Lecture 07
molecular sequences

Course: Practical Bioinformatics (BIOL 4220)
Instructor: Michael Landis
Email: michael.landis@wustl.edu

Lecture 07 outline

Last time: shell scripts

This time: molecular sequences

Topics

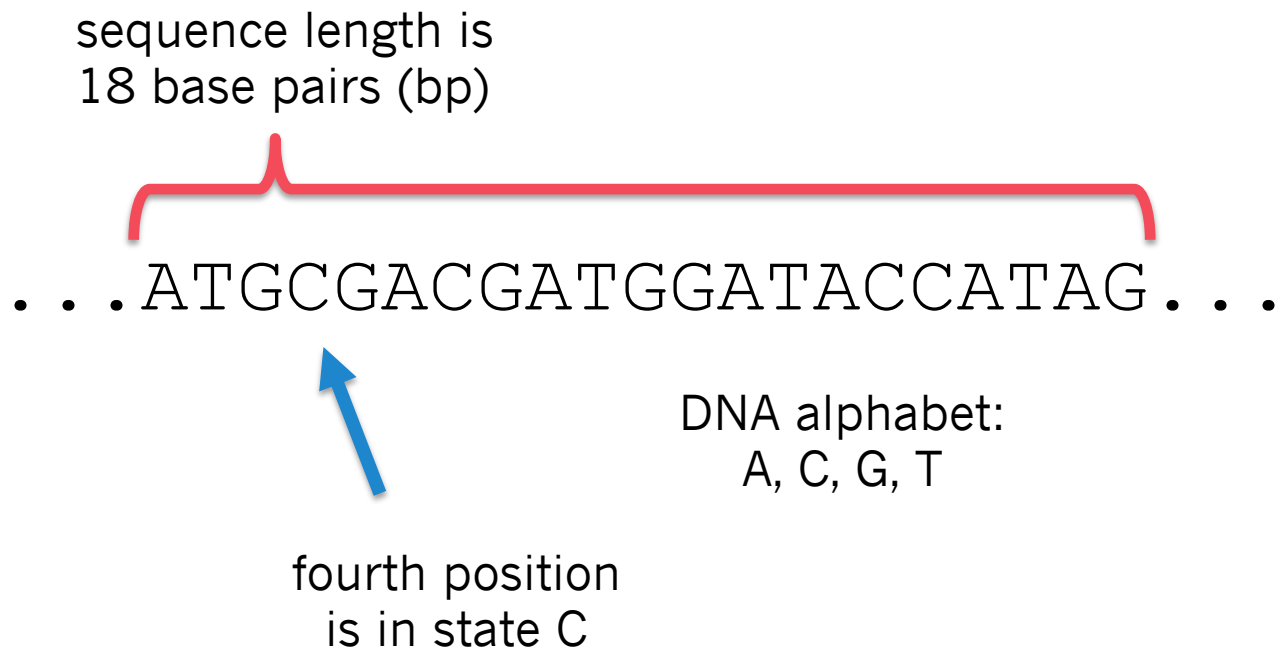
- sequence data
- GenBank
- BLAST

A ***molecular sequence*** is a string of characters from a molecular alphabet.

Examples: *DNA, RNA, amino acid sequences*

Sequences are key to understanding:

- disease mechanism
- gene expression
- developmental biology
- heredity and ancestry
- protein and cell function
- biodiversity patterns

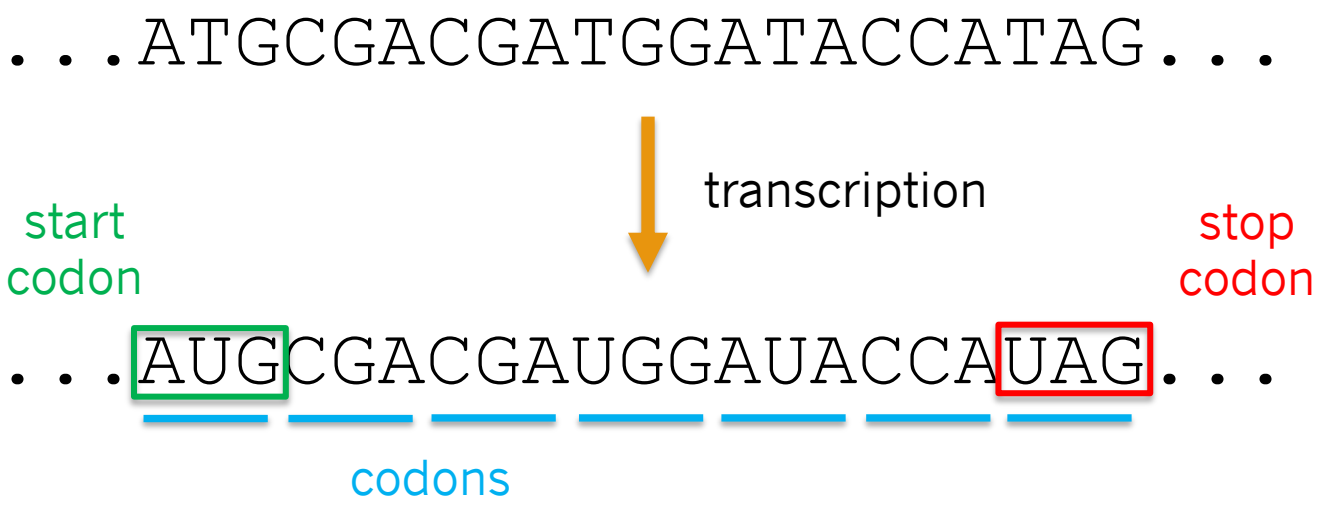

sequence length is
18 base pairs (bp)
...ATGCGACGATGGATACCATAG...
DNA alphabet:
A, C, G, T
fourth position
is in state C

...ATGCGACGATGGATACCATAG...

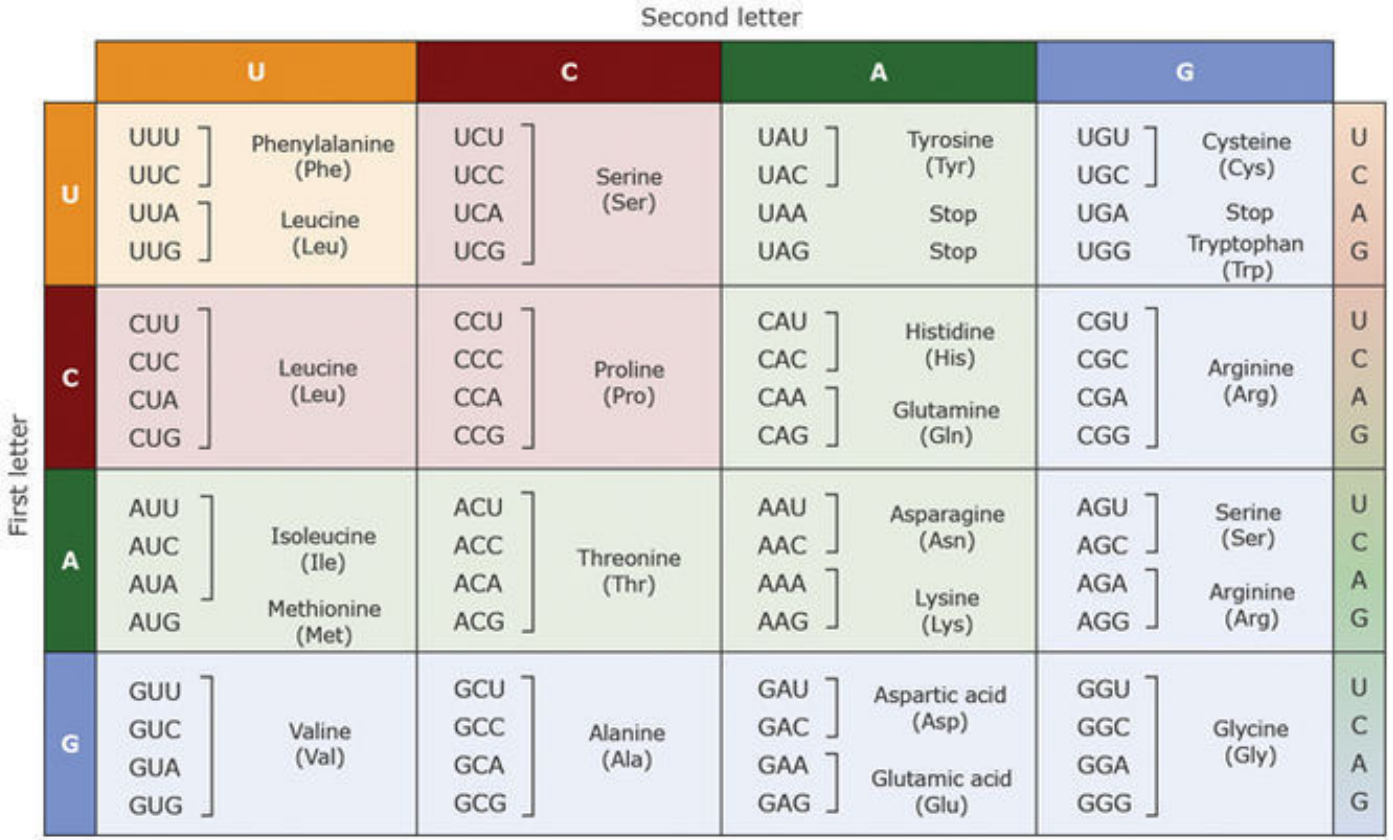
transcription

...AUGCGACGAUGGAUACCAUAG...

RNA alphabet:
A, C, G, U


...ATGCGACGATGGATACCATAG...
transcription
start codon
stop codon
...AUGCGACGAUGGAUACCAUAG...
codons

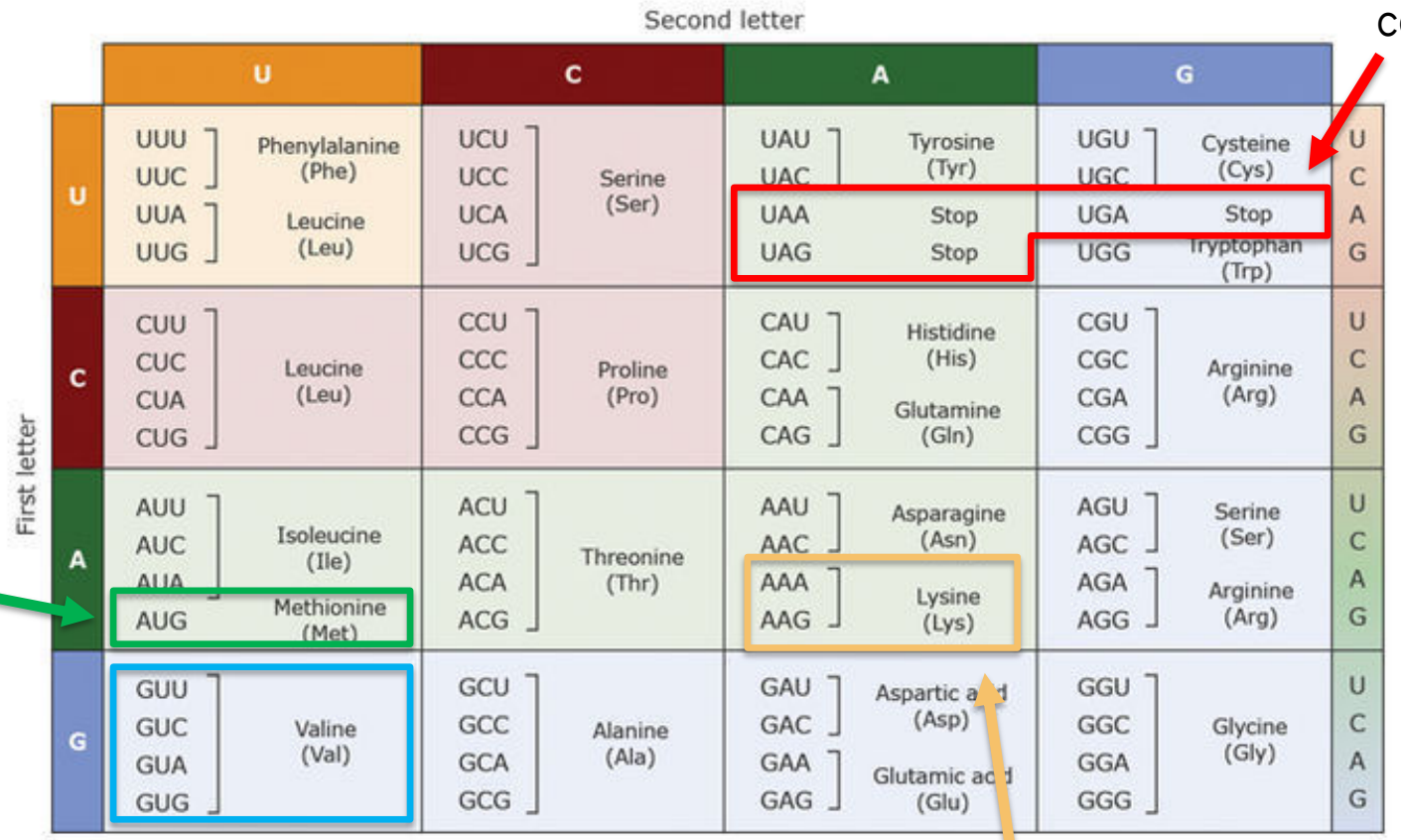
Standard genetic code



Standard genetic code

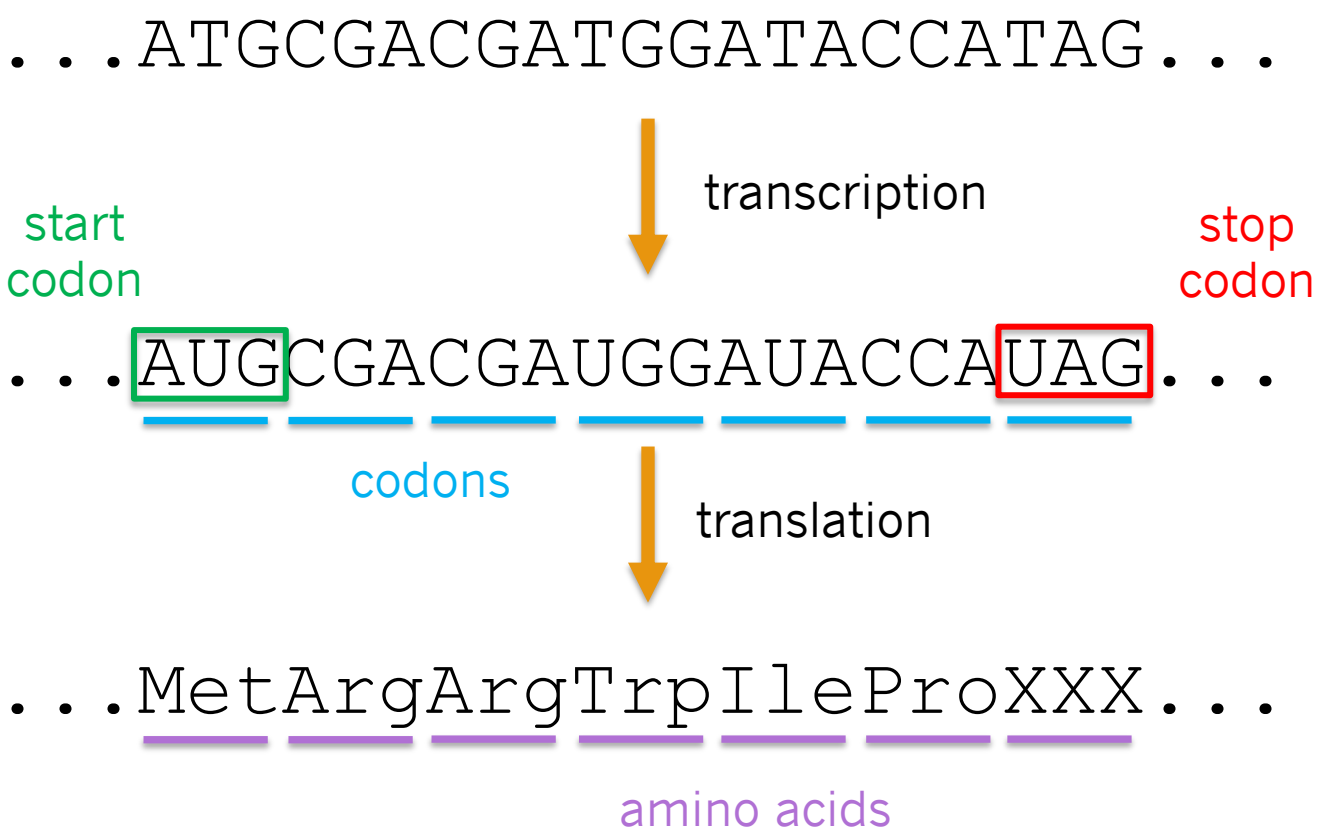
stop codons

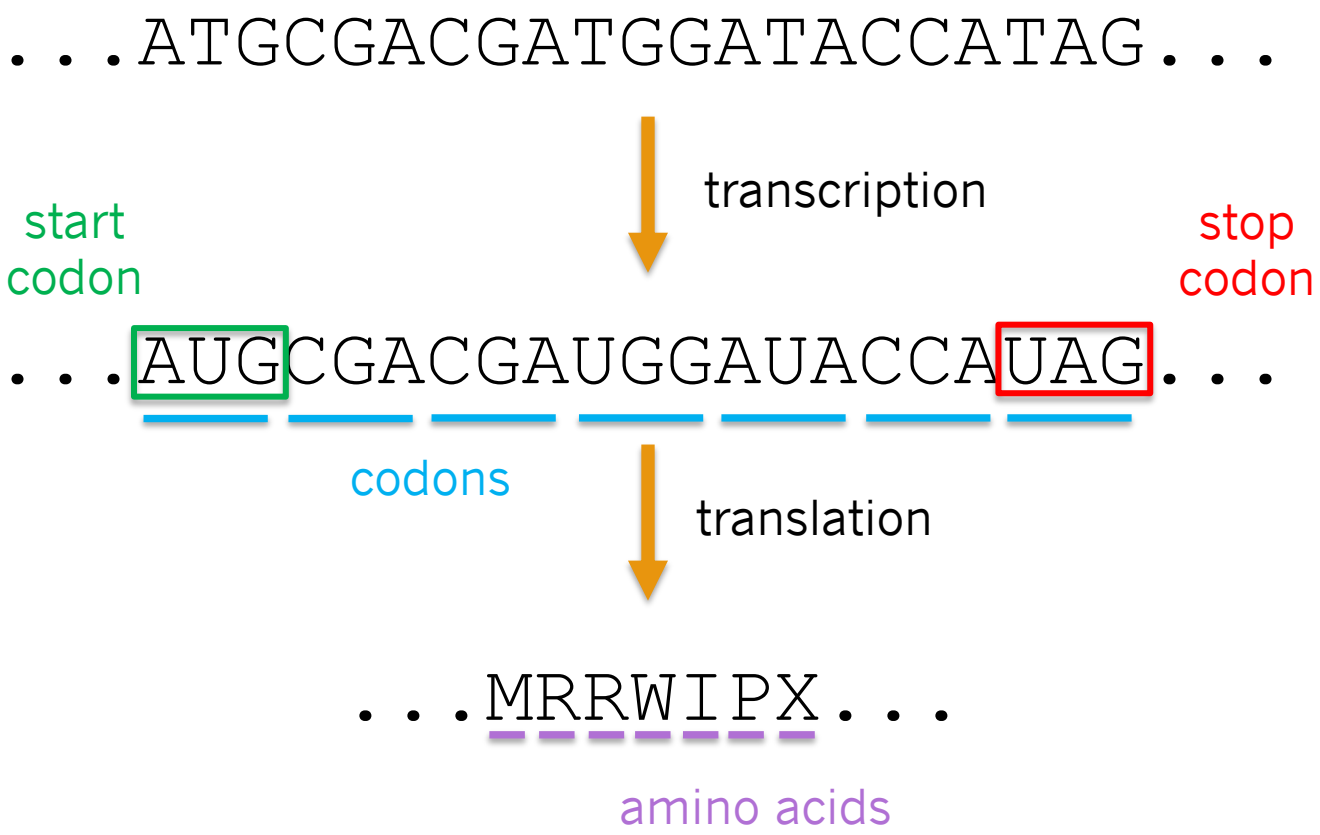
start codon



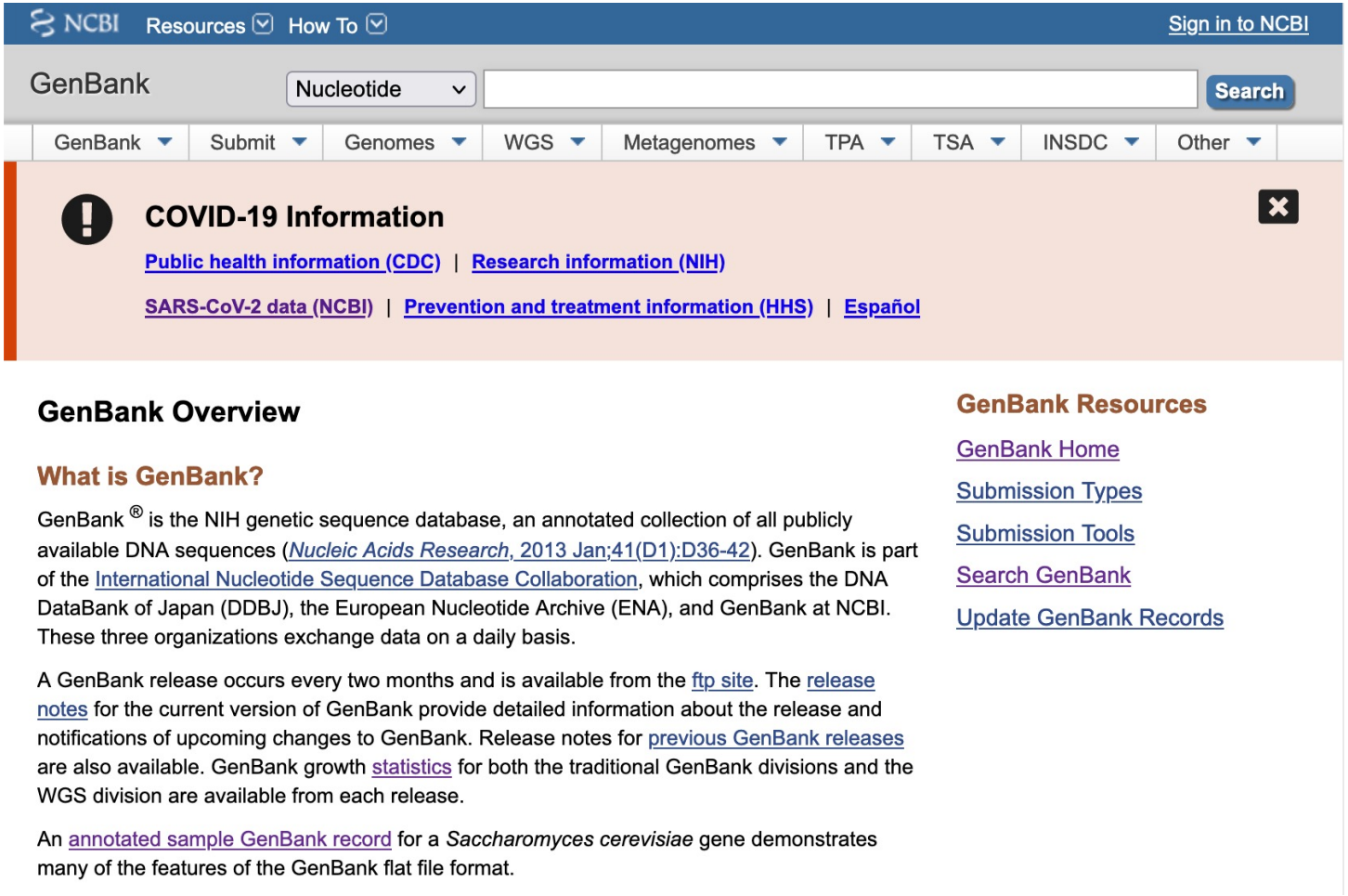
GUU, GUC, GUA, GUG all encode Valine

AAA and AAG encode Lysine


...ATGCGACGATGGATACCATAG...
transcription
start codon
stop codon
...AUGCGACGAUGGAUACCAUAG...
codons
translation
...MetArgArgTrpIleProXXX...
amino acids


...ATGCGACGATGGATACCATAG...
transcription
start codon
stop codon
...AUGCGACGAUGGAUACCAUAG...
codons
translation
...MRRWIPX...
amino acids

NCBI GenBank



https://www.ncbi.nlm.nih.gov/genbank/

GenBank sequences

GenBank publicly hosts (Aug ‘21)

- 231,982,592 sequences
- 940,513,260,726 bases

NCBI sequences are used extensively

- identifying anonymous sequences
- inferring gene function
- searching for drug targets
- expanding datasets
- met-analyses

Homo sapiens isolate C15542T cytochrome b (cytb) gene, partial cds; mitochondrial

GenBank: JN034136.1

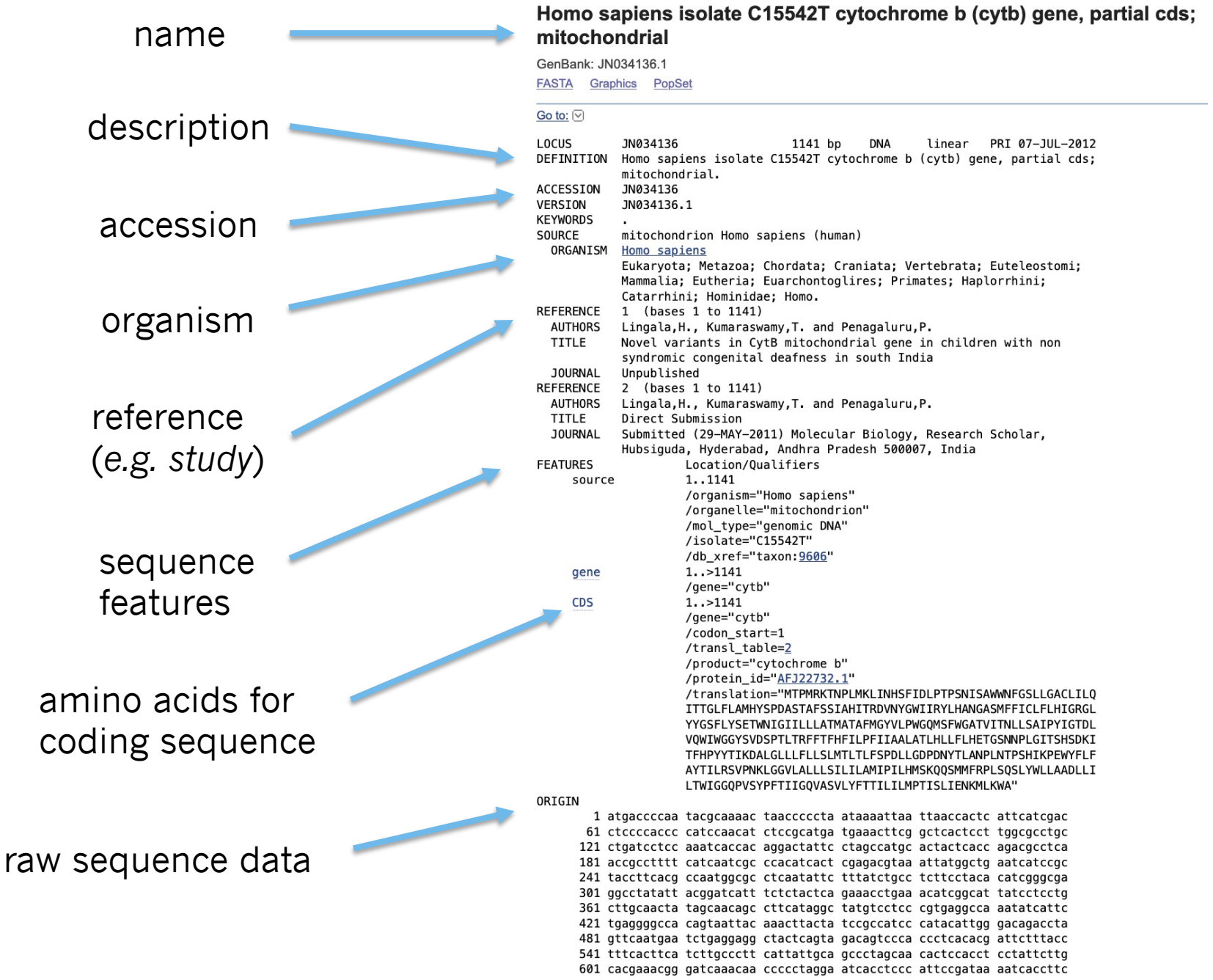
FASTA Graphics PopSet

Go to:

```
LOCUS       JN034136                1141 bp    DNA     linear   PRI 07-JUL-2012
DEFINITION  Homo sapiens isolate C15542T cytochrome b (cytb) gene, partial cds;
            mitochondrial.
ACCESSION   JN034136
VERSION     JN034136.1
KEYWORDS    .
SOURCE      mitochondrion Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1141)
  AUTHORS   Lingala,H., Kumaraswamy,T. and Penagaluru,P.
  TITLE     Novel variants in CytB mitochondrial gene in children with non
            syndromic congenital deafness in south India
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1141)
  AUTHORS   Lingala,H., Kumaraswamy,T. and Penagaluru,P.
  TITLE     Direct Submission
  JOURNAL   Submitted (29-MAY-2011) Molecular Biology, Research Scholar,
            Hubsiguda, Hyderabad, Andhra Pradesh 500007, India
```

```
FEATURES             Location/Qualifiers
     source          1..1141
                     /organism="Homo sapiens"
                     /organelle="mitochondrion"
                     /mol_type="genomic DNA"
                     /isolate="C15542T"
                     /db_xref="taxon:9606"
     gene            1..>1141
                     /gene="cytb"
     CDS             1..>1141
                     /gene="cytb"
                     /codon_start=1
                     /transl_table=2
                     /product="cytochrome b"
                     /protein_id="AFJ22732.1"
                     /translation="MTPMRKTNPLMKLINHSFIDLPTPSNISAWWNFGSLLGACLILQ
                     ITTGLFLAMHYSPDASTAFSSIAHITRDVNYGWIIRYLHANGASMFFICLFLHIGRGL
                     YYGSFLYSETWNIGIILLLATMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTDL
                     VQWIWGGYSVDSPTLTRFFTFHFILPFIIAALATLHLLFLHETGSNNPLGITSHSDKI
                     TFHPYYTIKDALGLLLFLLSLMTLTLFSPDLLGDPDNYTLANPLNTPSHIKPEWYFLF
                     AYTILRSVPNKLGGVLALLLSILILAMIPILHMSKQQSMMFRPLSQSLYWLLAADLLI
                     LTWIGGQPVSYPFTIIGQVASVLYFTTILILMPTISLIENKMLKWA"
```

```
ORIGIN
        1 atgaccccaa tacgcaaaac taacccccta ataaaattaa ttaaccactc attcatcgac
       61 ctccccaccc catccaacat ctccgcatga tgaaacttcg gctcactcct tggcgcctgc
      121 ctgatcctcc aaatcaccac aggactattc ctagccatgc actactcacc agacgcctca
      181 accgcctttt catcaatcgc ccacatcact cgagacgtaa attatggctg aatcatccgc
      241 taccttcacg ccaatggcgc ctcaatattc tttatctgcc tcttcctaca catcgggcga
      301 ggcctatatt acggatcatt tctctactca gaaacctgaa acatcggcat tatcctcctg
      361 cttgcaacta tagcaacagc cttcataggc tatgtcctcc cgtgaggcca aatatcattc
      421 tgaggggcca cagtaattac aaacttacta tccgccatcc catacattgg gacagaccta
      481 gttcaatgaa tctgaggagg ctactcagta gacagtccca ccctcacacg attctttacc
      541 tttcacttca tcttgccctt cattattgca gccctagcaa cactccacct cctattcttg
      601 cacgaaacgg gatcaaacaa ccccctagga atcacctccc attccgataa aatcaccttc
      661 cacccttact acacaatcaa agacgccctc ggcttacttc tcttccttct ctccttaatg
      721 acattaacac tattctcacc agacctccta ggcgacccag acaattatac cctagccaac
      781 cccttaaaca ccccttccca catcaagccc gaatgatatt tcctattcgc ctacacaatt
      841 ctccgatccg tccctaacaa actaggaggc gtccttgccc tattactatc catcctcatc
      901 ctagcaataa tccccatcct ccatatatcc aaacaacaaa gcataatatt tcgcccacta
      961 agccaatcac tttattgact cctagccgca gacctcctca ttctaacctg aatcggagga
     1021 caaccagtaa gctacccttt taccatcatt ggacaagtag catccgtact atacttcaca
     1081 acaatcctaa tcctaatacc aactatctcc ctaattgaaa acaaaatact caaatgggcc
     1141 t
//
```


name
description
accession
organism
reference
(e.g. study)
sequence
features
amino acids for
coding sequence
raw sequence data
Homo sapiens isolate C15542T cytochrome b (cytb) gene, partial cds; mitochondrial
GenBank: JN034136.1
FASTA Graphics PopSet
Go to:
LOCUS JN034136 1141 bp DNA linear PRI 07-JUL-2012
DEFINITION Homo sapiens isolate C15542T cytochrome b (cytb) gene, partial cds; mitochondrial.
ACCESSION JN034136
VERSION JN034136.1
KEYWORDS .
SOURCE mitochondrion Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 1141)
AUTHORS Lingala,H., Kumaraswamy,T. and Penagaluru,P.
TITLE Novel variants in CytB mitochondrial gene in children with non syndromic congenital deafness in south India
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 1141)
AUTHORS Lingala,H., Kumaraswamy,T. and Penagaluru,P.
TITLE Direct Submission
JOURNAL Submitted (29-MAY-2011) Molecular Biology, Research Scholar, Hubsiguda, Hyderabad, Andhra Pradesh 500007, India
FEATURES Location/Qualifiers
source 1..1141
/organism="Homo sapiens"
/organelle="mitochondrion"
/mol_type="genomic DNA"
/isolate="C15542T"
/db_xref="taxon:9606"
gene 1..>1141
/gene="cytb"
CDS 1..>1141
/gene="cytb"
/codon_start=1
/transl_table=2
/product="cytochrome b"
/protein_id="AFJ22732.1"
/translation="MTPMRKTNPLMKLINHSFIDLPTPSNISAWWNFGSLLGACLILQ
ITTGLFLAMHYSPDASTAFSSIAHITRDVNYGWIIRYLHANGASMFFICLFLHIGRGL
YYGSFLYSETWNIGIILLLATMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTDL
VQWIWGGYSVDSPTLTRFFTFHFILPFIIAALATLHLLFLHETGSNNPLGITSHSDKI
TFHPYYTIKDALGLLLFLLSLMTLTLFSPDLLGDPDNYTLANPLNTPSHIKPEWYFLF
AYTILRSVPNKLGGVLALLLSILILAMIPILHMSKQQSMMFRPLSQSLYWLLAADLLI
LTWIGGQPVSYPFTIIGQVASVLYFTTILILMPTISLIENKMLKWA"
ORIGIN
1 atgaccccaa tacgcaaaac taacccccta ataaaattaa ttaaccactc attcatcgac
61 ctccccaccc catccaacat ctccgcatga tgaaacttcg gctcactcct tggcgcctgc
121 ctgatcctcc aaatcaccac aggactattc ctagccatgc actactcacc agacgcctca
181 accgcctttt catcaatcgc ccacatcact cgagacgtaa attatggctg aatcatccgc
241 taccttcacg ccaatggcgc ctcaatattc tttatctgcc tcttcctaca catcgggcga
301 ggcctatatt acggatcatt tctctactca gaaacctgaa acatcggcat tatcctcctg
361 cttgcaacta tagcaacagc cttcataggc tatgtcctcc cgtgaggcca aatatcattc
421 tgaggggcca cagtaattac aaacttacta tccgccatcc catacattgg gacagaccta
481 gttcaatgaa tctgaggagg ctactcagta gacagtccca ccctcacacg attctttacc
541 tttcacttca tcttgccctt cattattgca gccctagcaa cactccacct cctattcttg
601 cacgaaacgg gatcaaacaa ccccctagga atcacctccc attccgataa aatcaccttc

Genome collections

Viruses > Riboviria > Orthornavirae > Kitrinoviricota > Alsuviricetes > Martellivirales > Togaviridae >

Alphavirus - 34 complete genomes

Retrieve sequences: -- Select data set from the list --

Species [33] unclassified Alphavirus [1]

* The list view for each taxonomy node shows only the next level of sublineages.
* Unclassified/unassigned names are written in copper

Sort the genomes list by: taxonomy | Filter by host: all hosts | >Download

Show / hide all segment lists

Genome	Accession	RefSeq type	Source information	Segm	Length	Protein	Neighbors	Host	Created	Updated
Aura virus	NC_003900	complete		-	11824 nt	3	1		02/09/1999	08/13/2018
Barmah Forest virus	NC_001786	complete	strain:BH2193	-	11488 nt	4	35	human, invertebrates, vertebrates	01/14/1997	08/13/2018
Bebaru virus	NC_016962	complete		-	11877 nt	3	-		03/09/2012	08/13/2018
Caaingua virus	NC_055569	complete	isolate:MS681	-	12096 nt	2	-	invertebrates	06/01/2021	06/07/2021
Cabassou virus	NC_038670	complete	strain:CaAr 508	-	11385 nt	2	-		08/24/2018	08/24/2018
Chikungunya virus	NC_004162	complete	strain:S27-African prototype	-	11826 nt	2	922	human, invertebrates, vertebrates	09/06/2002	10/29/2018
Eastern equine encephalitis virus	NC_003899	complete	strain:ssp. North American variant	-	11675 nt	4	453	human, invertebrates, vertebrates	12/16/1991	08/13/2018
Eilat virus	NC_018615	complete	isolate:EO329	-	11634 nt	3	1	invertebrates	09/20/2012	08/13/2018
Everglades virus	NC_038671	complete	strain:Everglades Fe3-7c	-	11395 nt	2	-		08/24/2018	08/24/2018
Fort Morgan virus	NC_013528	complete	isolate:CM4-146	-	11381 nt	3	1	invertebrates, vertebrates	11/25/2009	08/13/2018
Getah virus	NC_006558	complete	isolate:swine	-	11597 nt	3	46	invertebrates, vertebrates	12/17/2004	08/13/2018
Highlands J virus	NC_012561	complete	isolate:585-01	-	11526 nt	3	9	invertebrates, vertebrates	04/15/2009	08/13/2018
Madariaga virus	NC_023812	complete	strain:MADV/Cebus apella/BRA/BEAN5122/1956	-	11624 nt	2	29	human, invertebrates, vertebrates	03/20/2014	08/13/2018
Mayaro virus	NC_003417	complete		-	11411 nt	4	41	human, invertebrates, vertebrates	02/22/2002	08/13/2018
Middelburg virus	NC_024887	complete	isolate:ArB-8422	-	11550 nt	2	11	invertebrates, vertebrates	09/17/2014	08/13/2018
Mosso das Pedras virus	NC_038857	complete	strain:78V-3531	-	11465 nt	2	-		08/24/2018	08/24/2018
Mucambo virus	NC_038672	complete	strain:Mucambo BeAn 8	-	11391 nt	2	1	invertebrates	08/24/2018	08/24/2018
Ndumu virus	NC_016959	complete		-	11688 nt	4	1	invertebrates, vertebrates	03/09/2012	08/13/2018
Onyong-nyong virus	NC_001512	complete		-	11835 nt	3	3	human, invertebrates	08/02/1993	08/13/2018
Pixuna virus	NC_038673	complete	strain:Pixuna BeAr 35645	-	11344 nt	2	-	vertebrates	08/24/2018	08/24/2018

BLAST: Basic Local Alignment Search Tool

What if the identity of your sequence is unknown?

Matches a ***query sequence*** against any number of ***target*** sequences in a database (*e.g.* GenBank)

Simple, versatile, and fast

Used to
- identify species
- identify gene function
- locate domains in gene
- locate gene in genome
- expand datasets
- filter out noisy data
- perform meta-analysis

BLAST searches

- For example, the search sequence word "WVH" might score above threshold with these indexed sequences:

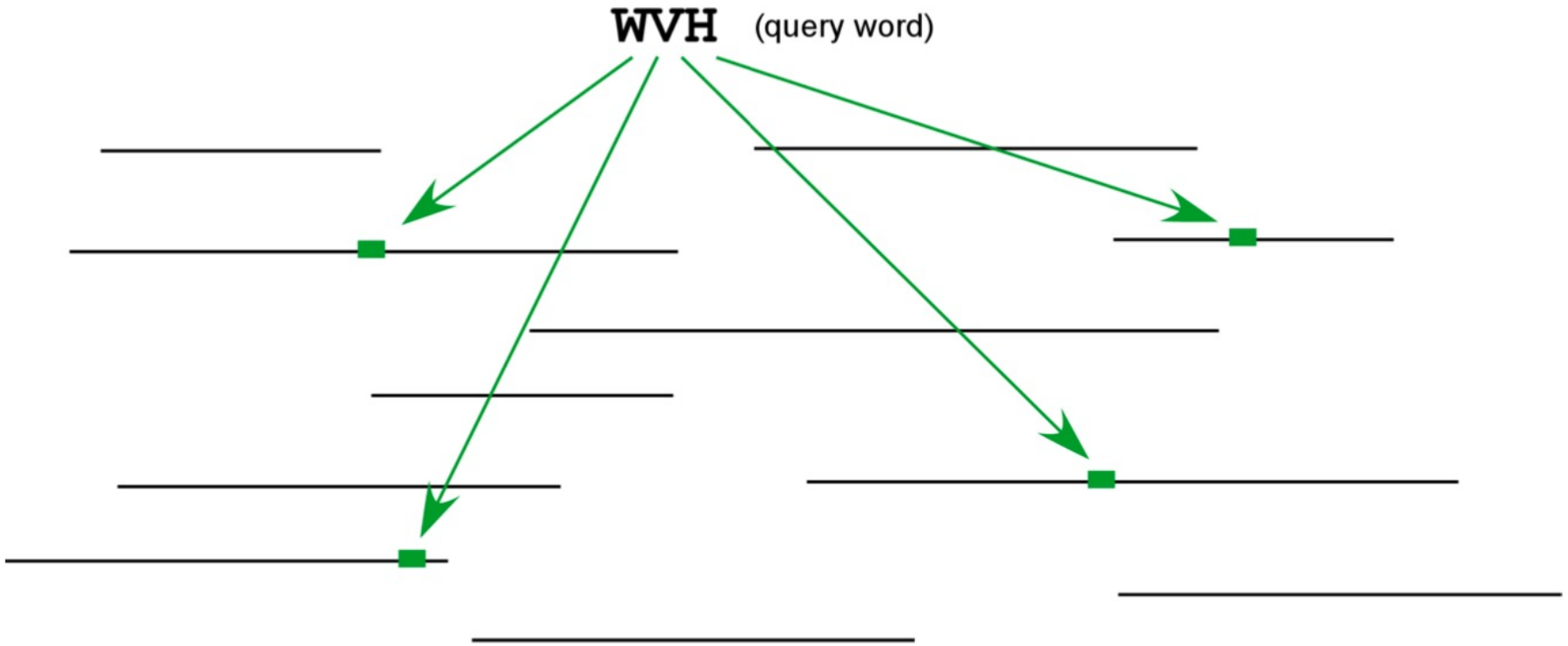
Indexed word	Score
WVH	23
WIH	22
WVY	17
WIY	16

- Target sequences around each indexed word hit are retrieved and match is extended in both directions:

```
...VFEWVHLLP...      your sequence
   ⟵ WIY ⟶           database (many sites)
```

from: http://faculty.washington.edu/jht/GS559_2017/

Schematic of indexed matches



Result – instead of aligning these 3 amino acids to <u>everything</u>, they are aligned only with the tiny fraction of sequence regions that are good candidates for a valid alignment.

from: http://faculty.washington.edu/jht/GS559_2017/

Extension and scoring

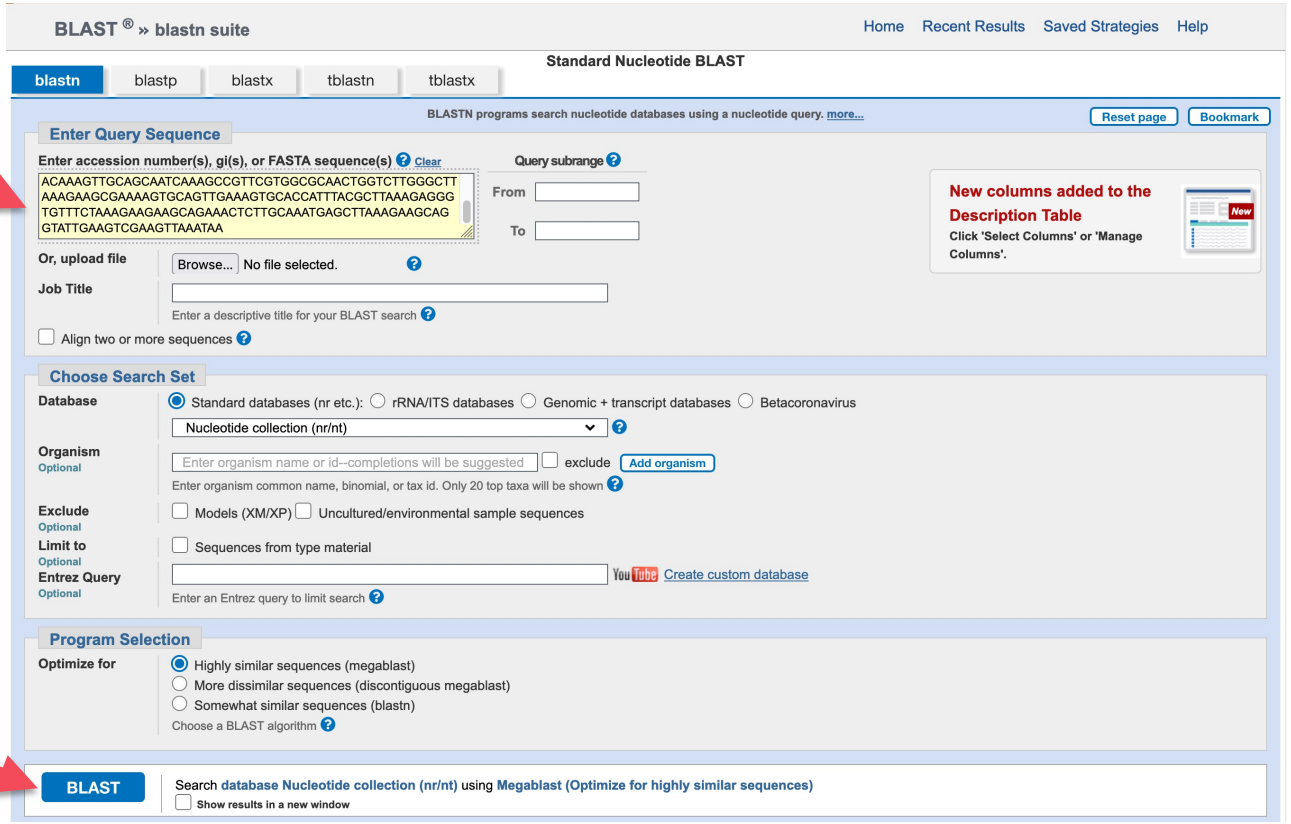
	Match Score:	Total Score:
...QSVFEWVHLLPGA... ..WIY..	16	16
...QSVFEWVHLLPGA... ..WIY**Q**..	-3	13
...QSVFEWVHLLPGA... ..WIYQ**K**..	-2	11
...QSVFEWVHLLPGA... ..WIYQK**A**..	-1	10

from: http://faculty.washington.edu/jht/GS559_2017/

NCBI BLAST interface

Input your ***query sequence*** as text

Perform BLAST against ***subject sequences***



https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch

gene from Kerfeld and Scott (PLoS Biology 2011)

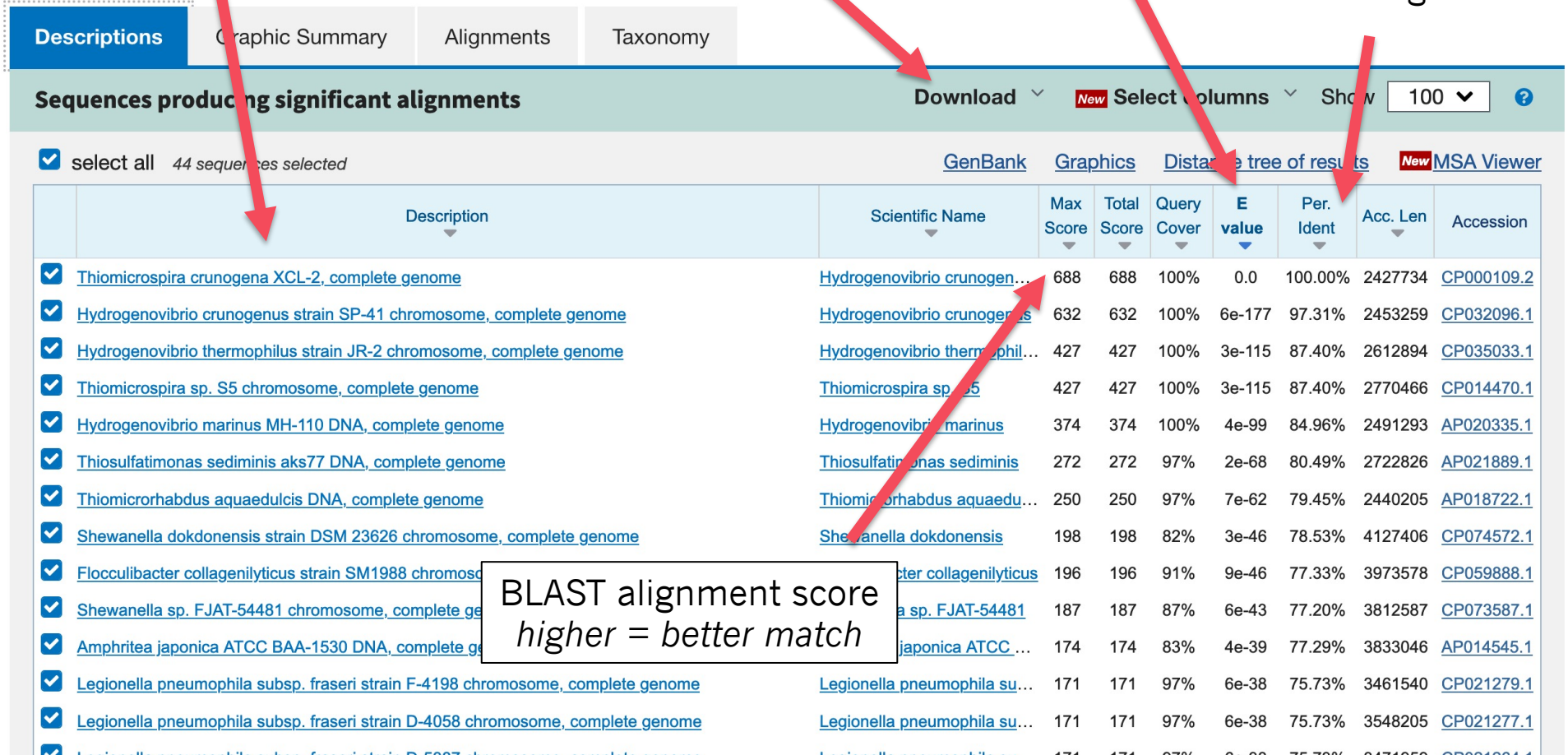
“BLAST hits”

Sequence name
e.g. species/gene ID

Download sequences

Expect value
lower = more significant

% identical characters in alignment



Descriptions | Graphic Summary | Alignments | Taxonomy

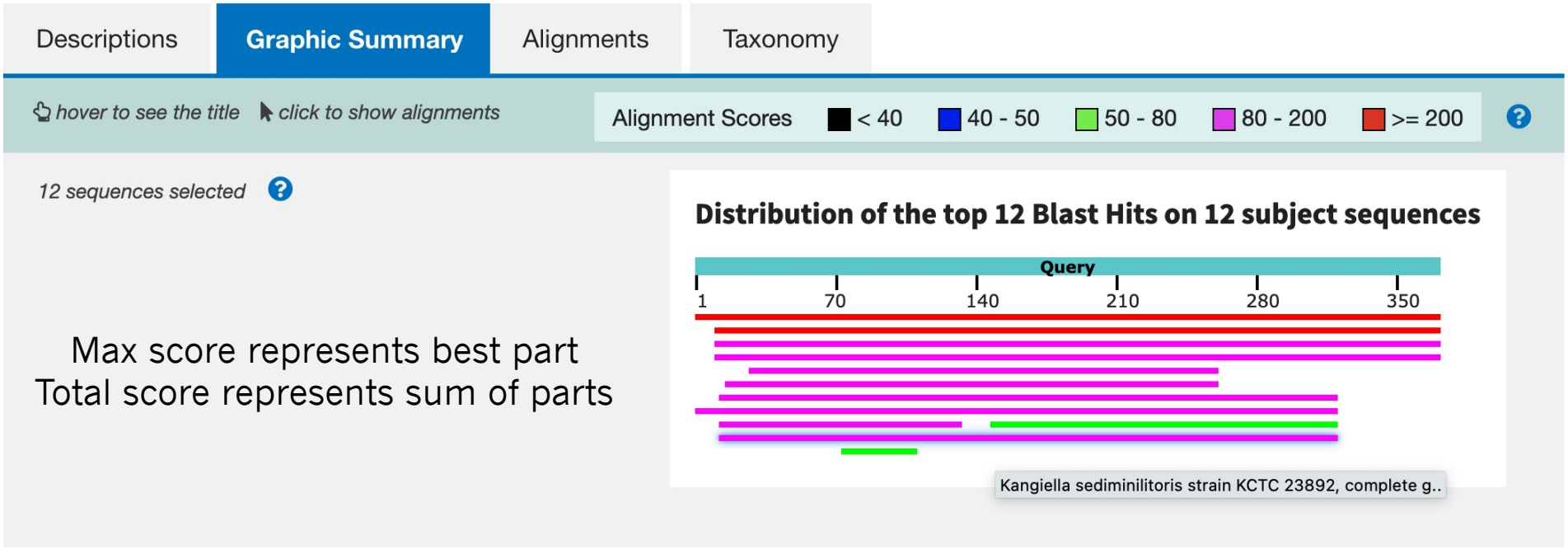
Sequences producing significant alignments Download | New Select columns | Show 100

select all *44 sequences selected* GenBank | Graphics | Distance tree of results | New MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Thiomicrospira crunogena XCL-2, complete genome	Hydrogenovibrio crunogen…	688	688	100%	0.0	100.00%	2427734	CP000109.2
Hydrogenovibrio crunogenus strain SP-41 chromosome, complete genome	Hydrogenovibrio crunogen…	632	632	100%	6e-177	97.31%	2453259	CP032096.1
Hydrogenovibrio thermophilus strain JR-2 chromosome, complete genome	Hydrogenovibrio thermophil…	427	427	100%	3e-115	87.40%	2612894	CP035033.1
Thiomicrospira sp. S5 chromosome, complete genome	Thiomicrospira sp. S5	427	427	100%	3e-115	87.40%	2770466	CP014470.1
Hydrogenovibrio marinus MH-110 DNA, complete genome	Hydrogenovibrio marinus	374	374	100%	4e-99	84.96%	2491293	AP020335.1
Thiosulfatimonas sediminis aks77 DNA, complete genome	Thiosulfatimonas sediminis	272	272	97%	2e-68	80.49%	2722826	AP021889.1
Thiomicrorhabdus aquaedulcis DNA, complete genome	Thiomicrorhabdus aquaedu…	250	250	97%	7e-62	79.45%	2440205	AP018722.1
Shewanella dokdonensis strain DSM 23626 chromosome, complete genome	Shewanella dokdonensis	198	198	82%	3e-46	78.53%	4127406	CP074572.1
Flocculibacter collagenilyticus strain SM1988 chromosc	cter collagenilyticus	196	196	91%	9e-46	77.33%	3973578	CP059888.1
Shewanella sp. FJAT-54481 chromosome, complete ge	a sp. FJAT-54481	187	187	87%	6e-43	77.20%	3812587	CP073587.1
Amphritea japonica ATCC BAA-1530 DNA, complete g	japonica ATCC …	174	174	83%	4e-39	77.29%	3833046	AP014545.1
Legionella pneumophila subsp. fraseri strain F-4198 chromosome, complete genome	Legionella pneumophila su…	171	171	97%	6e-38	75.73%	3461540	CP021279.1
Legionella pneumophila subsp. fraseri strain D-4058 chromosome, complete genome	Legionella pneumophila su…	171	171	97%	6e-38	75.73%	3548205	CP021277.1

BLAST alignment score
higher = better match

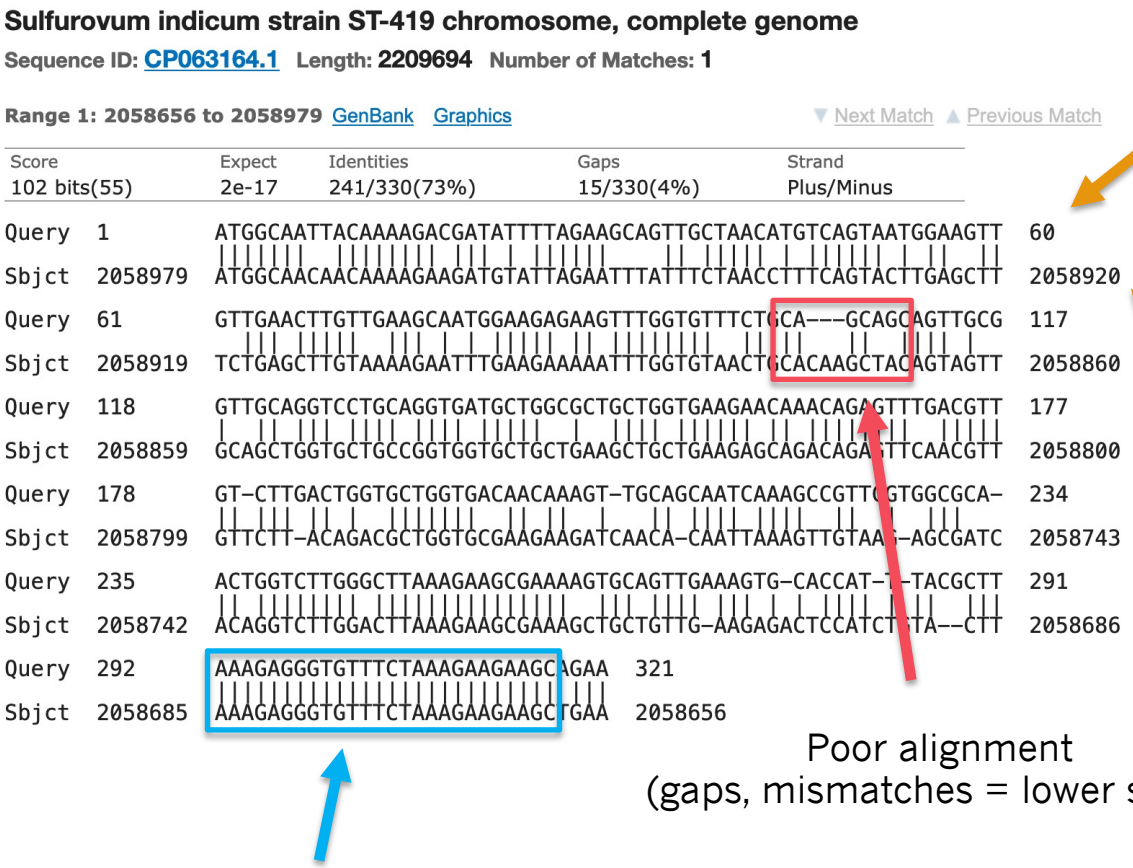
Graphic summary of BLAST scores



Max score represents best part
Total score represents sum of parts

Scores determined by alignment costs
for match/mismatch/indel reconciliation
(covered in a later lab)

BLAST generates a summary for each subject sequence



Sulfurovum indicum strain ST-419 chromosome, complete genome

Sequence ID: CP063164.1 **Length:** 2209694 **Number of Matches:** 1

Range 1: 2058656 to 2058979 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
102 bits(55)	2e-17	241/330(73%)	15/330(4%)	Plus/Minus

```
Query  1        ATGGCAATTACAAAAGACGATATTTTAGAAGCAGTTGCTAACATGTCAGTAATGGAAGTT  60
Sbjct  2058979  ATGGCAACAACAAAAGAAGATGTATTAGAAATTTATTTCTAACCTTTCAGTACTTGAGCTT  2058920
Query  61       GTTGAACTTGTTGAAGCAATGGAAGAGAAGTTTGGTGTTTCTGCA---GCAGCAGTTGCG  117
Sbjct  2058919  TCTGAGCTTGTAAAAGAATTTGAAGAAAAATTTGGTGTAACTGCACAAGCTACAGTAGTT  2058860
Query  118      GTTGCAGGTCCTGCAGGTGATGCTGGCGCTGCTGGTGAAGAACAAACAGAGTTTGACGTT  177
Sbjct  2058859  GCAGCTGGTGCTGCCGGTGGTGCTGCTGAAGCTGCTGAAGAGCAGACAGAGTTCAACGTT  2058800
Query  178      GT-CTTGACTGGTGCTGGTGACAACAAAGT-TGCAGCAATCAAAGCCGTTGGTGGCGCA-  234
Sbjct  2058799  GTTCTT-ACAGACGCTGGTGCGAAGAAGATCAACA-CAATTAAAGTTGTAAG-AGCGATC  2058743
Query  235      ACTGGTCTTGGGCTTAAAGAAGCGAAAAGTGCAGTTGAAAGTG-CACCAT-T-TACGCTT  291
Sbjct  2058742  ACAGGTCTTGGACTTAAAGAAGCGAAAGCTGCTGTTG-AAGAGACTCCATCTGTA--CTT  2058686
Query  292      AAAGAGGGTGTTTCTAAAGAAGAAGCAGAA  321
Sbjct  2058685  AAAGAGGGTGTTTCTAAAGAAGAAGCTGAA  2058656
```

Query sequence position

Subject sequence position

Poor alignment
(gaps, mismatches = lower score)

Perfect alignment
(more matches = higher score)

Overview for Lab 07