Alineamiento local: búsqueda de homologías

Guillermo Barturen
gbarturen@ugr.es

¿Por qué necesitamos alineadores locales?

Generalmente en biología molecular lo que obtenemos en algunos experimentos no son genes o proteínas completos bien caracterizados, sino secuencias individuales y aisladas de su contexto como por ejemplo un oligonucleótido:

TACAGCAGATAGCAGCCATAGCCGCATACGTCGCGACTAC

O bien de un oligopéptido:

PTWRVPGRMEKWHALVKYLKYRTKDLEEVR

¿Cómo podemos obtener en este caso información de dicha secuencia?

Para responder a esto, necesitamos hacer un 'rastreo' de la base de datos. El alineamiento global de dos secuencias (**Needleman-Wunsch**) es muy preciso y garantiza obtener el alineamiento óptimo. Sin embargo, es muy lento. Siendo el **tiempo de cálculo proporcional al producto de la longitud de las secuencias a alinear** o en este caso a la longitud de nuestra secuencia problema y de la base de datos al completo.

Por el contrario, los algoritmos de **alineamiento local** son mucho más rápidos.

Alineamiento local de secuencias

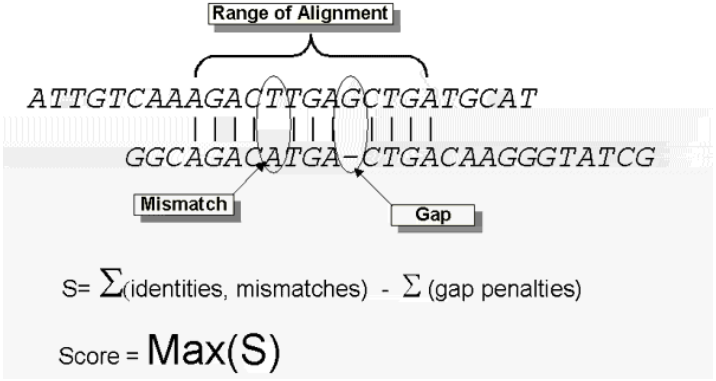
Se localizan todas las subsecuencias similares entre las dos secuencias:

```
Query: 181 acgatagcagatagcgcatagcgactagcgactgcagctacgcagcatagcagcagcaga 240
                     ||||||               |||          |||||
Sbjct: 189 tgagctagagatagctacgacgcatcagcgatagcagctaggcagctgcagcgactagca 247
```

El alineamiento se trata de extender en los dos sentidos mediante alineamiento global:

```
Query: 181 acgatagcagatagcgcatagcgactagcgactgcagctacgcagcatagcagcagcaga 240
                     ||||||               |||          |||||
Sbjct: 189 tgagctagagatagctacgacgcatcagcgatagcagctaggcagctgcagcgactagca 247
```

Y se calcula una puntuación de alineamiento para las posibles combinaciones, eligiéndose la máxima:




```
AACGTTTCCAGTCCAAATAGCTAGGC
|||**|||   |*|||*||*||||||
AACCGTTC---TACAATTACCTAGGC
```

| Emparejamientos (+1): 18
* Desemparejamientos (-2): 5
- Huecos (existencia-2, extension -1): 1 de longitud 3

Puntuación = [18 * 1] + [5 * (-2)] + [(– 2) + 2*(-1)] = 4

Diferencias alineamiento global y local

Global Sequence Alignment	Local Sequence Alignment
In global alignment, an attempt is made to align the entire sequence (end to end alignment)	Finds local regions with the highest level of similarity between the two sequences.
A global alignment contains all letters from both the query and target sequences	A local alignment aligns a substring of the query sequence to a substring of the target sequence.
If two sequences have approximately the same length and are quite similar, they are suitable for global alignment.	Any two sequences can be locally aligned as local alignment finds stretches of sequences with high level of matches without considering the alignment of rest of the sequence regions.
Suitable for aligning two closely related sequences.	Suitable for aligning more divergent sequences or distantly related sequences.
Global alignments are usually done for comparing homologous genes like comparing two genes with same function (in human vs. mouse) or comparing two proteins with similar function.	Used for finding out conserved patterns in DNA sequences or conserved domains or motifs in two proteins.
A general global alignment technique is the Needleman–Wunsch algorithm.	A general local alignment method is Smith–Waterman algorithm.
Examples of Global alignment tools: • EMBOSS Needle • Needleman-Wunsch Global Align Nucleotide Sequences (Specialized BLAST)	**Examples of Local alignment tools:** • BLAST • EMBOSS Water • LALIGN

Significación estadística de un alineamiento: Test de randomización

Las **puntuaciones** de los alineamientos son útiles para determinar cuál de ellos es el **más probable** entre todos los posibles. Sin embargo, el valor numérico resultante **no nos informa de hasta qué punto las similitudes de ese alineamiento pueden producirse por azar**, para ello debe realizarse un test de aleatorización.

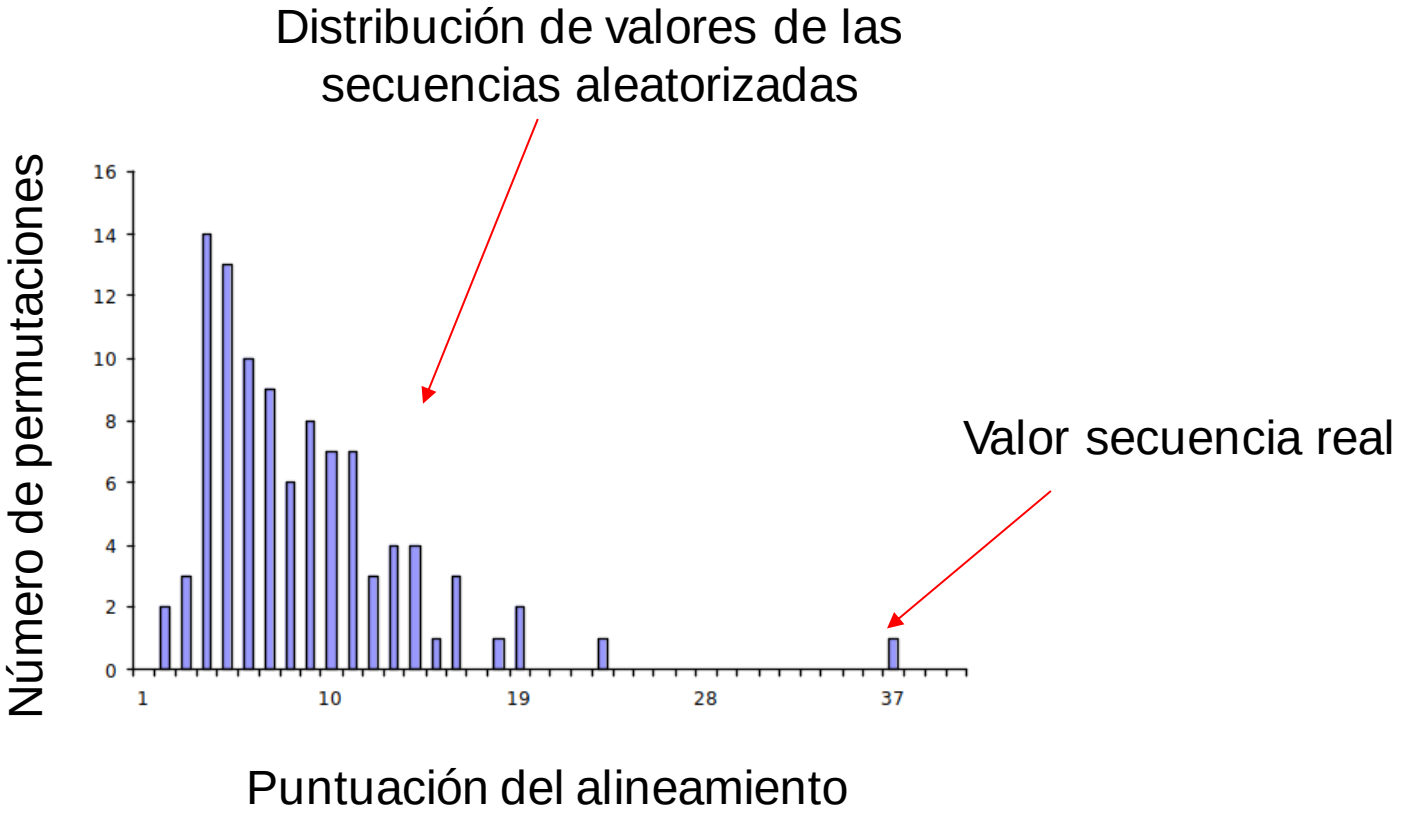
1. Se alinean las dos proteínas y se obtiene una puntuación real para el alineamiento obtenido:

```
RBP:          26  RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWD- 84
                  + K++ +  + +GTW++MA      + L   + A   V  T  +        +L+  W+
glycodelin:   23  QTKQDLELPKLAGTWHSMAMA-TNNISLMATLKAPLRVHITSLLPTPEDNLEIVLHRWEN 81
```

2. Se aleatoriza la segunda secuencia n veces, permutando al azar (**‘shuffling’**) las posiciones que ocupan los aminoácidos (manteniendo por tanto la longitud de la secuencia y la composición de aminoácidos)
3. Se alinea cada secuencia aleatorizada con la primera secuencia y se obtienen n puntuaciones aleatorias.
4. Cabe esperar que la puntuación real sea mucho más grande que las puntuaciones aleatorias.

Significación estadística de un alineamiento: Test de randomización

```
RBP:          26  RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWD- 84
                  + K++ +  + +GTW++MA      + L   + A   V  T  +        +L+  W+
glycodelin:   23  QTKQDLELPKLAGTWHSMAMA-TNNISLMATLKAPLRVHITSLLPTPEDNLEIVLHRWEN 81
```



En este caso, el test de aleatorización muestra que el oligonucleótido RBP muestra una similitud de secuencia alta no debida al azar

Alineamiento local: FASTA (Fast Algorithm, Pearson & Lipman, 1988)

https://www.ebi.ac.uk/Tools/sss/fasta/nucleotide.html

https://www.ebi.ac.uk/Tools/sss/fasta/

Nucleotide Similarity Search

This tool provides sequence similarity searching against nucleotide databases using the FASTA suite of programs. FASTA provides a heuristic search with a nucleotide query. TFASTX and TFASTY translate the DNA database for searching with a protein query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

STEP 1 - Select your databases

NUCLEOTIDE DATABASES

111 Databases Selected — ✗ Clear Selection

- **ENA Sequences (formerly EMBL-Bank)**
  - ☑ ENA Sequences
  - ☐ ENA Coding Sequences
  - ☐ ENA Non-coding Sequences
  - ☐ ENA Ribosomal Sequences
  - ☐ ENA Geospatial Sequences
  - ☐ ENA Spacer Sequences
  - **Others**
- **IMGT**
- **Patents**
- **Structure**
- **COVID-19 Data Portal**
- **Vectors**

Protein Similarity Search

This tool provides sequence similarity searching against protein databases using the FASTA suite of programs. FASTA provides a heuristic search with a protein query. FASTX and FASTY translate a DNA query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

STEP 1 - Select your databases

PROTEIN DATABASES

1 Database Selected — ✗ Clear Selection

- ☐ UniProt Knowledgebase (The UniProt Knowledgebase includes UniProtKB/Swiss-Prot and UniProtKB/TrEMBL)
- ☑ UniProtKB/Swiss-Prot (The manually annotated section of UniProtKB)
- ☐ UniProtKB/Swiss-Prot isoforms (The manually annotated isoforms of UniProtKB/Swiss-Prot)
- ☐ UniProtKB/TrEMBL (The automatically annotated section of UniProtKB)
- ☐ UniProtKB Reference Proteomes plus Swiss-Prot
- ☐ UniProtKB COVID-19
- **UniProtKB Taxonomic Subsets**
- **UniProt Clusters**
- **Patents**
- **Structures**
- **Other Protein Databases**

STEP 3 - Set your parameters

PROGRAM: FASTA

MATCH/MISMATCH SCORES	GAP OPEN	GAP EXTEND	KTUP	EXPECTATION UPPER	EXPECTATION LOWER
+5/-4	-14	-4	6	10	0 (default)

FASTA

MATRIX	GAP OPEN	GAP EXTEND	KTUP	EXPECTATION	EXPECTATION
BLOSUM50	-10	-2	2	10	0 (default)

DNA STRAND	HISTOGRAM	FILTER	STATISTICAL ESTIMATES
N/A	no	none	Regress

SCORES	ALIGNMENTS	SEQUENCE RANGE	DATABASE RANGE	MULTI HSPs
50	50	START-END	START-END	no

SCORE FORMAT	ANNOTATION FEATURES
Default	no

Alineamiento local: FASTA

Implementaciones del algoritmo FASTA

For Protein FASTA:

Program Name	Description	Abbreviation
FASTA	Scan a protein or DNA sequence library for similar sequences.	fasta
FASTX	Compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.	fastx
FASTY	Compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.	fasty
SSEARCH	Compare a protein or DNA sequence to a sequence database using the Smith-Waterman algorithm.	ssearch
GGSEARCH	Compare a protein or DNA sequence to a sequence database using a global alignment (Needleman-Wunsch)	ggsearch
GLSEARCH	Compare a protein or DNA sequence to a sequence database with alignments that are global in the query and local in the database sequence (global-local).	glsearch

For Nucleotide/Genome/Whole Genome Shotgun

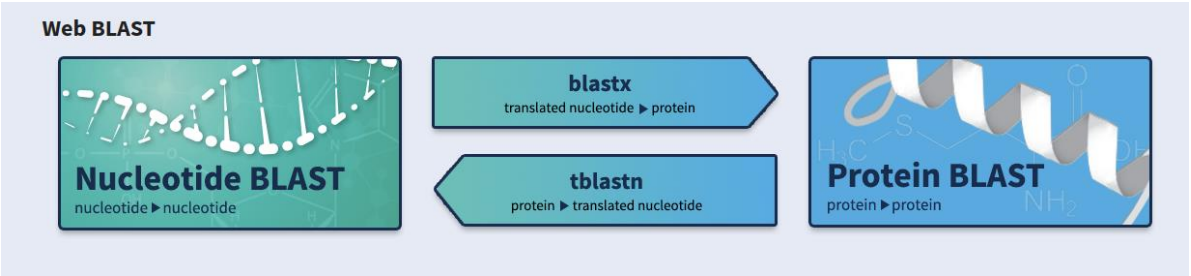
Program Name	Description	Abbreviation
FASTA	Scan a protein or DNA sequence library for similar sequences.	fasta
SSEARCH	Compare a protein or DNA sequence to a sequence database using the Smith-Waterman algorithm.	ssearch
GGSEARCH	Compare a protein or DNA sequence to a sequence database using a global alignment (Needleman-Wunsch)	ggsearch
GLSEARCH	Compare a protein or DNA sequence to a sequence database with alignments that are global in the query and local in the database sequence (global-local).	glsearch
TFASTX	Compare a protein sequence to a DNA sequence database, calculating similarities with frameshifts to the forward and reverse orientations.	tfastx
TFASTY	Compare a protein sequence to a DNA sequence database, calculating similarities with frameshifts to the forward and reverse orientations.	tfasty

Align.	DB:ID	Source	Length	Score (Bits)	Identities %	Positives %	E()
1	SP:P18097	Virion infectivity factor OS=Human immunodeficiency virus type 2 subtype A (isolate BEN) OX=11714 GN=vif PE=2 SV=1 *Cross-references and related information in:* ▶ Bioactive molecules ▶ Nucleotide sequences ▶ Literature ▶ Samples & ontologies ▶ Protein families ▶ Protein sequences	215	310.0	100.0	100.0	1.0E-83
2	SP:P17758	Virion infectivity factor OS=Human immunodeficiency virus type 2 subtype A (isolate D194) OX=11713 GN=vif PE=2 SV=1 *Cross-references and related information in:* ▶ Bioactive molecules ▶ Nucleotide sequences ▶ Literature ▶ Samples & ontologies ▶ Protein families ▶ Protein sequences	215	295.4	93.9	98.2	2.6E-79
3	SP:P04595	Virion infectivity factor OS=Human immunodeficiency virus type 2 subtype A (isolate ROD) OX=11720 GN=vif PE=2 SV=1 *Cross-references and related information in:* ▶ Bioactive molecules ▶ Nucleotide sequences ▶ Literature ▶ Samples & ontologies ▶ Protein families ▶ Protein sequences	215	288.9	90.2	98.2	2.2E-77
4	SP:P18043	Virion infectivity factor OS=Human immunodeficiency virus type 2 subtype A (isolate Ghana-1) OX=11717 GN=vif PE=2 SV=1 *Cross-references and related information in:* ▶ Bioactive molecules ▶ Nucleotide sequences ▶ Literature ▶ Samples & ontologies ▶ Protein families ▶ Protein sequences	215	286.1	91.5	97.6	1.6E-76
5	SP:P20878	Virion infectivity factor OS=Human immunodeficiency virus type 2 subtype A (isolate ST) OX=11721 GN=vif PE=2 SV=1 *Cross-references and related information in:* ▶ Bioactive molecules ▶ Nucleotide sequences ▶ Literature ▶ Samples & ontologies ▶ Protein families ▶ Protein sequences	215	282.3	89.0	95.7	2.3E-75
6	SP:P12452	Virion infectivity factor OS=Human immunodeficiency virus type 2 subtype A (isolate SBLISY) OX=11718 GN=vif PE=2 SV=1 *Cross-references and related information in:* ▶ Bioactive molecules ▶ Nucleotide sequences ▶ Literature ▶ Samples & ontologies ▶ Protein families ▶ Protein sequences	215	281.8	86.5	96.9	3.3E-75

Valor P: Probabilidad de que un suceso ocurra por azar.

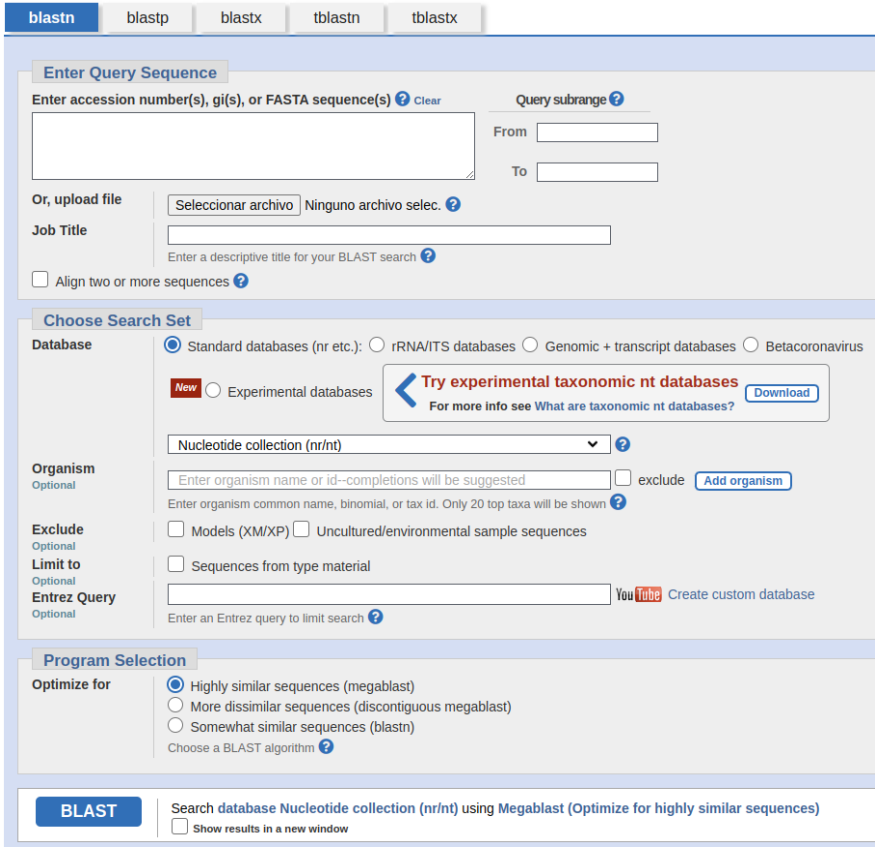
Valor E (expectation value): Corrección del valor P para ensayos múltiples. Cuanto más bajo el valor E, más significativa es la puntuación obtenida para un alineamiento.

Alineamiento local: BLAST (Basic Local Alignment Search Tool, Altschul, S.F. et al. 1990)



Selección de algoritmos

Parámetros optimizados



Bases de datos

Alineamiento local: BLAST (Basic Local Alignment Search Tool, Altschul, S.F. et al. 1990)

Puntuación

Cobertura de la secuencia problema sobre la base de datos

Valor E

Porcentaje se similitud de secuencia

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments — Download — Select columns — Show 100

select all *100 sequences selected* — GenBank — Graphics — Distance tree of results — MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Human beta-actin related pseudogene H-beta-Ac-psi-2, (Probably derived from a backtranslated mRNA), I...	Homo sapiens	3003	3003	100%	0.0	100.00%	1665	V00481.1
Human DNA sequence from clone RP11-459O1 on chromosome 6, complete sequence	Homo sapiens	2975	2975	100%	0.0	99.70%	107745	AL139042.15
Homo sapiens ACTB pseudogene 8 (ACTBP8) on chromosome 6	Homo sapiens	2546	2546	84%	0.0	99.93%	1969	NG_000840.7
Macaca mulatta chromosome UNK clone CH250-53H12, complete sequence	Macaca mulatta	1911	1911	100%	0.0	85.65%	162603	AC188451.1
PREDICTED: Hylobates moloch actin beta (ACTB), mRNA	Hylobates moloch	1728	1728	78%	0.0	88.55%	1863	XM_032755826.2
PREDICTED: Aotus nancymaae actin beta (ACTB), mRNA	Aotus nancymaae	1725	1725	78%	0.0	88.52%	1911	XM_012455575.2
PREDICTED: Piliocolobus tephrosceles actin beta (ACTB), mRNA	Piliocolobus tep...	1722	1722	78%	0.0	88.41%	1912	XM_023197872.2
PREDICTED: Callithrix jacchus actin beta (ACTB), mRNA	Callithrix jacchus	1716	1716	78%	0.0	88.38%	1854	XM_035292064.2
PREDICTED: Callithrix jacchus actin, cytoplasmic 1 (LOC100412618), mRNA	Callithrix jacchus	1709	1709	78%	0.0	88.34%	1641	XM_002751780.6
PREDICTED: Nomascus leucogenys actin beta (ACTB), mRNA	Nomascus leuc...	1708	1708	78%	0.0	88.17%	1923	XM_030796883.1
PREDICTED: Symphalangus syndactylus actin beta (ACTB), transcript variant X1, mRNA	Symphalangus ...	1708	1708	78%	0.0	88.17%	1865	
PREDICTED: Piliocolobus tephrosceles actin, cytoplasmic 1-like (LOC111546247), mRNA	Piliocolobus tep...	1704	1704	78%	0.0	88.17%	1809	
Pongo abelii mRNA: cDNA DKFZp459A127 (from clone DKFZp459A127)	Pongo abelii	1703	1703	78%	0.0	88.20%	1833	
PREDICTED: Pongo pygmaeus actin beta (ACTB), mRNA	Pongo pygmaeus	1701	1701	78%	0.0	88.21%	1825	
Pan troglodytes actin beta (ACTB), mRNA	Pan troglodytes	1700	1700	78%	0.0	87.91%	1850	
PREDICTED: Chlorocebus sabaeus ac						88.04%	2076	
Homo sapiens actin beta (ACTB), mRN						87.83%	1812	
Homo sapiens cDNA FLJ25290 fis, clo						87.83%	1804	

Range 1: 2482 to 4142 GenBank Graphics — Next Match — Previous Ma

Score	Expect	Identities	Gaps	Strand
2975 bits(3299)	0.0	1660/1665(99%)	4/1665(0%)	Plus/Minus

```
Query  1     CTACAGTGAGCCGAGGTCATGCCATTGCACTCCAATCTGGGCGACAAGAGTGAAACTCCG  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4142  CTACAGTGAGCCGAGGTCATGCCATTGCACTCCAATCTGGGCGACAAGAGTGAAACTCCG  4083

Query  61    TCaaaagaaagaaagaaagagacaaagagagttagaaagaaagaaagagagagagagaga  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4082  TCAAAAGAAAGAAAGAAAGAGACAAAGAGAGTTAGAAAGAAAGAAAGAGAGAGAGAGA  4023

Query  121   aaggaaggaaggaaggaaaaaagaaagaaaaagaaagaaaagagaaagaaagaaagagaaaga  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4022  AAGGAAGGAAGGAAGAAAAAGAAAGAAAAAGAAAGAAAGAGAAAGAAAGAAAGAGAAAGA  3963

Query  181   aagaaagaaagaaagaaagaaagaaagaaagaaagaaagaaaaagaaagaaagaaagaaaa  240
             ||||||||||||||||||||||||||||||||||||||||  |||||  |||||||||||
Sbjct  3962  AAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAA--AAGAAA--GAAAGAAAGAAAGAAA  3907

Query  241   gaaagaaagaaagaaagaaagaaagaaagaaagaaagaaagaaaggaaggaaagaaagaaGCAAG  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||  |||
Sbjct  3906  GAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGGAAGGAAAGAAAGAGCAAG  3847

Query  301   TTACTATAGCGGTAGGGGAGATGTTGTAGAAATATATATAAACCTCCTTACACCGCGGAG  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3846  TTACTATAGCGGTAGGGGAGATGTTGTAGAAATATATATAAACCTCCTTACACCGCGGAG  3787

Query  361   ACCGCGTCAGCCCAGCGAGCACAGAACCTTGTCCTTGCCGCTGCGCCTTGCGTCCGCACC  420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3786  ACCGCGTCAGCCCAGCGAGCACAGAACCTTGTCCTTGCCGCTGCGCCTTGCGTCCGCACC  3727

Query  421   CGCCGCCAGCTCACCATGGATGATGCTATCACCGCGCTCGTCGTCGTCGACAACTGCTCC  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3726  CGCCGCCAGCTCACCATGGATGATGCTATCACCGCGCTCGTCGTCGTCGACAACTGCTCC  3667

Query  481   AGCATGCGCAAGGCTCCCCAGGCCGTCTTCCCCTCCATTGTGGGGCACCCTAGGCACCAG  540
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3666  AGCATGCGCAAGGCTCCCCAGGCCGTCTTCCCCTCCATTGTGGGGCACCCTAGGCACCAG  3607
```