáticos y tiene por objeto encontrar la mejor similitud entre ellas.

ALINEAMIENTO DE SECUENCIAS POR PARES (PSA).

Como su nombre lo indica, consiste en comparar pares de secuencias. La utilización de este método no posibilita encontrar información crítica sobre la función de la proteína. En contraste, a partir del alineamiento de múltiples secuencias es posible sugerir un gen funcional.

ALINEAMIENTO DE MÚLTIPLES SECUENCIAS (MSA).

Los paquetes de programa disponibles en internet permiten deducir perfiles desde un alineamiento múltiple de secuencias. Un perfil es una matriz de sustitución específica para cada posición de la secuencia (position specific scoring matrix). Esta matriz tiene como dimensiones 20xL, siendo la longitud del alineamiento múltiple. A partir del mismo es construida dicha matriz teniendo en cuenta la frecuencia de los aminoácidos en cada posición así como sus propiedades fisicoquímicas.



UNIVERSIDAD DE CARTAGENA

N	N
Q	D
H	H
P	P
K	K
P	P
Y	Y
S	T
F	M
A	L
I	I
N	N
A	A
F	F
S	S
M	T
S	Y
K	H

N	N
D	Q
H	H
P	P
K	K
P	P
Y	Y
T	S
M	F
L	A
I	I
N	N
A	A
F	F
S	S
T	M
Y	S
H	K

La mayoría de las familias de secuencias conservan ciertos residuos críticos y motivos. Esta información permite incrementar la sensibilidad en la búsqueda de bases de datos. La mayor parte de programas de perfiles está basada en los modelos de Markov ocultos (HMMs: Hiden Markov models). Un HMM es entrenado a partir de diversas observaciones en las que puede esperarse que las posibles variaciones hayan sido generadas. Su principal ventaja es que tienen una base probabilística muy sólida (Eddy *et al.*, 1998).

Algunas herramientas bioinformáticas de Internet que emplean métodos de alineamiento de secuencias de proteínas son descritas a continuación:



Blast (Herramientas de Búsqueda de Alineamientos Locales).
URL: <http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi>

ENTIDAD ADMINISTRADORA

Centro Nacional de Información en Biotecnología (NCBI).

DESCRIPCIÓN

Blast es un conjunto de programas que tiene por objeto obtener similitudes entre secuencias alineadas. Está diseñado para explorar todas las bases de datos disponibles independientemente de que sean proteínas o ADN. El fundamento de los algoritmos de BLAST es comparar secuencias creando matrices de sustitución generales, como por ejemplo Blosum 62 (Block Substitution Matrices), en las que son propuestos cuáles son los aminoácidos que menos difieren y las mutaciones más frecuentes. A partir de estas matrices es establecida una puntuación (Score), la cual indica el grado de similitud entre pares de secuencias (McGinnis *et al.*, 2004).

Un resultado típico obtenido a partir de BLAST al someter la secuencia de la Actinidina en formato Fasta, descrita con anterioridad, es presentado en la Figura 3.11.3.1



Figura 3.1. El encabezado cita la versión del algoritmo empleado y su referencia, el nombre y la longitud de la secuencia analizada y la base de datos utilizada como blanco. Aparece además, el código RID (Request ID), el cual permite recuperar los resultados de búsqueda dentro de las 24 horas siguientes (Figura 3.1A).

En la segunda parte figuran identificados los dominios putativos de la proteína blanco, en este caso el dominio Peptidase_C:1A, presente en cisteína proteasas (CPs) similares a la Papaína (Figura 3.1B).

BLASTP 2.2.12 [Aug-07-2005]

Reference:

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.* 25:3389-3402.

RID: 1135109549-7486-189822402772.BLASTQ1

Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples
3,114,327 sequences; 1,070,851,917 total letters

If you have any problems or questions with the results of this search please refer to the [BLAST FAQ](#) [Taxonomy reports](#)

Query=
(254 letters)

3.1A.



UNIVERSIDAD DE CARTAGENA

K	H
S	Y
M	T
S	S
F	F
A	A
N	N
I	I
A	L
F	M
S	T
Y	Y
P	P
K	K
P	P
H	H
Q	D
N	N

N	N
D	Q
H	H
P	P
K	K
P	P
Y	Y
T	S
M	F
L	A
I	I
N	N
A	A
F	F
S	S
T	M
Y	S
H	K

La tercera sección muestra una representación gráfica de los alineamientos diferenciando por colores los porcentajes de identidades (Figura 3.1C). La cuarta parte del informe presenta un resumen de todas las secuencias que produjeron alineamientos significativos junto con un valor de puntuación (Score), lo cual, expresa el grado de similitud entre pares de secuencias. El valor E para una determinada puntuación indica cuántos alineamientos esperamos que por azar alcancen un valor igual o mayor y está dado por la Ecuación 1:

$$E = Kmn e^{-\lambda S}$$

ECUACIÓN 1.

K y lambda (λ) son dos parámetros determinados empíricamente, m y n correspondan las longitudes de las secuencias y S es la puntuación del alineamiento). Por lo tanto un valor de E de 0.001 (valor establecido por defecto para las búsquedas con Blast) significa que hasta uno de cada 1000 alineamientos pueden haberse dado al azar. Es recomendable obtener valores inferiores a 0.00001. Sin embargo, lo mejor siempre es examinar los alineamientos en detalle, hacer alineamientos múltiples y emplear análisis filogenéticos para confirmar si las secuencias involucradas en el estudio están relacionadas de manera evolutiva. Como es posible apreciar en nuestro ejemplo, los valores E de los alineamientos están en el orden de 1E-70, lo cual es menos frecuente de encontrar, y resulta muy confiable en cuanto a que las secuencias alineadas estén evolutivamente relacionadas. Como era de esperarse, todas estas secuencias pertenecen a una misma familia denominada cisteína proteasas (Figura 3.1D).

La quinta parte muestra los alineamientos detallando los valores de las puntuaciones, identidad y valor E (Figura 3.1E). La última pantalla del informe señala los parámetros utilizados en la búsqueda (Figura 3.1F).



> [gi|15984|emb|CAA34486.1](#) unnamed protein product [Actinidia deliciosa]
[gi|113285|sp|F00785|ACTN_ACTCH](#) Actinidain precursor (Actinidin) (Allergen Act c 1)
 Length=380

Score = 520 bits (1338), Expect = 1e-146
 Identities = 254/254 (100%), Positives = 254/254 (100%), Gaps = 0/254 (0%)

```

Query 1  LPSYVDWRSAGAVVDIKSQEGCGWAFSAIATVEGINKIVTGVLIISLSEQELIDCGRTQ 60
Sbjct 127 ..... 186

Query 61  NTRGCNGGYITDGFQFYXXXXXXXXXXTEENYPYTAQDGECNLDLQNEKYVITIDTYENVPYNN 120
Sbjct 187 ..... 246

Query 121  EWALQTAVTYQPVSVALDAAGDAFKHYSSGIFTGPGCTAIDHAVTIVGYGTEGGIDYIV 180
Sbjct 247 ..... 306

Query 181  KNSWDTT@GEEGYMRILRNVGAGTCGIATMPSYYPVKYNNQNHKPYSSLINEPPAFSMK 240
Sbjct 307 ..... 366

Query 241  DGPVGVDDGQRYSA 254
Sbjct 367 ..... 380
  
```

3.1E.

```

Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+ERF excluding
environmental samples
Posted date: Dec 19, 2005 0:04 AM
Number of letters in database: 534,625,651
Number of sequences in database: 1,367,685
Lengths: K H
0.315 0.135 0.428
Gapped
Lengths: K H
0.267 0.0410 0.140
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 1367685
Number of Hits to DB: 54579315
Number of extensions: 2294193
Number of successful extensions: 6199
Number of sequences better than 1f: 285
Number of HSP's better than 10 without gapping: 273
Number of HSP's gapped: 5605
Number of HSP's successfully gapped: 310
Number of extra gapped extensions for HSP's above 10: 5281
Length of query: 254
Length of database: 534605651
Length adjustment: 122
Effective length of query: 132
Effective length of database: 534605651
Effective search space: 0056744392
Effective search space used: 48542746692
P: 11
A: 40
X1: 16 (7.3 bits)
X2: 58 (14.6 bits)
X3: 64 (24.7 bits)
s1: 42 (23.8 bits)
s2: 72 (32.3 bits)
  
```

3.1F.

Figura 3.1. Resultado de la búsqueda Blast para una secuencia de proteína (Actinidina de 30 kD).

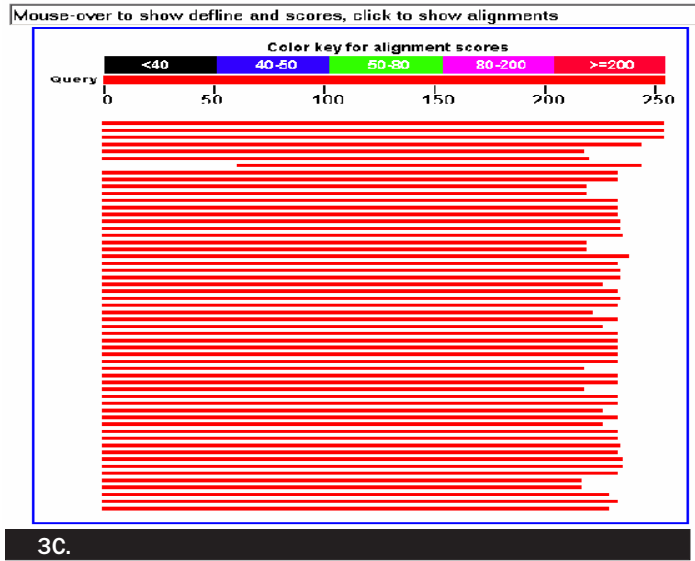


UNIVERSIDAD DE CARTAGENA

K	H
S	Y
M	T
S	S
F	F
A	A
N	N
I	I
A	L
F	M
S	T
Y	Y
P	P
K	K
P	P
H	H
Q	D
N	N

N Q
 D Q
 H H
 P P
 K K
 P P
 Y Y
 T S
 M F
 L A
 I I
 N N
 A A
 A F
 F F
 S S
 S S
 T M
 Y S
 H K

Distribution of 501 Blast Hits on the Query Sequence



Related Structures

Sequences producing significant alignments:	Score (Bits)	E Value
gi 15984 emb CAA34486.1 unnamed protein product [Actinidia d...	520	1e-146
gi 12744965 qb AAK06862.1 actinidin protease [Actinidia chinens	515	3e-145
gi 21444501 pir TAGB actinidain (EC 3.4.22.14) precursor - ki...	508	3e-143
gi 15957 emb CAA31435.1 actinidin precursor [Actinidia chine...	481	4e-135
gi 442619 pdb IABC Actinidin (E.C.3.4.22.14) Complex With T...	439	2e-122
gi 230417 pdb 2ACT Actinidin (Sulphydryl Proteinase) (E.C. Num	423	1e-117
gi 15959 emb CAA31529.1 actinidin precursor [Actinidia chine...	371	8e-102
gi 2425066 qb AAB88263.1 cysteine proteinase Mir3 [Zea mays]	281	8e-75
gi 2828252 emb CAA05894.1 CYP1 [Lycopersicon esculentum] >gi...	278	5e-74
gi 255032 qb AAB23155.1 COT44=cysteine proteinase homolog [B...	278	5e-74
gi 81682 pir JQ1121 cysteine proteinase (EC 3.4.22.-) COT44 ...	278	5e-74
gi 6682829 dbj BAA88898.1 cysteine protease component of pro...	278	9e-74
gi 4731372 qb AAD28476.1 papain-like cysteine protease [Sanders	277	2e-73
gi 18141285 qb AAL60580.1 senescence-associated cysteine protea	274	1e-72
gi 50929537 ref XP_474296.1 OSJNBa0043A12.33 [Oryza sativa (...	273	2e-72
gi 62526575 qb AAK84673.1 cysteine protease CP1 [Manihot escule	273	2e-72
gi 30141019 dbj BAC75923.1 cysteine protease-1 [Helianthus annu	273	3e-72
gi 15234557 ref NP_195406.1 cysteine-type endopeptidase/ cys...	272	5e-72
gi 15290508 qb AAK92229.1 cysteine proteinase [Arabidopsis thal	272	5e-72
gi 50355617 dbj BAD29957.1 cysteine protease [Daucus carota]	271	9e-72
gi 5777889 emb CAB53515.1 cysteine protease [Solanum tuberosum]	271	1e-71
gi 2425064 qb AAB88262.1 cysteine proteinase Mir2 [Zea mays]	270	1e-71
gi 218183 dbj BAA14403.1 unnamed protein product [Oryza sati...	270	1e-71
gi 218181 dbj BAA14402.1 unnamed protein product [Oryza sati...	270	1e-71

3D.



Direcciones de otros métodos de alineamiento por pares de secuencias de proteínas.

NCBI **BLAST 2 sequences** BLAST Entrez ?

Blast 2 Sequences

URL: <http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi>
(Tatusova et al., 1999).

ch.EMBnet.org
Home Services Courses Links Contacts

LALIGN - find multiple matching subsegments in two sequences



LALIGN

URL: http://www.ch.embnet.org/software/LALIGN_form.html
(Huang y Miller, 1991)



UNIVERSIDAD DE CARTAGENA

K	H
S	Y
M	T
S	S
F	F
A	A
N	N
I	I
A	L
F	M
S	T
Y	Y
P	P
K	K
P	P
H	H
Q	D
N	N

N	N
D	Q
H	H
P	P
K	K
P	P
Y	Y
T	S
M	F
L	A
I	I
N	N
A	A
F	F
S	S
T	M
Y	S
H	K

MÉTODOS DE ALINEAMIENTOS MÚLTIPLES



Clustal W

URL: <http://www.ebi.ac.uk/Tools/clustalw/>

ENTIDAD ADMINISTRADORA

Laboratorio Europeo de Biología Molecular (EMBL), Heidelberg, Alemania.

DESCRIPCIÓN

Clustal W es un programa que aplica métodos de alineamiento globales de alta velocidad (Wilbur y Lipman, 1983) para calcular los niveles de semejanza entre las secuencias. Por ello no es aconsejable alinear secuencias con largos sectores disímiles. Además, ha mostrado no funcionar muy bien en secuencias que presentan baja homología o en secuencias de dominios conservados en medio de zonas de baja homología (Lassmann y Sonnhammer, 2002). Sin embargo, la calidad de los alineamientos es aceptable, y permite alinear algunos cientos de proteínas (Barton y Stenberg, 1987; Taylor, 1986).

Ejemplo: Con el objeto de comparar las secuencias reportadas de las estructuras cristalinas de la Actinidina depositadas en las base de dato Protein Data Bank (PDB), bajo los códigos 1aec y 2act, con la secuencia alergénica almacenada en la base de datos SDAP (Structural Database of Allergenic Proteins) con código de acceso Act c 1 será realizado un alineamiento de múltiples secuencias.

Las secuencias de la Actinidina, 1aec, 2act y act c 1, son muy parecidas por tratarse de una misma proteína y por ende es aconsejable utilizar una metodología de alineamiento global, proporcionada por Clustal W.



El archivo de salida del programa Clustal W es presentado en la Figura 3.2. El reporte está dividido en tres partes, la primera consiste en un resumen de los resultados de búsqueda, en donde es indicado el número, formato y tipo de secuencia utilizada como dato de entrada y los archivos de salida, entre otros. Genera una lista de vínculos que proveen los archivos de alineamientos (Figura 3.2A). En la segunda aparece una tabla de identidades obtenidas de los alineamientos entre las secuencias, la cual aparece clasificada por el puntaje de identidad, número, nombre y longitud de A.A. en la secuencia.

Results of search	
Number of sequences	3
Alignment score	4060
Sequence format	Pearson
Sequence type	aa
ClustalW version	1.82
JaView	<input type="button" value="JaView"/>
Output file	clustalw-20060127-21165588_output
Alignment file	clustalw-20060127-21165588_aln
Guide tree file	clustalw-20060127-21165588_dnd
Your input file	clustalw-20060127-21165588_input
<input type="button" value="SUBMIT ANOTHER JOB"/>	

3.2A.

Scores Table

Sort by

SeqA Name	Len(aa)	SeqB Name	Len(aa)	Score
1 Act	254	2 1AEC.	218	99
2 1AEC.	218	3 2ACT.	220	93
1 Act	254	3 2ACT.	220	92

3.2B.



K H
S Y
M T
S S
A F A F
A A
N N
I I
A L
F M
S T
Y Y
P P
K K
P P
H H
Q D
N N

N	N
D	Q
H	H
P	P
K	K
P	P
Y	Y
T	S
M	F
L	A
I	I
N	N
A	A
F	F
S	S
T	M
Y	S
H	K

Alignment

Hide Colors

View Alignment File

CLUSTAL W (1.82) multiple sequence alignment

```

Act      LPSYVDWRSAGAVVDIKSQGECGGCWAFSAIATVEGINKIVTGVLSLSEQLIDCGRTQ 60
1AEC.    LPSYVDWRSAGAVVDIKSQGECGGCWAFSAIATVEGINKIVTGVLSLSEQLIDCGRTQ 60
2ACT.    LPSYVDWRSAGAVVDIKSQGECGGCWAFSAIATVEGINKITSGSLISLSEQLIDCGRTQ 60
*****:*****

Act      NTRGCNGGYITDGFQFIINNGGINTEENYPYTAQDGECLDLQNEKYVTIDTYENVPYNN 120
1AEC.    NTRGCNGGYITDGFQFIINNGGINTEENYPYTAQDGECLDLQNEKYVTIDTYENVPYNN 120
2ACT.    NTRGCNGGYITDGFQFIINNGGINTEENYPYTAQDGDVALQDQKYVTIDTYENVPYNN 120
*****:*****

Act      EWALQTAVTYQPVSVLDAAGDAFKHYSSGIFTGPCGTVIDHAVTIVGYGTEGGIDYWIV 180
1AEC.    EWALQTAVTYQPVSVLDAAGDAFKQYSSGIFTGPCGTVIDHAVTIVGYGTEGGIDYWIV 180
2ACT.    EWALQTAVTYQPVSVLDAAGDAFKQYASGIFTGPCGTVIDHAVTIVGYGTEGGVDYWIV 180
*****:*****

Act      KNSWDTIWGEEGYMRILRNVGGAGTCGIATMPSPVKYNNQNHKPYSSLINPPAFMSMK 240
1AEC.    KNSWDTIWGEEGYMRILRNVGGAGTCGIATMPSPVKY----- 218
2ACT.    KNSWDTIWGEEGYMRILRNVGGAGTCGIATMPSPVKYNN----- 220
*****:*****

Act      DGPVGVDDGQRYS 254
1AEC.    -----
2ACT.    -----

```

3.2C.

Figura 3.2. Resultado de un alineamiento de múltiples secuencias por el servidor Clustal W. A) Resultado de búsqueda. B) Tabla de Puntuaciones o identidades (Score). C) Alineamiento.

Los colores presentan la propiedad fisicoquímica del residuo. El rojo, verde y azul corresponden a residuos hidrofóbicos, hidrofílicos y polares con carga, respectivamente. Además, cada residuo está asociado con un símbolo que representa el tipo de alineamiento. Así, el asterisco (*) indica que existe alineamiento correcto entre los residuos, los dos puntos (:) designan residuos superpuestos con propiedades estructurales y físicoquímicas similares, el punto (.) sugiere que no hay superposición entre todas las secuencias involucradas en el alineamiento y la línea (-) es empleada para mostrar que existe por lo menos un residuo que no está alineado.



Los resultados del alineamiento múltiple son presentados en la última sección. Muestran que la 1 Aec tiene una mayor identidad (99%) que la 2 Act (92.2%) con respecto a la secuencia de la Actinidina de 30 kD (Act). Esto es debido a que la 2 Act tiene 17 sustituciones y la 1 Aec solamente presenta 2 en las posiciones 100 y 146 (Figuras 3.2B y 3.2C).

Una ventaja importante de Clustal W es que permite apreciar el alineamiento entre las secuencias con opciones a color o blanco y negro.

MAFFT: A Program for Multiple Sequence Alignment



MAFFT: Un Programa para Alineamiento de Secuencias Múltiples.
URL: <http://bioinformatics.uams.edu/mafft/>

ENTIDAD ADMINISTRADORA

Instituto de Tecnología Microbial, India. Desarrollado por el grupo del Dr G. Raghava.

DESCRIPCIÓN

MAFFT inicialmente realiza un alineamiento por métodos progresivos y luego es refinado por métodos iterativos. Para los métodos de alineamiento progresivo es utilizada una aproximación de las transformadas de Fourier (FT). MAFFT es uno de los métodos más rápidos entre las herramientas de alineamiento múltiple actualmente disponibles y ha sido utilizado en varios proyectos tales como Pfam (base de datos de familia de proteínas), Astral (Compendio de bases de datos y herramientas para el análisis de estructuras proteicas) y Merops (Base de datos de peptidasas).

La versión 5 de MAFFT generó una mayor exactitud que otros métodos de amplio uso, incluyendo la versión 2 de Tcoffe y Clustal W en un test de prueba consistente en más de 50 secuencias alineadas (Katoch *et al.*, 2002; Katoh *et al.*, 2005).



UNIVERSIDAD DE CARTAGENA

K	H
S	Y
M	T
S	S
F	F
A	A
N	N
I	I
A	L
F	M
S	T
Y	Y
P	P
K	K
P	P
H	H
Q	D
N	N

N	N
D	Q
H	H
P	P
K	K
P	P
Y	Y
T	S
M	F
L	A
I	I
N	N
A	A
F	F
S	S
T	M
Y	S
H	K



FACULTÉS UNIVERSITAIRES
NOTRE-DAME DE LA PAIX
NAMUR



Servidor Web Match-Box 1.3

URL: http://www.fundp.ac.be/sciences/biologie/bms/matchbox_submit.shtml

ENTIDAD ADMINISTRADORA

Unidad de Recursos de Biología Molecular, Universidad de Namur, Bélgica.

DESCRIPCIÓN

El programa Match-Box propone ser una herramienta para el análisis de múltiples secuencias basadas en un estricto criterio estadístico. Es particularmente conveniente para encontrar y alinear motivos estructurales conservados.

PROBCONS



PROBCONS

URL: <http://probcons.stanford.edu/>

ENTIDAD ADMINISTRADORA

Departamento de Ciencias Computacionales, Universidad de Stanford, California. USA. Grupo de Sarafim Batzoglou.

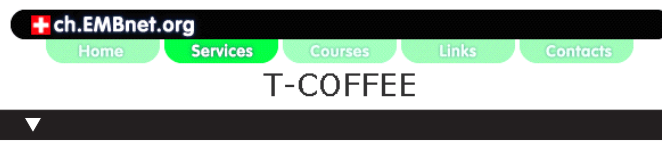
DESCRIPCIÓN

ProbCons es una herramienta basada en una combinación de modelos probabilísticos y técnicas de alineamiento, los cuales emplean una nueva función de puntuación para comparar múltiples secuencias. El alineamiento producido por ProbCons posee una mejor significancia estadística que los programas actuales, generando en promedio 7%, 11% y 14% más de columnas correctamente alineadas que los programas T-Coffe, CLUSTAL W y DIALING, respectivamente (Do *et al.*, 2005).

UNIVERSIDAD DE CARTAGENA



Existen servidores que emplean métodos combinados de alineamiento por pares y locales (Notredame *et al.*, 2000), comparan dos alineamientos múltiples diferentes (Morgenstem *et al.*, 2003), producen gráficos multivalentes que combinan alineamientos, filogenia, análisis estructural e información de la estructura secundaria (Joachimiak *et al.*, 2002; Simossis *et al.*, 2005; y Zhou *et al.*, 2005) y realizan estudios ontológicos (Thom-pson *et al.*, 2005), entre otros, tales como:



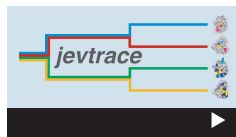
T-COFFEE
 URL: <http://www.ch.embnet.org/software/TCoffee.html>
 (Notredame *et al.*, 2000).



AltAVist
 URL: <http://bibiserv.techfak.uni-bielefeld.de/altavist/>
 (Morgenstem *et al.*, 2003).



PRALINE
 URL: <http://zeus.cs.vu.nl/programs/pralinewww/>
 (Simossis *et al.*, 2005)



JEvTrace
 URL: <http://www.cmpharm.ucsf.edu/~marcinj/JEvTrace/>
 (Joachimiak *et al.*, 2002).



UNIVERSIDAD DE CARTAGENA

K	H
S	Y
M	T
S	S
F	F
A	A
N	N
I	I
A	L
F	M
S	T
Y	Y
P	P
K	K
P	P
H	H
Q	D
N	N

N	N
D	Q
H	H
P	P
K	K
P	P
Y	Y
T	S
M	F
L	A
I	I
N	N
A	A
F	F
S	S
T	M
Y	S
H	K



MAO

URL: <http://bips.ustrasbg.fr/LBGI/MAO/mao.html> (Thompson et al., 2005).



SPEM

URL: http://sparks.informatics.iupui.edu/Softwares-Services_files/spem.htm (Zhou et al., 2005)

VISUALIZACIÓN DE ALINEAMIENTOS

Visualizar los alineamientos con buenos programas siempre facilita el análisis. Algunos ejemplos de Visores de Alineamientos son:



INSTITUTO DE KAROLINSKA, Centro de Investigación en Genómica.
URL: <http://bioinformatics.abc.hu/tothg/biocomp/other/Belvu.html>

ENTIDAD ADMINISTRADORA

Universidad de Queen. Ontario, Canada.

DESCRIPCIÓN

Belvu es un visor para alineamiento de múltiples secuencias. Una de sus ventajas es que posee un sistema de distinción de residuos conservados y por tipo de residuos en el alineamiento.

LALNVIEW :

A graphical viewer for pairwise alignments



LALNVIEW

URL: <http://www.expasy.ch/tools/lalnview.html>

UNIVERSIDAD DE CARTAGENA



ENTIDAD ADMINISTRADORA

Departamento de Bioquímica Médica, Universidad de Ginebra, Suiza.

DESCRIPCIÓN

Lalnview es un programa gráfico para visualizar alineaciones locales entre dos secuencias. Las secuencias son representadas por rectángulos coloreados para dar un bosquejo total de las semejanzas entre las mismas. Los bloques de semejanza entre las dos secuencias aparecen coloreados según el grado de identidad entre los dos segmentos (Duret *et al.*, 1996).



ESPrict 1.8 (Easy Sequencing in Postscript)
URL: <http://bioinfo.hku.hk/doc/ESPrict/>

ENTIDAD ADMINISTRADORA

Creado por Patrice Gouet, Laboratorio de Biofísica (Oxford, OX1 3QU, UK) y Frédéric Metz, Instituto de Biología Molecular (Francia).

DESCRIPCIÓN

El programa ESPrict permite la visualización rápida de secuencias alineadas de programas populares tales como Clustal W o GCG PILEUP en un archivo de salida en formato Postscript. Puede leer archivos generados por diferentes métodos de asignación de estructuras secundarias tales como DSSP (Data base of Secondary Structure in Proteins) (Kabsch y Sander, 1983), STRIDE (Frishman y Argos, 1996) y PHD (Rost, 1996). El archivo de salida del programa ESPrict 1.8 muestra datos de estructuras secundarias, secuencias alineadas, una puntuación que indica el grado de similaridad entre pares de residuos alineados, datos de accesibilidad, hidropaticidad, contactos intermoleculares, entre otros (Gouet *et al.*, 1999).



UNIVERSIDAD DE CARTAGENA

K	H
S	Y
M	T
S	S
F	F
A	A
N	N
I	I
A	L
F	M
S	T
Y	Y
P	P
K	K
P	P
H	H
Q	D
N	N

N	N
D	Q
H	H
P	P
K	K
P	P
Y	Y
T	S
M	F
L	A
I	I
N	N
A	A
F	F
S	S
T	M
Y	S
H	K

REFERENCAS

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215(3):403-410.
- Barton, G.J., Sternberg, M.J.E. (1987). Evaluation and improvements in the automatic alignment of protein sequences. *Protein Eng.* 1(2):89-94.
- Caffrey, D.R., Dana, P.H., Mathur, V., Ocano, M., Hong, E.J., Wang, Y.E., Somaroo, S., Caffrey, B.E., Potluri, S., Huang, E.S. (2007). PFAAT version 2.0 : A tool for editing, annotating, and analyzing multiple sequence alignments. *BMC Bioinformatics.* 8(1):381 [Epub ahead of print].
- Chung, Y.S., Lee, W.H., Tang, C.Y., Lu C.L. (2007). RE-MuSiC: a tool for multiple sequence alignment with regular expression constraints. *Nucleic Acids Res.* 35(Web Server issue):639-44.
- Do, C., Mahabhashyam, M., Brudno, M., Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* 15(2):330-340.
- Duret, L., Gasteiger, E., Perriere, G. (1996). LALNVIEW: a graphical viewer for pairwise sequence alignments. *Comput. Appl. Biosci.* 12(6):507-510.
- Eddy, S.R. (1998). Profile hidden Markov models. *Bioinformatics.* 14(9):755-763.
- Frishman, D., Argos, P. (1996). Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng.* 9(2):133-142.
- Gaskell, J.G. (2000). Multiple Sequence Alignment Tools on the Web. *BioTechniques. Appl. Microbiol. Biotechnol.* 57(1):579-592.
- Goode, M.G., Rodrigo, A.G. (2007). SQUINT: a multiple alignment program and editor. *Bioinformatics.* 23(12):1553-1555.
- Gouet, P., Courcelle, E., Stuart, D.I., Metz, F. (1999). ESPript: multiple sequence alignments in PostScript. *Bioinformatics.* 15(4): 305-308.
- Huang X, y Miller W (1991) A time-efficient, linear-space local similarity algorithm. *Adv Appl Math* 12: 337-357



• Huska, M.R., Buschmann, H., Andrade-Navarro M.A. 2007. BiasViz: Visualization of amino acid biased regions in protein alignments. *Bioinformatics*. Oct 6; [Epub ahead of print].

• Joachimiak, M., Cohen, F. (2002). JEvTrace: refinement and variations of the evolutionary trace in JAVA. *Genome Biol.* 3(12):1-12.

• Kabsch, W., Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* 22(12):2577-637.

• Katoh, K., Misawa, K., Kuma, K., Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30(14):3059-3066.

• Katoh, K., Kuma, K., Toh, H., Miyata T. (2005). MAFFT version 5: improvement in accuracy of Multiple sequence alignment. *Nucleic Acids Res.* 33(2):511-518.

• Lassmann, T., Sonnhammer, E.L. (2002). Quality assessment of multiple alignment programs. *FEBS Lett.* 529(1):126-130.

• Lassmann, T., Sonnhammer, E.L. (2007). Automatic extraction of reliable regions from multiple sequence alignments. *BMC Bioinformatics.* 8 Suppl 5:S9.

• McGinnis, S., Madden, T.L. (2004). BLAST: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* 32(Web Server issue):20-25.

• Morgenstern, B., Goel, S., Sczyrba, A., Dress, A. (2003). AltAVisT: Comparing alternative multiple alignments. *Bioinformatics.* 19(3):425-426.

• Notredame, C., Higgins, D.G., Heringa, J. (2000). T-Coffee: A novel method for multiple sequence alignments. *J. Mol. Biol.* 302(1): 205-217.

• Pearson, W.R., Lipman D.J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA.* 85(8):2444-2448.

• Pei, J., Grishin, N.V. (2006). MUMMALS: multiple sequence alignment improved by using hidden Markov models with local structural information. *Nucleic Acids Res.* 34:4364-4374

• Pei, J., Grishin N.V. (2007). PROMALS: towards accurate multiple sequence alignment of distantly related proteins. *Bioinformatics.* 23(7):802-808.



UNIVERSIDAD DE CARTAGENA

K	H
S	Y
M	T
S	S
F	F
A	A
N	N
I	I
A	L
F	M
S	T
Y	Y
P	P
K	K
P	P
H	H
Q	D
N	N

N	N
D	Q
H	H
P	P
K	K
P	P
Y	Y
T	S
M	F
L	A
I	I
N	N
A	A
F	F
S	S
T	M
Y	S
H	K

•Rehm, B.H.A. (2001). Bioinformatic tools for DNA/protein sequence analysis, functional assignment of genes and protein classification. *Appl. Microbiol. Biotechnol.* 57(5-6):579-592.

•Rost, M. (1996). PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* 266:525-539.

•Simossis, V., Heringa, J. (2005). PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Research.* 33 (Web Server issue):289-294.

•Smith, F., Waterman, S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* 147(1):195-197.

•Taylor, W.R. (1986). Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.* 188(2):233-258.

•Tatusova, T.A., Madden, T.L. (1999). Blast 2 sequences - a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* 174(2):247-250.

•Thompson, J., Holbrook, S., Katoh, K., Koehl, P., Moras, D., Westhof, E., Poch, O. (2005). MAO: a Multiple Alignment Ontology for nucleic acid and protein sequences. *Nucleic Acids Res.* 33(13): 4164-4171.

•Webb, B., Liu, J., Lawrence, C. (2002). BALSAs: Bayesian algorithm for local sequence alignment. *Nucleic Acids Res.* 5(5): 1268-1277.

•Wilbur, W.J., Lipman, D.J. (1983). Improved tools for biological sequence Comparison. *Proc. Natl. Acad. Sci. USA.* 80:726-730.

•Zhou, H., Zhou, Y. (2005). SPEM: Improving multiple-sequence alignment with sequence profiles and predicted secondary structures. *Bioinformatics.* 21(18):3615-3621.

•Zhou H., Zhou, Y. 2007. SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. *Bioinformatics.* 21:3615-3621.

