T.C. Fırat Üniversitesi
Bilgisayar Mühendisliği Bölümü

BMÜ450 - Biyoinformatik Dersi
BLAST İle Dizi Hizalama Ödevi Raporu

Ders Sorumlusu: Prof. Dr. Mehmet KAYA

Öğrenci İsim: Mert İNCİDELEN
Öğrenci No.: 170260101

ÖDEV TANIMI

Ödevde, BLAST Algoritmasının araştırılması ve çalışma prensibinin anlaşılması; BLAST kullanan bir araç üzerinde örnek dizilerin hizalanıp, nasıl yapıldığının gösterilmesi beklenmektedir.

1. BLAST ALGORİTMASI

BLAST (Basic Local Alignment Search Tool), biyoinformatik çalışmalarında yaygın kullanılan veri tabanı arama ve hizalama aracıdır. Dizi veri tabanlarını aramak için birçok algoritma mevcuttur, ancak BLAST Algoritmaları hızı itibariyle diğerlerine kıyasla en popüler olanlarıdır. BLAST Algoritmasının temel stratejisi amino asitlerden oluşan sorgu (BLASTP) üzerinden örneklenmiştir.

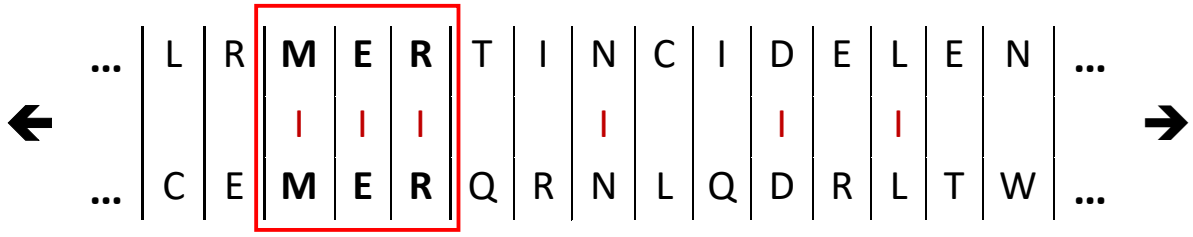
"… **M E R T I N C I D E L E N** …" sorgusu, boyutu 3 olan (BLASTP için varsayılan boyut 3 amino asittir) kelimelere ayrılır. Böylece MER, ERT, RTI, TIN, INC, NCI, CID, IDE, DEL, ELE, LEN kelimeleri elde edilir. Bir sorgu için elde edilen kelime n-2 adet olmaktadır. Proteinler 20 farklı amino asit içerdiğinden, bir kelimenin her 20^3 tripeptitte bir kez (8000 tripeptit, 24000 amino asit içinde) rastgele oluşması beklenir, bu da herhangi bir proteinden daha uzundur ve dolayısıyla idealdir.

Sorgu:	…	M	E	R	T	I	N	C	I	D	E	L	E	N	…
Kelimeler:	…	M	E	R										
			E	R	T									
				R	T	I								
					T	I	N							
						I	N	C						
							N	C	I					
								C	I	D				
									I	D	E			
										D	E	L		
											E	L	E	
												L	E	N	…

Elde edilen boyutu 3 olan bu kelimelerin komşu kelimeleri bulunur ve bir eşik değeri belirlenir. Bir skor tablosu kullanarak eşik değerinin üzerinde skora sahip komşu kelimeler de sorgulamaya dahil edilir. Sonraki aşamada "MER" kelimesi için komşu kelimeler, eşik değeri 11 belirlenerek BLOSUM 62 tablosuna göre bulunmuştur. Eşik değeri düşürülerek arama kapsamı genişletilebilir.

KELİME	SKOR (BLOSUM62)
MER	**15**
MEK	12
LER	12
MDR	12
MQR	12
MEQ	11
IER	11
VER	11
MKR	11
~~MEN~~	~~10~~
~~QER~~	~~10~~
. . .	
Eşik değeri = 11	

Daha sonra hedef dizilerde aranan kelimeler bulunarak, isabetin olduğu bölgede her iki yönde eşleştirmeler genişletilir.



Genişletilen bölgenin HSP (high-scoring segment pair) puanı, skor tablosu üzerinden hesaplanır ve en iyi eşleşmeler bu şekilde saptanabilir. Yukarıdaki bölge için HSP puanı BLOSUM62 matrisine göre hesaplanmıştır.

EŞ.	SKOR
L-C	-1
R-E	0
M-M	5

EŞ.	SKOR
E-E	5
R-R	5
T-Q	-1

EŞ.	SKOR
I-R	-3
N-N	6
C-L	-1

EŞ.	SKOR
I-Q	-3
D-D	6
E-R	0

EŞ.	SKOR
L-L	4
E-T	-1
N-W	-4

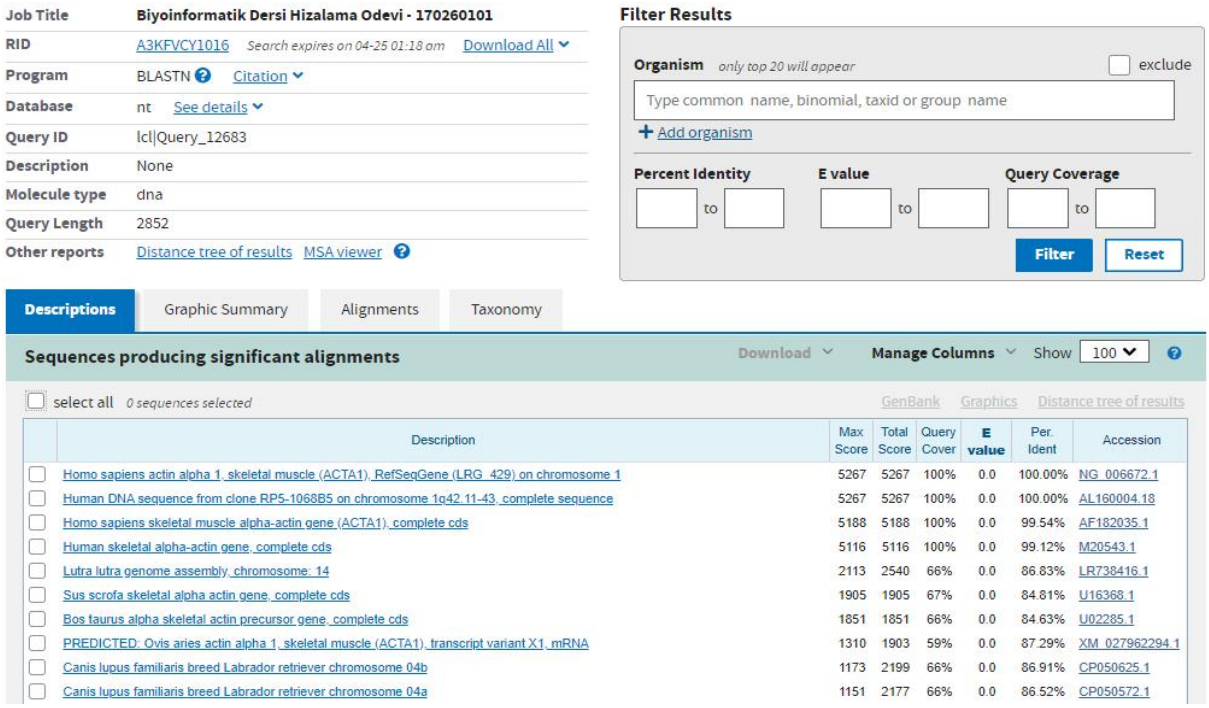
Pozitif skorlu eşleşmelerin değerlerini toplarsak (5+5+5+6+6+4 = 31), HSP puanı 31 olarak bulunur. Negatif skorlu eşleşmeler de dahil edilerek toplam puan 17 olarak bulunabilir.

2. BLAST İle Dizi Hizalama Uygulamaları

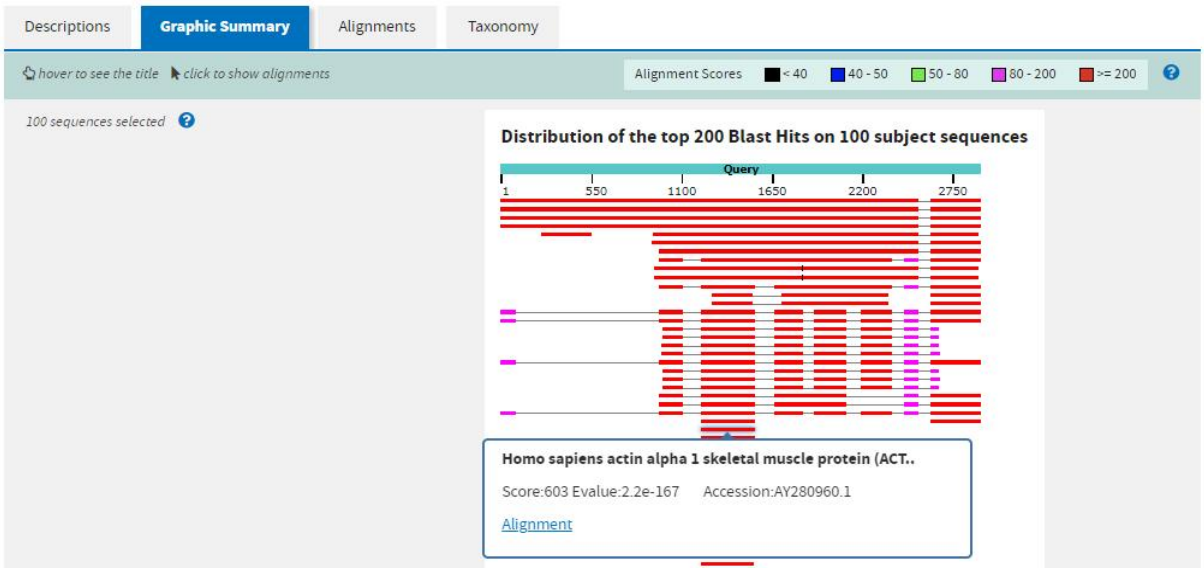
2.1. BLASTn İle Uygulama

Dizi hizalama işlemi için **blast.ncbi.nlm.nih.gov/Blast.cgi** adresinde bulunan araçtan faydalanılmıştır. Eşleştirme için ikinci örnek uygulamada dizi olarak **ACTA1** kimliğine sahip (ID No. 58) 2852 nükleotit uzunluğunda bir dizi kullanılmıştır. BLASTN (nükleotit-nükleotit karşılaştırma ve eşleştirme) ile ACTA1 dizisi standart veri tabanı üzerinde benzerlik derecesi yüksek olacak şekilde sorgulanmış ve aşağıdaki veriler elde edilmiştir.

Sorgulama sonrasında karşılaştırılan dizilerden benzerlik oranı yüksek olandan düşüğe doğru sonuç ekranında aşağıda görüldüğü şekilde sıralanmıştır.



Bu diziler bir grafik üzerinde hizalanma konumları ile kabaca gösterilmiştir. Kırmızı renk ile gösterilenler eşleşme skoru 200 ve üzerinde olan hizalamalardır.



Eşleşen dizilerden bir tanesi için eşleşen kısımların bir bölümünün sonuç dökümü ise aşağıdaki gibi gösterilmiştir.

Lutra lutra genome assembly, chromosome: 14

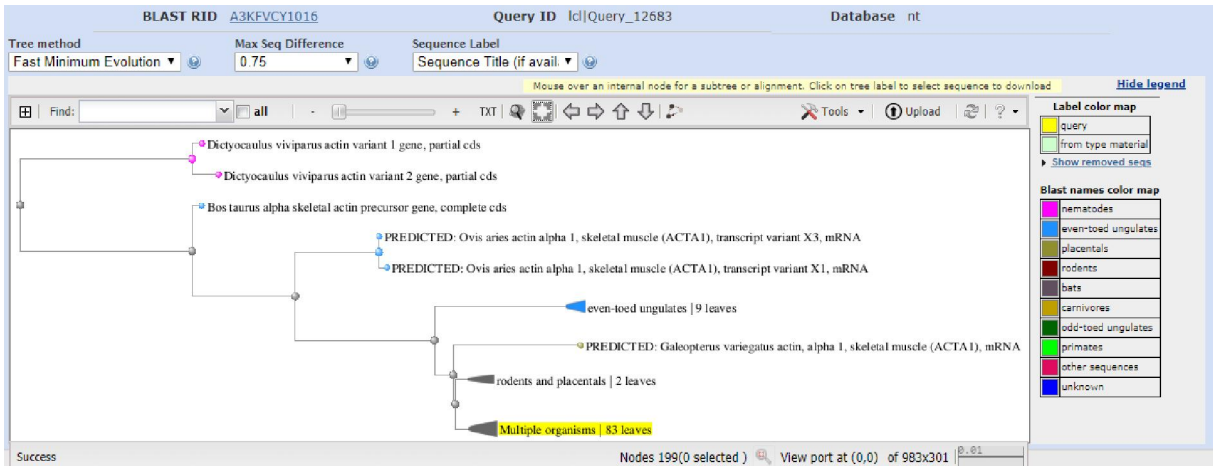
Sequence ID: LR738416.1 Length: 89080780 Number of Matches: 4

Range 1: 1388448 to 1390078 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
1879 bits(1017)	0.0	1452/1655(88%)	57/1655(3%)	Plus/Minus

```
Query  924      CTGCGCTGATGCACCGCGCCTCTTCGCGGTCTCCCTGTCCTT-GCAGAAACTAGACACAA  982
                |||  |||||  ||  |||||   |||  |  |||  ||  |||  |||||  |||||||||  ||||||  |
Sbjct  1390078  CTGTGCTGACGCGCCGCGTGTCTCCACGGCCTTCCTATCCTTTGCAGAAACCAGACACCA  1390019

Query  983      TGTGCGACGAAGACGAGACCACCGCCCTCGTGTGCGACAATGGCTCCGGCCTGGTGAAAG  1042
                ||||  ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1390018  TGTGTGACGAAGACGAGACCACCGCCCTCGTGTGCGACAATGGCTCCGGCCTGGTGAAAG  1389959

Query  1043     CCGGCTTCGCCGGGGATGACGCCCCTAGGGCCGTGTTCCCGTCCATCGTGGGCCGCCCCC  1102
                |||||||||||||  ||  |||||||||||||||||||||||||  ||||||||||||||  |
Sbjct  1389958  CCGGCTTCGCCGGCGACGACGCCCCTAGGGCCGTGTTCCCTTCCATCGTGGGCCGCCCTC  1389899

Query  1103     GACACCAGGTCAGGCTGCCCCTCCGCAGAGGGAGCCGGCTCGGGGTCC-CCGCGTAAGCC  1161
                |  ||||||||||||||||  |||  ||  ||||||||||  |||||  ||  |||  |  ||||
Sbjct  1389898  GCCACCAGGTCAGGCTGCCGCTCGGCGGAGGGAGCCGGGCCGGGGACCTCCGTG-GAGCC  1389840

Query  1162     -AGCCTGGTGCCACCCGGAGCGGCGTTAACGGGTGCGTGGTGTCTCGGCTCTGCAGGGCG  1220
                 |||  |||||||  |  ||||  |||  |||||  |  ||||||||||||  ||||  |||||||||
Sbjct  1389839  GGGCCCTGTGCCACTCTGAGCCGCGGTAACGTGGGCGTGGTGTCTCCGCTCCGCAGGGCG  1389780

Query  1221     TCATGGTCGGTATGGGTCAGAAAGATTCCTACGTGGGCGACGAGGCTCAGAGCAAGAGAG  1280
                |||||||  ||||||||||||||||||||||  ||||||||||||||||||||||||||||
Sbjct  1389779  TCATGGTGGGTATGGGTCAGAAAGATTCCTATGTGGGCGACGAGGCTCAGAGCAAGAGAG  1389720

Query  1281     GTATCCTGACCCTGAAGTACCCTATCGAGCACGGCATCATCACCAACTGGGATGACATGG  1340
                |  ||||||||||||||||||  |||||||||||||||||||||||||||||||  |||||||
Sbjct  1389719  GCATCCTGACCCTGAAGTACCCCATCGAGCACGGCATCATCACCAACTGGGACGACATGG  1389660

Query  1341     AGAAGATCTGGCACCACACCTTCTACAACGAGCTTCGCGTGGCTCCCGAGGAGCACCCCA  1400
                ||||||||||||||||||||||||||||||||||||  ||||||||  ||  ||||||||||||
Sbjct  1389659  AGAAGATCTGGCACCACACCTTCTACAACGAGCTCCGCGTGGCCCCTGAGGAGCACCCCA  1389600

Query  1401     CCCTGCTCACCGAGGCCCCCCTCAATCCCAAGGCCAACCGCGAGAAGATGACCCAGATCA  1460
                ||||||||||  ||||||||||||||||  |||||  ||||||||||||||||||||||||||||
Sbjct  1389599  CCCTGCTCACGGAGGCCCCCCTCAACCCCAAAGCCAACCGCGAGAAGATGACCCAGATCA  1389540

Query  1461     TGTTTGAGACCTTCAACGTGCCCGCCATGTACGTGGCCATCCAGGCCGTGCTGTCCCTCT  1520
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1389539  TGTTTGAGACCTTCAACGTGCCCGCCATGTACGTGGCCATCCAGGCCGTGCTGTCCCTCT  1389480
```

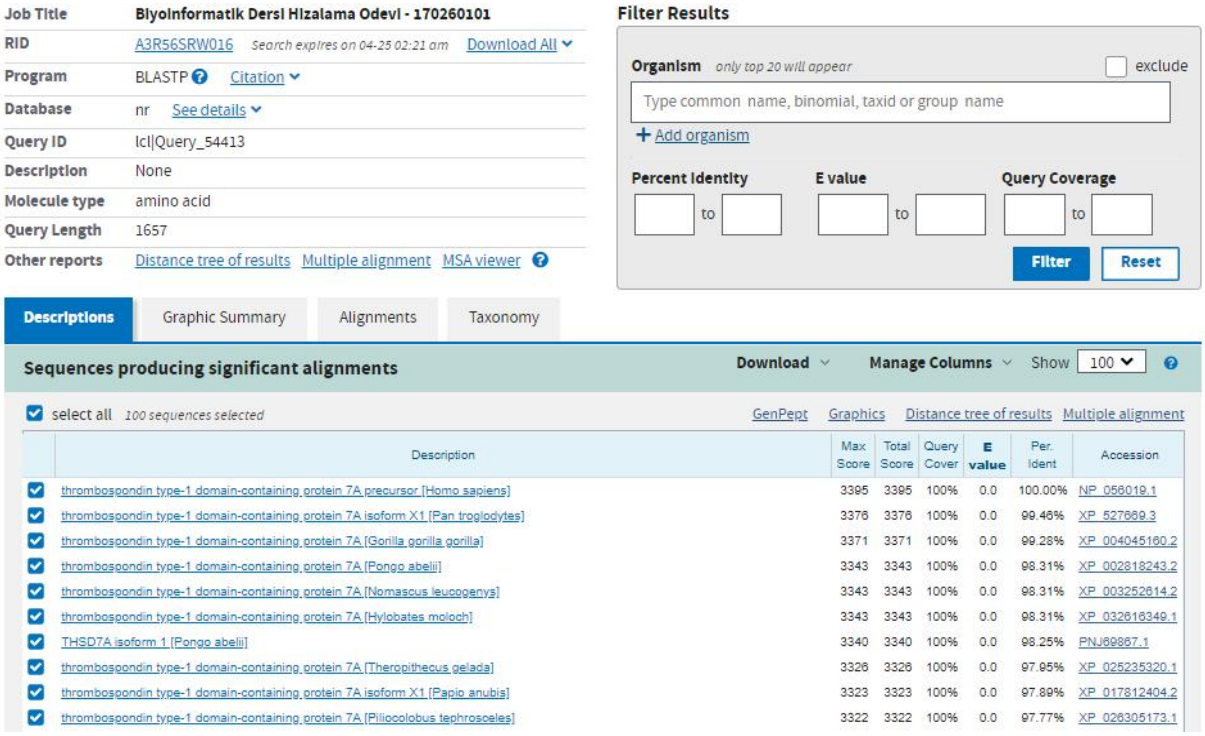
Benzerlikleri esas alınarak sorgu sonucunda elde edilen verilere göre dizilerin filogenetik haritası ise aşağıdaki gibidir.



2.2. BLASTp İle Uygulama

Eşleştirme için ikinci örnek uygulamada dizi olarak **Thrombospondin Type-1** kimliğine sahip 1657 aminoasit içeren bir dizi kullanılmıştır. BLASTp (aminoasid-aminoasid karşılaştırma ve eşleştirme) ile Thrombospondin Type-1 dizisi veri tabanı üzerinde sorgulanmış ve aşağıdaki veriler elde edilmiştir.

Sorgulama sonrasında karşılaştırılan dizilerden benzerlik oranı yüksek olandan düşüğe doğru sonuç ekranında aşağıda görüldüğü şekilde sıralanmıştır.



Job Title	Biyoinformatik Dersi Hizalama Odevi - 170260101
RID	A3R56SRW016 Search expires on 04-25 02:21 am Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_54413
Description	None
Molecule type	amino acid
Query Length	1657
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism only top 20 will appear exclude

Type common name, binomial, taxid or group name

Add organism

Percent Identity to | E value to | Query Coverage to

Filter Reset

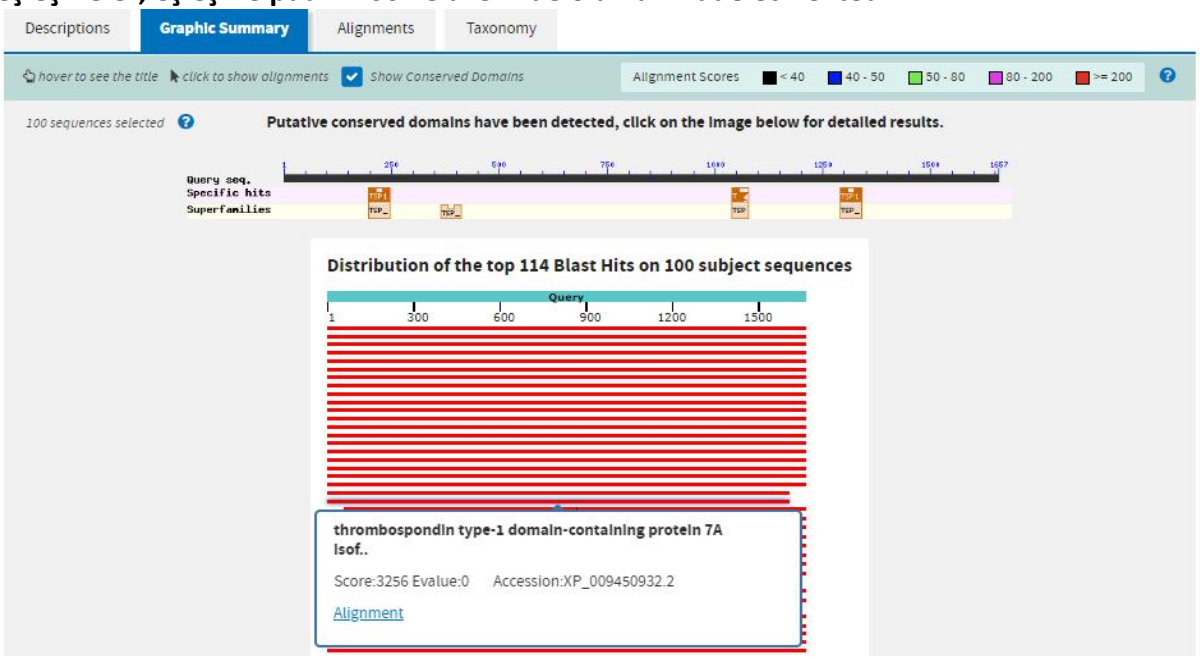
Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments — Download — Manage Columns — Show 100

select all 100 sequences selected — GenPept Graphics Distance tree of results Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
thrombospondin type-1 domain-containing protein 7A precursor [Homo sapiens]	3395	3395	100%	0.0	100.00%	NP_056019.1
thrombospondin type-1 domain-containing protein 7A isoform X1 [Pan troglodytes]	3376	3376	100%	0.0	99.46%	XP_527669.3
thrombospondin type-1 domain-containing protein 7A [Gorilla gorilla gorilla]	3371	3371	100%	0.0	99.28%	XP_004045160.2
thrombospondin type-1 domain-containing protein 7A [Pongo abelii]	3343	3343	100%	0.0	98.31%	XP_002818243.2
thrombospondin type-1 domain-containing protein 7A [Nomascus leucogenys]	3343	3343	100%	0.0	98.31%	XP_003252614.2
thrombospondin type-1 domain-containing protein 7A [Hylobates moloch]	3343	3343	100%	0.0	98.31%	XP_032616349.1
THSD7A isoform 1 [Pongo abelii]	3340	3340	100%	0.0	98.25%	PNJ69867.1
thrombospondin type-1 domain-containing protein 7A [Theropithecus gelada]	3326	3326	100%	0.0	97.95%	XP_025235320.1
thrombospondin type-1 domain-containing protein 7A isoform X1 [Papio anubis]	3323	3323	100%	0.0	97.89%	XP_017812404.2
thrombospondin type-1 domain-containing protein 7A [Piliocolobus tephrosceles]	3322	3322	100%	0.0	97.77%	XP_026305173.1

Bu diziler için hizalama grafiği kabaca aşağıdaki gibi elde edilmiştir. Kırmızı renkli eşleşmeler, eşleşme puanı 200 ve üzerinde olanları ifade etmektedir.



Eşleşen dizilerden bir tanesi için eşleşen kısımların bir bölümünün sonuç dökümü ise aşağıdaki gibi elde edilmiştir.

PREDICTED: thrombospondin type-1 domain-containing protein 7A isoform X1 [Chlorocebus sabaeus]

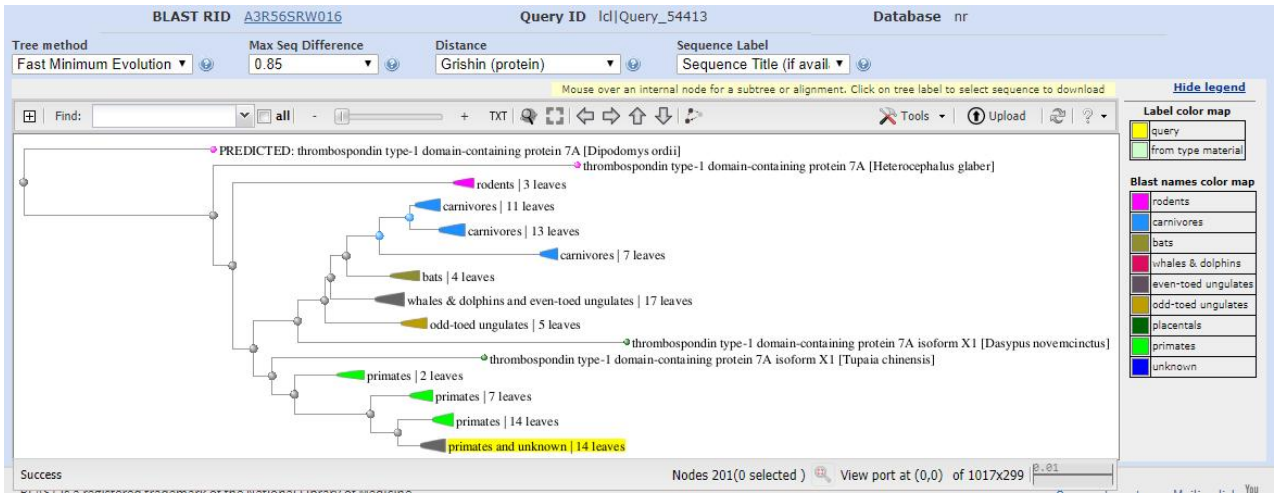
Sequence ID: XP_007980223.1 Length: 1654 Number of Matches: 1

Range 1: 1 to 1654 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
3314 bits(8592)	0.0	Compositional matrix adjust.	1620/1658(98%)	1637/1658(98%)	5/1658(0%)

```
Query  1     MGLQARRWASGSRGAAGPRRGVLQLLPLPLPLLLLLLLRPGA-GRAAAQGEAEAPTLY  59
             MGLQAR WASGSRGAAGPRRGVLQLLPL L LPLLLLL   PGA GRAAAQGEAE PTLY
Sbjct  1     MGLQARCWASGSRGAAGPRRGVLQLLPLLL-LPLLLLL---PGAAGRAAAQGEAEVPTLY  56

Query  60    LWKTGPWGRCMGDECGPGGIQTRAVWCAHVEGWTTLHTNCKQAERPNNQQNCFKVCDWHK  119
             LWKTGPWGRCMGDECGPGGIQTRAVWCAHVEGWTTLHTNCKQAERPNNQQNCFKVCDWHK
Sbjct  57    LWKTGPWGRCMGDECGPGGIQTRAVWCAHVEGWTTLHTNCKQAERPNNQQNCFKVCDWHK  116

Query  120   ELYDWRLGPWNQCQPVISKSLEKPLECIKGEEGIQVREIACIQKDKDIPAEDIICEYFEP  179
             ELYDWRLGPWNQCQPVISKSLEKPLECIKGEEGIQVREI CIQKDKDIPAEDIICEYFEP
Sbjct  117   ELYDWRLGPWNQCQPVISKSLEKPLECIKGEEGIQVREIVCIQKDKDIPAEDIICEYFEP  176

Query  180   KPLLEQACLIPCQQDCIVSEFSAWSECSKTCGSGLQHRTRHVVAPPQFGGSGCPNLTEFQ  239
             KPLLEQACLIPCQQDCIVSEFSAWSECSKTCGSGLQHRTRHVVAPPQFGGSGCPNLTEFQ
Sbjct  177   KPLLEQACLIPCQQDCIVSEFSAWSECSKTCGSGLQHRTRHVVAPPQFGGSGCPNLTEFQ  236

Query  240   VCQSSPCEAEELRYSLHVGPWSTCSMPHSRQVRQARRRGKNKEREKDRSKGVKDPEAREL  299
             VCQSSPCEAEEL YSLHVGPWSTCSMPHSRQVRQARRRGKNKEREKDRSKGVKDPEAREL
Sbjct  237   VCQSSPCEAEELMYSLHVGPWSTCSMPHSRQVRQARRRGKNKEREKDRSKGVKDPEAREL  296

Query  300   IKKKRNRNRQNRQENKYWDIQIGYQTREVMCINKTGKAADLSFCQQEKLPMTFQSCVITK  359
             IKKKRNRNRQNRQENKYWDIQIGYQTREVMCINKTGKAADLSFCQQEKLPMTFQSCVITK
Sbjct  297   IKKKRNRNRQNRQENKYWDIQIGYQTREVMCINKTGKAADLSFCQQEKLPMTFQSCVITK  356

Query  360   ECQVSEWSEWSPCSKTCHDMVSPAGTRVRTRTIRQFPIGSEKECPEFEEKEPCLSQGDGV  419
             ECQVSEWSEWSPCSKTCHDMVSPAGTRVRTRTIRQFPIGSEKECPE EEKEPCLSQGDGV
Sbjct  357   ECQVSEWSEWSPCSKTCHDMVSPAGTRVRTRTIRQFPIGSEKECPELEEKEPCLSQGDGV  416

Query  420   VPCATYGWRTTEWTECRVDPLLSQQDKRRGNQTALCGGGIQTREVYCVQANENLLSQLST  479
              PCATYGWRTTEWTECRVDPLLSQQDKRRGNQTALCGGGIQTREVYCVQANENLLSQL+T
Sbjct  417   APCATYGWRTTEWTECRVDPLLSQQDKRRGNQTALCGGGIQTREVYCVQANENLLSQLNT  476

Query  480   HKNKEASKPMDLKLCTGPIPNTTQLCHIPCPTECEVSPWSAWGPCTYENCNDQQGKKGFK  539
             HKNKEASKP++LKLCTGPIPNTTQLCHIPCPTECEVSPWSAWGPCTYENCNDQQGKKGFK
Sbjct  477   HKNKEASKPVNLKLCTGPIPNTTQLCHIPCPTECEVSPWSAWGPCTYENCNDQQGKKGFK  536

Query  540   LRKRRITNEPTGGSGVTGNCPHLLEAIPCEEPACYDWKAVRLGNCEPDNGKECGPGTQVQ  599
             LRKRRITNEPTGGSG TGNCPHLLEAIPCEEPACYDWKAVRLG+CEPDNGKECGPGTQVQ
Sbjct  537   LRKRRITNEPTGGSGGTGNCPHLLEAIPCEEPACYDWKAVRLGDCEPDNGKECGPGTQVQ  596

Query  600   EVVCINSDGEEVDRQLCRDAIFPIPVACDAPCPKDCVLSTWSTWSSCSHTCSGKTTEGKQ  659
             EVVCINSDGEEVDRQLCRDAIFPIPVACDAPCPKDCVLSTWSTWSSCSHTCSGKTTEGKQ
Sbjct  597   EVVCINSDGEEVDRQLCRDAIFPIPVACDAPCPKDCVLSTWSTWSSCSHTCSGKTTEGKQ  656
```

Benzerliklerine göre sorgu sonucunda elde edilen veriler için dizilerin benzerlik haritası ise aşağıdaki gibidir.



3. YARARLANILAN KAYNAKLAR

[1]http://resources.qiagenbioinformatics.com/manuals/clcmainworkbench/800/index.php?manual=How_does_BLAST_work.html

[2]https://bio.libretexts.org/Bookshelves/Cell_and_Molecular_Biology/Book%3A_Investigations_in_Molecular_Cell_Biology_(O'Connor)/9%3A_Protein_Conservation/9.3%3A_BLAST_algorithms_are_used_to_search_databases

[3]https://en.wikipedia.org/wiki/BLAST_(biotechnology)

[4]https://biotechgo.org/tr/?view=article&id=247:lo2&catid=136

[5]https://www.ccg.unam.mx/~vinuesa/tlem/pdfs/Bioinformatics_explained_BLAST.pdf

[6]http://www.yasinhoca.com/2016/11/dna-dizilerini-karslastrma-blast.html

[7]https://www.comp.nus.edu.sg/~ksung/algo_in_bioinfo/slides/Ch5_database.pdf

[8]https://bioinf.comav.upv.es/courses/biotech3/theory/blast.html

[9]https://www.youtube.com/watch?v=4AcnZnZRss8

[10]https://www.youtube.com/watch?v=g0nSH17psDc

[11]https://www.youtube.com/watch?v=NnY2f5111FU