Tema 2: Alineamientos pareados y búsqueda de homólogos en bases de datos mediante BLAST

Introducción a la filoinformática – pan-genómica y filogenómica. Diplomado en Bioinformática UAS, Sinaloa, México, Octubre 2021

Introducción a la Filoinformática: Pan-genómica y Filogenómica – Diplomado en Bioinformática, Univ. Autónoma de Sinaloa, Oct. 2021
http://econtinua.uas.edu.mx/diplomados/2510-2700-18_001.htm

Pablo Vinuesa (vinuesa[at]ccg . unam . mx; @pvinmex)
Centro de Ciencias Genómicas, CCG-UNAM, Cuernavaca, México
http://www.ccg.unam.mx/~vinuesa/

Todo el material del curso (presentaciones, tutorales y datos) lo encontrarás en:
https://github.com/vinuesa/intro-filoinfo_UAS

- **Tema 2: alineamientos pareados y búsqueda de homólogos en bases de datos mediante BLAST**
- evolución de secuencias y **clasificación de mutaciones**
- **indeles** y **gaps**
- **alineamientos globales** (Needleman-Wunsch) vs. **locales** (Smith-Waterman);
- **programación dinámica**;
- **dot plots**;
- **matrices de costo de sustitución, penalización de gaps** y cuantificación de la similitud;
- evaluación estadística de la similitud entre pares de secuencias;
- escrutinio de bases de datos mediante **BLAST**; Búsquedas a nivel de **DNA vs. AA**;
- la **familia BLAST** e interpretación de resultados de **búsqueda de secuencias homólogas**
- prácticas: uso de **NCBI BLAST en línea**

Protocolo básico para un análisis filogenético de secuencias moleculares

Tema 2: alineamientos pareados, búsquedas de homólogos en bases de datos

Colección de secuencias homólogas
⇩ · BLAST, diamond

Alineamiento múltiple de secuencias
⇩ · clustalo, mafft, muscle ...

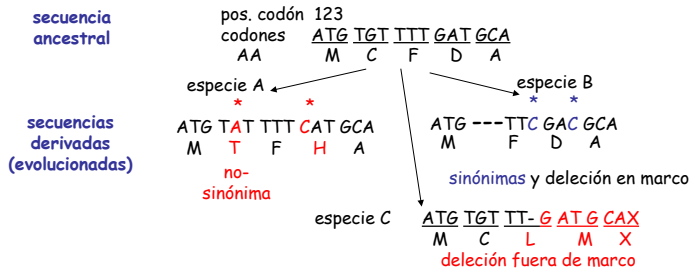
Análisis evolutivo del alineamiento y selección del modelo de sustitución más ajustado
⇩ · tests de saturación, modeltest, ...

Estima filogenética
⇩ · NJ, ME, MP, ML, Bayes ...

Pruebas de confiabilidad de la topología inferida
⇩ · proporciones de bootstrap probabilidad posterior ...

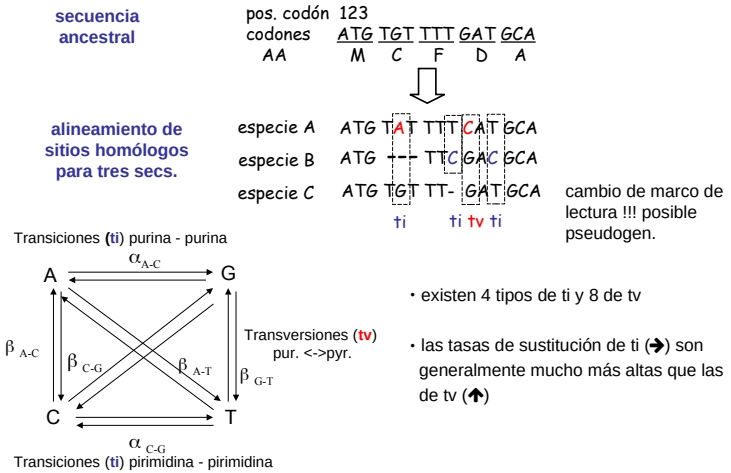
Interpretación evolutiva y aplicación de las filogenias

Homología entre secuencias de DNA y proteína: tipos de mutaciones en secs. codificadoras de proteínas



- Todas las mutaciones en 2as posiciones resultan en sustituciones no sinónimas
- 96% de mutaciones en 1as posiciones resultan en sustituciones no sinónimas
- Casi todas las sustituciones sinónimas ocurren en las 3as posiciones
- las deleciones o inserciones en secs. codificadoras de aa suceden generalmente en múltiplos de tres nt; de no ser así se generan cambios de marco de lectura corriente abajo de la mutación, con frecuencia generando un pseudogen no funcional

Homología entre secuencias de DNA y proteína: alineamiento y tipos de mutaciones



- existen 4 tipos de ti y 8 de tv
- las tasas de sustitución de ti (➔) son generalmente mucho más altas que las de tv (➚)

Tema 2: Alineamientos pareados y búsqueda de homólogos en bases de datos mediante BLAST

Introducción a la filoinformática – pan-genómica y filogenómica. Diplomado en Bioinformática UAS, Sinaloa, México, Octubre 2021

Alineamientos pareados y búsqueda de homólogos en bases de datos

Los alineamientos pareados son la base de lo métodos de búqueda de secuencias homólogas en bases de datos

- Si dos proteínas o genes se parecen mucho a lo largo de toda su longitud asumimos que se trata de proteínas o genes homólogos, es decir, descendientes de un mismo ancestro común (cenancestro).
- Por ello una de las técnicas más utilizadas para detectar potenciales homólogos en bases de datos de secuencias se basa en la **cuantificación de la similitud entre pares de secuencias** y la determinación de la **significancia estadística** de dicho parecido. Estas magnitudes son las que reportan los estadísticos de **BLAST**.

```
>gi|71548896|ref|ZP_00669120.1|  Translation elongation factor G:Small GTP-binding protein domain
[Nitrosomonas eutropha C71]
 gi|71486077|gb|EAO18626.1|  Translation elongation factor G:Small GTP-binding protein domain
[Nitrosomonas eutropha C71]
Length=696

 Score =  828 bits (2140),  Expect = 0.0
 Identities = 434/697 (62%), Positives = 541/697 (77%), Gaps = 9/697 (1%)

Query  1    MTREFSLEKTRNIGIMAHIDAGKTTTTERVLYYTGRIHKIGETHEGASQMDWMAQEQERG  60
            M++   LE+ RNIGIMAHIDAGKTTT+ER+L+YTG  HK+GE H+GA+ MDWM QEQERG
Sbjct  1    MSKRNPLERYRNIGIMAHIDAGKTTTSERILFYTGVSHKLGEVHDGAATMDWMEQEQERG  60

Query  61   XXXXXXXXXXWN-------DHRINIIDTPGHVDFTVEVERSLRVLDGAVAVLDAQSGVE  113
            ITITSAATT  W        +HRIN+IDTPGHVDFT+EVERSLRVLDGA  V  +  GV+
Sbjct  61   ITITSAATTCFWKGMAGNYPEHRINVIDTPGHVDFTIEVERSLRVLDGACTVFCSVGGVQ  120
```

(... truncado)

Programación dinámica y la generación de alineamientos pareados (globales y locales)

- Pares de secuencias pueden ser comparadas usando **alineamientos globales y locales**, dependiendo del objetivo de la comparación.

Un **alineamiento global** fuerza el alineamiento de ambas secuencias a lo largo de toda su longitud. Usamos aln. globales cuando estamos seguros de que la homología se extiende a lo largo de todas las secuencias a comparar. Este es el tipo de alineamientos que generan programas de alineamiento múltiple tales como clustal, T-Coffee o muscle.

(a)

```
P00001     1 MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNK---GI  58
               D  KG+ +F     QC T  +  K+  GP L G+ GRK G A G++Y+  N N    G+
P00090     1 Q-DAAKGEAVF----KQCMTCHRADKNMVGPALGGVVGRKAGTAAGFTYSPLNHNSGEAGL  56

P00001    59 IWGEDTLMEYLENPKKYIP--------------GTKMIFVGIKKEEERADLIAYLKKATNE  105
             +W ++ ++ YL +P  Y+                TKM F   +   ++R D+ AYL  AT +
P00090    57 VWTQENIIAYLPDPNAYLKKFLTDKGQADKATGSTKMTF-KLANDQQRKDVAAYL--ATLK  114
```

Alineamiento global óptimo del citocromo C humano (105 resíduos, SWISS-PROT acc. P00001) y citocromo C2 de Rhodopseudomonas palustris (114 resíduos, SWISS-PROT acc. P00090).

La matriz de puntuación o ponderación ("scoring matrix) empleada fue **BLOSUM62**, con **costo de gaps afines** de –(11 + k). La puntuación del alineamiento global es de 131, usando el algoritmo de **Needleman-Wunsch**.

Programación dinámica y la generación de alineamientos pareados (globales y locales)

Un **alineamiento local** sólo busca los segmentos con la puntuación más alta.
Se usa por ejemplo en el escrutinio de bases de datos de secuencias debido a que la homología entre pares de secuencias frecuentemente existe sólo a nivel de ciertos dominios, pero no a lo largo de toda la secuencia (**estructura modular de proteínas; genes discontínuos introns-exones; barajado de exones ...**).
BLAST y FASTA buscan alineamientos locales con alta puntuacion (**HSPs** ó high-scoring pairs)

(b)

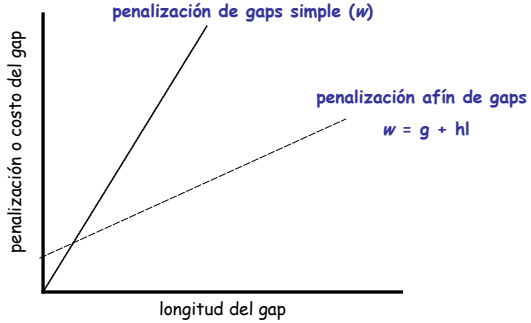
```
P13569  1221 EGGNAILENISFSISPGQRVGLLGRTGSGKSTLLSAFLRLL-----NTEGEIQIDGVS 1273
             +    ++  +S ++  G+ + L+G +GSGKS   +A L +L      T GEI  DG
P33593    13 QAAQPLVHGVSLTLQRGRVLALVGGSGSGKSLTCAATLGILPAGVRQTAGEILADGKP   70

P13569  1274 WDSITL---------QQWRKAFGVIPQKVFIFSGTFRKNLDPYEQWSDQEIWKVADEV 1322
               L           Q  R AF  +          +  + + +   + K AD+
P33593    71 VSPCALRGIKIATIMQNPRSAFNPL-----------HTMHTHARETCLALGKPADDA  116

P13569  1323 GLRSVIEQFP-GKLDFVLVDGGCVLSHGHKQLMCLARSVLSKAKILLLDEPSAHLDPV 1379
               L + IE         VL      +S G  Q M +A +VL ++  ++ DEP+  LD V
P33593   117 TLTAAIEAVGLENAARVLKLYPFEMSGGMLQRMMIAMAVLCESPFIIADEPTTDLDVV  174
```

Alineamiento local óptimo del regulador de conductancia transmembranal de fibrosis cística de humano (1480 resíduos, SWISS-PROT acc. P13569) y la proteína transportadora de Ni dependiente de ATP de E. coli (253 resíduos, SWISS-PROT acc. P33593).

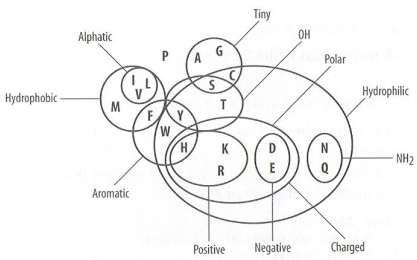
La matriz de puntuación o ponderación ("scoring matrix) empleada fue **BLOSUM62**, con **costo de gaps afines** de –(11 + k). La puntuación del alineamiento local es de 89, usando el algoritmo de **Smith-Waterman**.

alineamientos pareados y factores de penalización afines para gaps

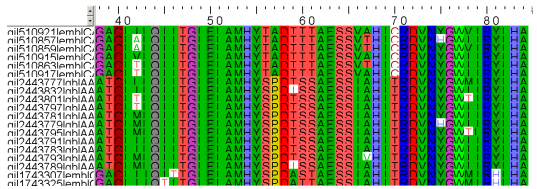
- Dado que **un sólo evento mutacional puede insertar o eliminar varios nucleótidos** de una secuencia, un indel largo no debe de ser penalizado mucho más que otro más corto ubicado en la misma región de un gen. De ahí el uso de **factores de penalización afines para gaps** (affine gap penalties or costs), que cobran una penalidad relativamente alta por abrir un gap y una penalidad más baja por cada posición sobre la que se extiende.
- La calidad de un alineamiento depende en gran medida de los **valores de apertura y extensión de gap** elegidos.



Similitud entre pares de secuencias de AA

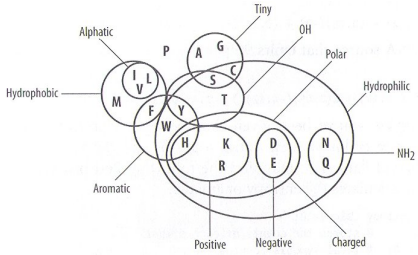


- Este diagrama muestra a los **aminoácidos** agrupados atendiendo a sus características químicas y físicas.
- Desde una perspectiva evolutiva **esperamos encontrar más sustituciones entre aas. similares que entre los menos relacionados.**
- Estos patrones puede observarse en alineamientos múltiples como el mostrado abajo

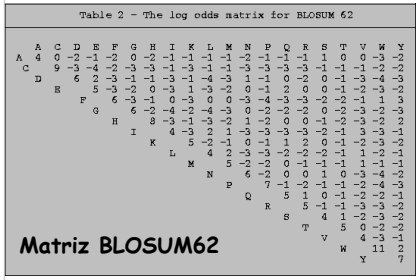


Segmento de un aln. múltiple de citocromos C de primates

Similitud entre pares de secuencias de AA



- **Las matrices empíricas de sustitución** entre AAs no reflejan necesariamente las relaciones químicas entre ellos. Se trata de una **definición púramente estadística basada en el análisis de frecuencias empíricas de sustituciones observadas en alineamientos de secs. con un grado de divergencia definido**



- Cada **score** de la matriz representa la tasa de sustitución esperada entre un par de AAs. Por tanto, los scores de los alineamientos pareados evaluados con estas matrices reflejan la distancia evolutiva existente entre las secuencias.

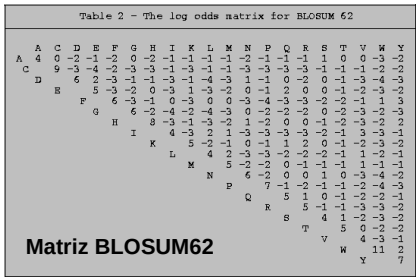
Es importante notar que los scores son evolutivamente simétricos al no conocerse la dirección del cambio evolutivo.

Similitud entre pares de secuencias de AA

- **Matrices de sustitución de AAs log-odds scores**

$$s(a,b) = (c) \log \frac{p_{ab}}{f_a f_b}$$

$s(a,b)$ = score del par a, b



p_{ab} = verosimilitud de la hipótesis a testar; **frecuencia esperada o diana**, probabilidad con la que esperamos encontrar *a* y *b* apareados en un alineamiento múltiple

$f_a f_b$ = verosimilitud de la hipótesis nula; **frecuencia de fondo**, probabilidad con la que esperamos encontrar *a* y *b* en cualquier proteína. Refleja su abundancia o frecuencia

c = Factor de escalamiento usado para multiplicar los lod scores (números reales) antes de ser redondeados a números enteros, tal y como se observa en la matriz. Los valores enteros redondeados resultantes se conocen como "raw scores".

Estadísticos de Karlin-Altschul de similitud entre secuencias: frecuencias diana, lambda y entropía relativa

Los atributos más importantes de una matriz de sustitución son sus **frecuencias esperadas o diana** implícitas para cada par de aa en sus respectivos scores crudos. Estas frecuencias esperadas **representan el modelo evolutivo subyacente**.
Los scores que han sido re-escalados y redondeados (scores representados en la matriz) son los **scores crudos** $s_{a,b}$. Para convertirlos a un **score normalizado** (log-odd score original) tenemos que mutiplicarlos por **λ**, una constante específica para cada matriz.
λ es aprox. igual al inverso del factor de escalamiento (c).

$$s(a,b) = \frac{1}{\lambda} \log \frac{p_{ab}}{f_a f_b} \qquad p_{ab} = f_a f_b \, e^{\lambda s_{ab}}$$ = **score normalizado**

por tanto, para despejar λ necesitamos $f_a f_b$ y encontrar el valor de λ para el que la suma de las frecuencias diana implícitas valga 1.

$$\sum_{a=1}^{n}\sum_{b=1}^{a} p_{ab} = \sum_{a=1}^{n}\sum_{b=1}^{a} f_a f_b \, e^{\lambda s_{ab}} = 1$$

Una vez calculada **λ**, se usa para calcular el **valor de expectación** (***E***) de cada **HSP** (**High Scoring Pair**) en el reporte de una búsqueda **BLAST**

Dado que las $f_a f_b$ de los resíduos de algunas proteínas difieren mucho de las frecuencias de resíduos empleadas para calcular las matrices PAM y BLOSUM, versiones recientes de BLASTP y PSI-BLAST incorporan una "**composition-based λ**" que es "**hit-específica**"

© Pablo Vinuesa 2021, @pvinmex
vinuesa[at]ccg[dot]unam[dot]mx;
http://www.ccg.unam.mx/~vinuesa/

Estadísticos de Karlin-Altschul para alineamientos locales

Karlin, S., and Altschul, S. F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A **87**: 2264-268.

$E = k\,m\,n\,e^{-\lambda S}$

Esta ecuación indica que el **número de alineamientos esperados por azar** (*E*) durante una búsqueda de similitud en una base de datos de secuencias está en función de:
el tamaño del espacio de búsqueda (*m*, *n*), el **score normalizado** (λ*S*) del HSP y una constante de valor pequeño (*k*)

E **Describe el ruido de fondo por azar presente en matches de dos secs.**

m = número de símbolos en la secuencia problema
n = número de símbolos en la base de datos
k ≈ 0.1 constante de ajuste para considerar HSPs altamente correlacionados

BLAST: Basic Local Alignment Search Tool

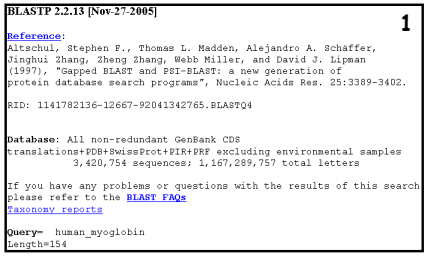
BLAST consta de una **familia de programas**. Los 5 ppales son:

BLASTN (nt-nt), **BLASTP** (p-p), **BLASTX** (translated nt-p),
TBLASTN (p-translated nt), usado en mapeo de prots contra DNA genómico
TBLASTX (translated nt - translated nt) usado en la predicción de genes

y **variantes de BLASTP** como **PSI-** y **PHI-BLAST**

BLAST: Basic Local Alignment Search Tool

· Anatomía de un reporte de NCBI-BLAST estándar



BLASTP 2.2.13 [Nov-27-2005] 1

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: 1141782136-12667-92041342765.BLASTQ4

Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples
3,420,754 sequences; 1,167,289,757 total letters

If you have any problems or questions with the results of this search please refer to the BLAST FAQs
Taxonomy reports

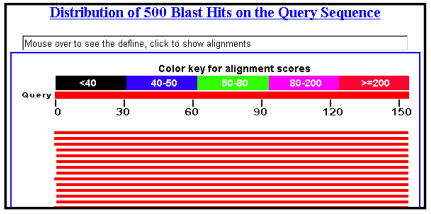
Query= human_myoglobin
Length=154

1.- **Encabezado**. Indica el programa de BLAST y su versión, con la fecha

Request ID

Indica la **BD** sobre la que se hizo la búsqueda, junto con el no. de secs contenida en ella y el no. de caracteres

Indica cual fue la query y su longitud



2.- **Resumen gráfico de distribución de hits con respecto a la query**.

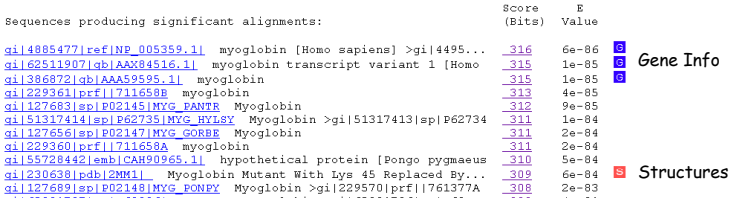
escala de color que indica el score de los HSPs

Las barras indican la distribución de los HSPs (coordenadas) con respecto a la secuencia problema (query), indicando en una escala de color el score de los alns. medidos en bits

BLAST: Basic Local Alignment Search Tool

· Anatomía de un reporte de NCBI-BLAST estándar

3. **Resúmenes de 1 linea**. Indican el nombre de la sec. junto con el score más alto y E value más bajo encontrado para un HSP o grupo de HSPs



Related Structures

Sequences producing significant alignments: Score (Bits) E Value

gi|4885477|ref|NP_005359.1| myoglobin [Homo sapiens] >gi|4495... 316 6e-86
gi|62511907|qb|AAX84516.1| myoglobin transcript variant 1 [Homo 315 1e-85
gi|386872|qb|AAA59595.1| myoglobin 315 1e-85
gi|229361|prf||711658B myoglobin 313 4e-85
gi|127683|sp|P02145|MYG_PANTR Myoglobin 312 9e-85
gi|51317414|sp|P62735|MYG_HYLSY Myoglobin >gi|51317413|sp|P62734 311 1e-84
gi|127656|sp|P02147|MYG_GORBE Myoglobin 311 2e-84
gi|229360|prf||711658A myoglobin 311 2e-84
gi|55728442|emb|CAH90965.1| hypothetical protein [Pongo pygmaeus 310 5e-84
gi|230638|pdb|2MM1| Myoglobin Mutant With Lys 45 Replaced By... 309 6e-84
gi|127689|sp|P02148|MYG_PONPY Myoglobin >gi|229570|prf||761377A 308 2e-83
gi|62901707|sp|P68086|MYG_ERYPA Myoglobin >gi|62901706|sp|P68... 300 4e-81

Gene Info

Structures

BLAST: Basic Local Alignment Search Tool

- **Anatomía de un reporte de NCBI-BLAST estándar**

4. **Alineamientos**. Representan la parte más voluminosa del reporte. Además de la información estadística, indica las coordenadas de inicio y fin de las secuencias query y subject. Si la búsqueda involucra secuencias de DNA, también se indica direccionalidad de las hebras Q/S (plus/plus; plus/minus).

```
>gi|47523546|ref|NP_999401.1|  myoglobin [Sus scrofa]
 gi|127688|sp|P02189|MYG_PIG  Myoglobin
 gi|164547|gb|AAA31073.1|  myoglobin
Length=154    normalized score
                      raw score
 Score =  296 bits (758),  Expect = 5e-80
 Identities = 144/154 (93%), Positives = 148/154 (96%), Gaps = 0/154 (0%)

Query  1    MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE  60
            MGLSDGEWQLVLNVWGKVEAD+ GHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE
Sbjct  1    MGLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE  60

Query  61   DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH  120
            DLKKHG TVLTALGGILKKKGHHEAE+ PLAQSHATKHKIPVKYLEFISE IIQVLQSKH
Sbjct  61   DLKKHGNTVLTALGGILKKKGHHEAELTPLAQSHATKHKIPVKYLEFISEAIIQVLQSKH  120

Query  121  PGDFGADAQGAMNKALELFRKDMASNYKELGFQG  154
            PGDFGADAQGAM+KALELFR DMA+ YKELGFQG
Sbjct  121  PGDFGADAQGAMSKALELFRNDMAAKYKELGFQG  154
```