BLAST

CISS6500– 16 June 2021

Sequence Alignment

Pairwise Sequence Alignment

Local Alignment

Query Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| |||||||||||||||
5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Subject Sequence

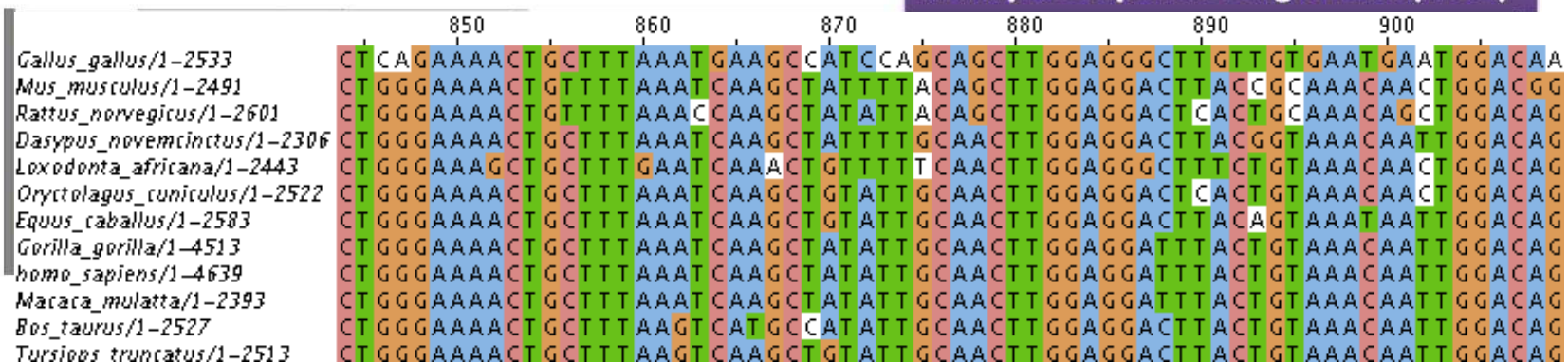
Global Alignment

Query Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  ||||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Subject Sequence

Multiple Sequence Alignment (MSA)

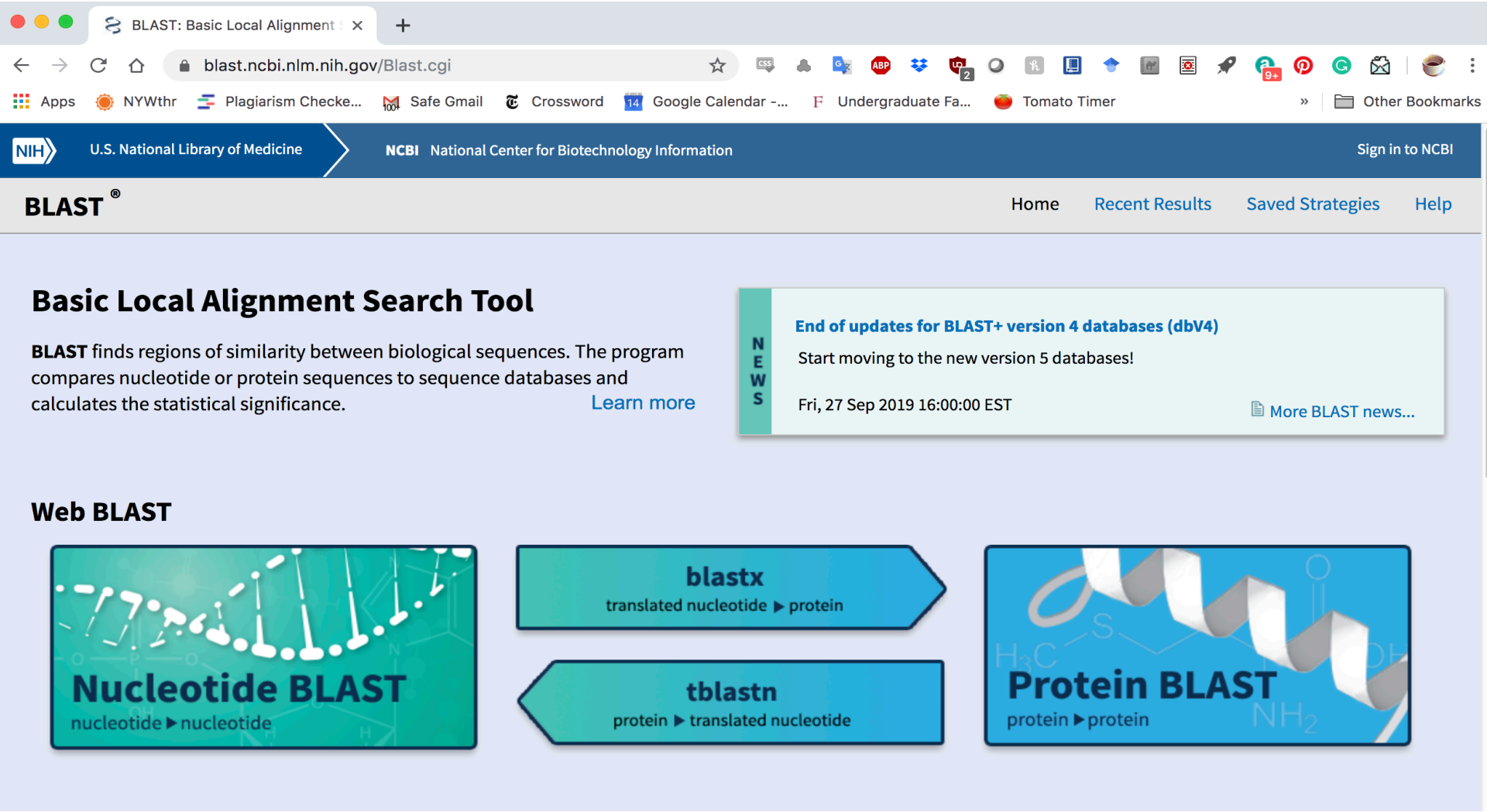


Sequence Alignment

- Algorithms compare sequences and give matches a score
  - Perfect matches get good score
  - Penalties are given for:
    - Mismatches
    - Opening a gap
    - Extending a gap
  - Total scores balance mismatches with gaps
  - ***Global alignment algorithms*** search for the optimal score for matches over an entire sequence
  - ***Local alignment algorithm***s search for the highest scoring matching regions which can be smaller portions of similarity, the entire sequence length does not need to be matched.
- Uses:
  - Multiple sequence alignments to make phylogenies
  - MSAs to find variants for medical and evolutionary studies
  - Pairwise alignments for indexing and searching databases of gene and protein sequences. -> BLAST

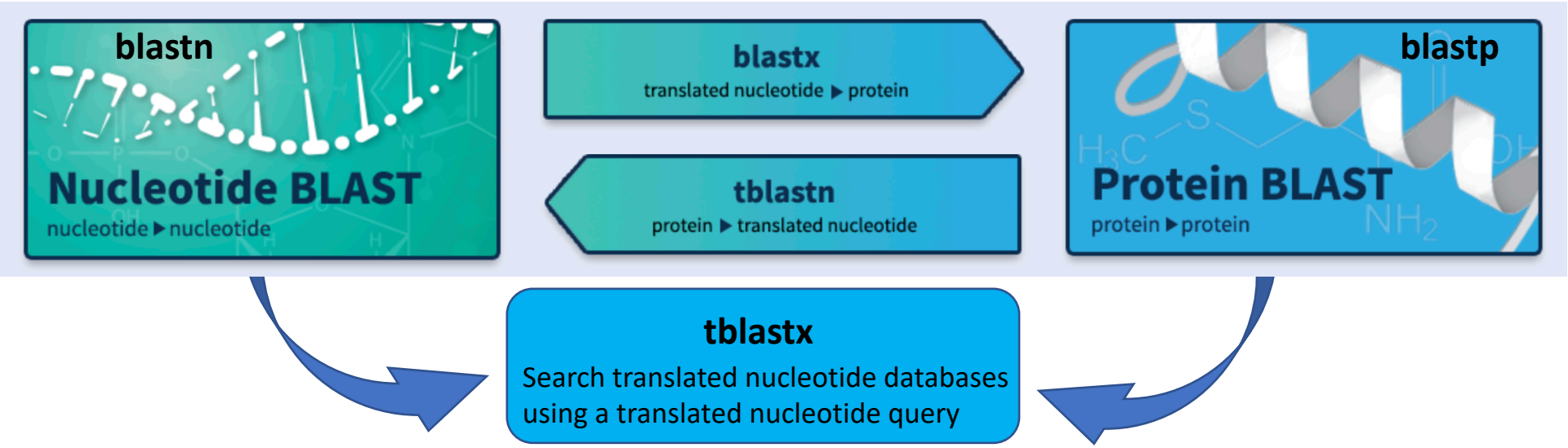
Sequence Alignment - BLAST

- Basic Local Alignment Search Tool



Sequence Alignment - BLAST

- Basic Local Alignment Search Tool
  - The most frequently used alignment program
  - Pairwise alignments
  - Used on the web and from the command line
  - Based on matching a **sequence**, the "query" with **databases** of sequence, the "subject"
  - Uses heuristics, so it is very fast



Some Genetics Review: DNA -> RNA ->Protein

ATGGTGTCTCTGTGTGTATGTCCAAATCTCCTTTTCTTATAAGGATACCAGTAAGATTGGATTAGAAC
CCACTCTAATGACCCCATCTTAACTTAATCAACCCATTTAAGGCTTGTGATGGTTAATCTTATGTGTC
AGCTTGACTGGGCGAAGGGATACCCATATAGCTACTAAAACGATTCTGGGTATGTCTGTGTTCATACT
TCTGGAAGAGATTAGCATTTAAGTCAGTAGACTGAGTAAAGAAGATCCACCCTCACCAATGTGGAT

↓

ATG•GTG•TCT•TGT•GTG•TAT•GTC•CAA•ATC•TCC•TTT•TCTTACCAGTAAGATTGGATTAGA
ACCCACTCTAATGACCCCATCTTAACTTAATCAACCCATTTAAGGCTTGTGATGGTTAATCTTATGT

↓

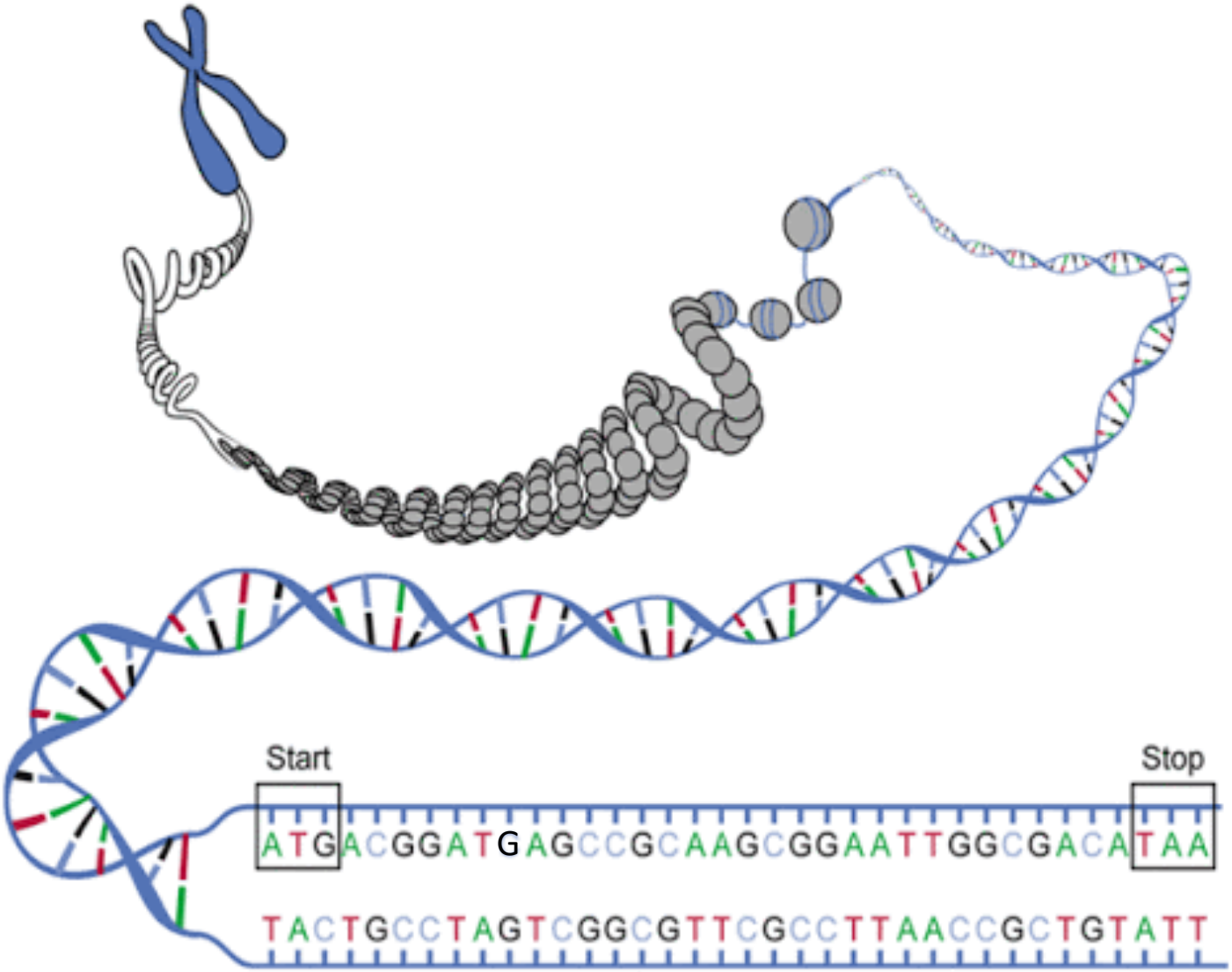
MAASFPPTLGLSSAPDEIQHPHIKFSEWKFKLFRVRSFEKTP

Codons

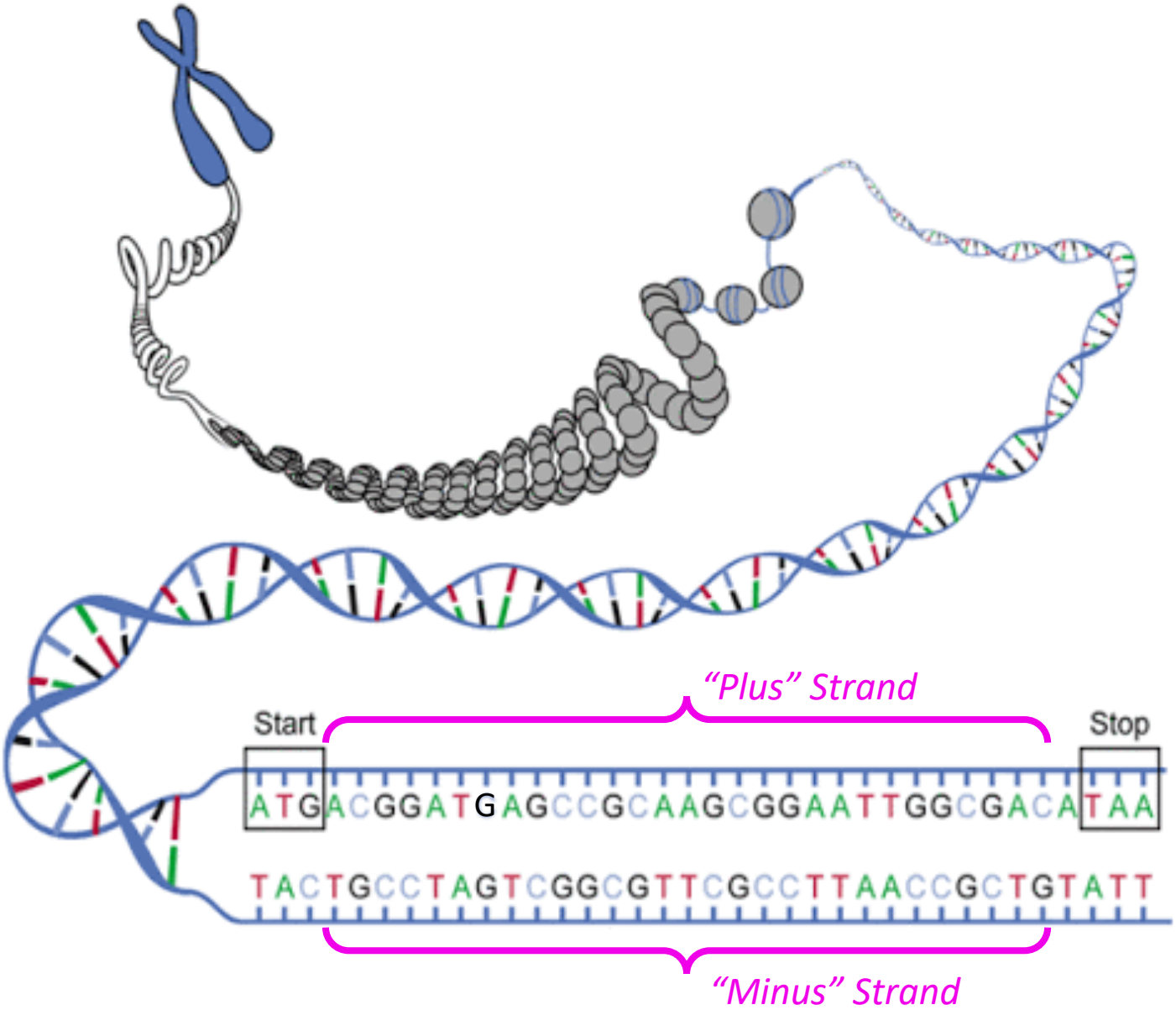
Second Nucleotide

First Nucleotide	U	C	A	G	Third Nucleotide
U	UUU, UUC Phe; UUA, UUG Leu	UCU, UCC, UCA, UCG Ser	UAU, UAC Tyr; UAA Stop; UAG Stop	UGU, UGC Cys; UGA Stop; UGG Trp	U, C, A, G
C	CUU, CUC, CUA, CUG Leu	CCU, CCC, CCA, CCG Pro	CAU, CAC His; CAA, CAG Gln	CGU, CGC, CGA, CGG Arg	U, C, A, G
A	AUU, AUC, AUA Ile; AUG Met	ACU, ACC, ACA, ACG Thr	AAU, AAC Asn; AAA, AAG Lys	AGU, AGC Ser; AGA, AGG Arg	U, C, A, G
G	GUU, GUC, GUA, GUG Val	GCU, GCC, GCA, GCG Ala	GAU, GAC Asp; GAA, GAG Glu	GGU, GGC, GGA, GGG Gly	U, C, A, G

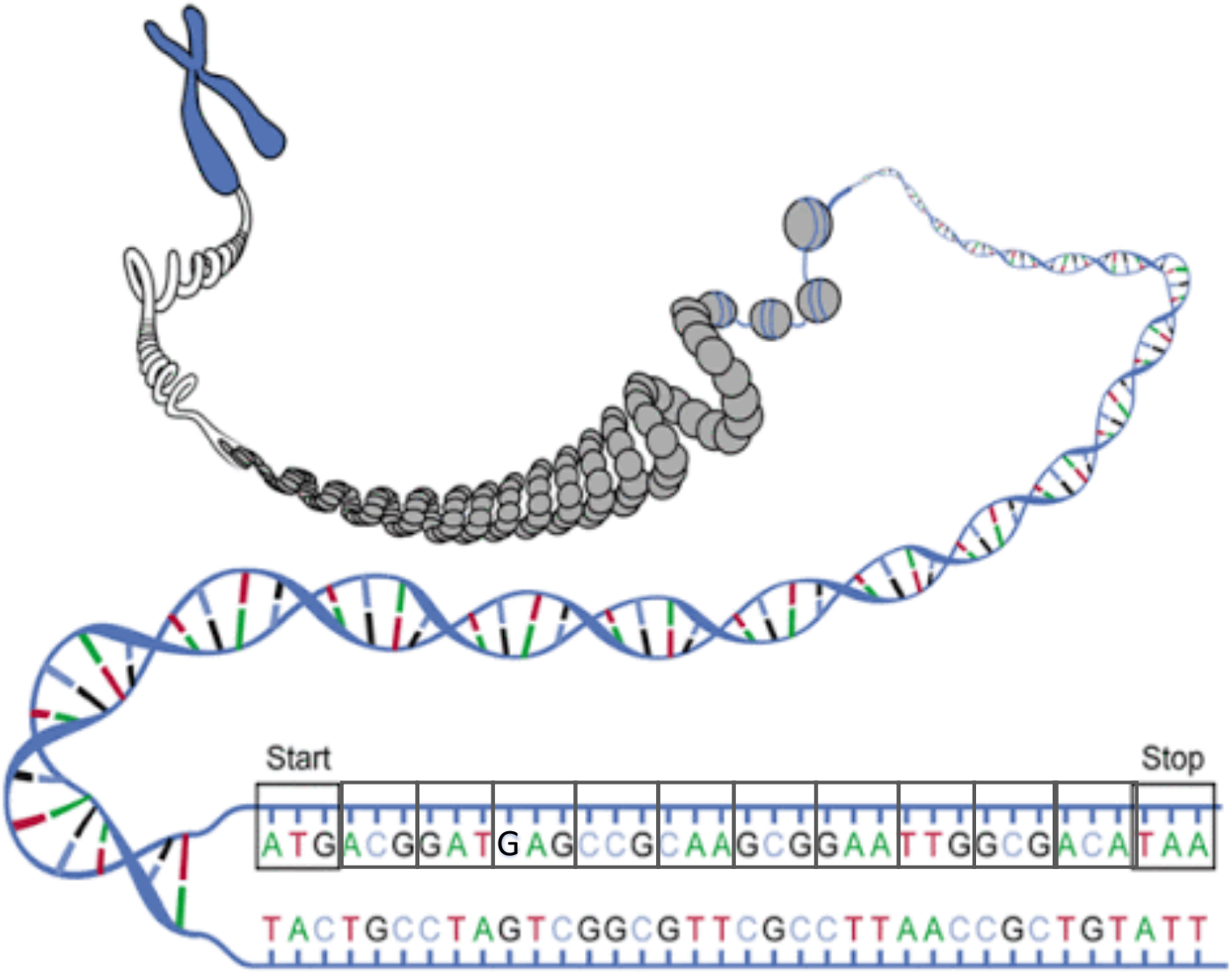
A bit of review: DNA -> RNA ->Protein



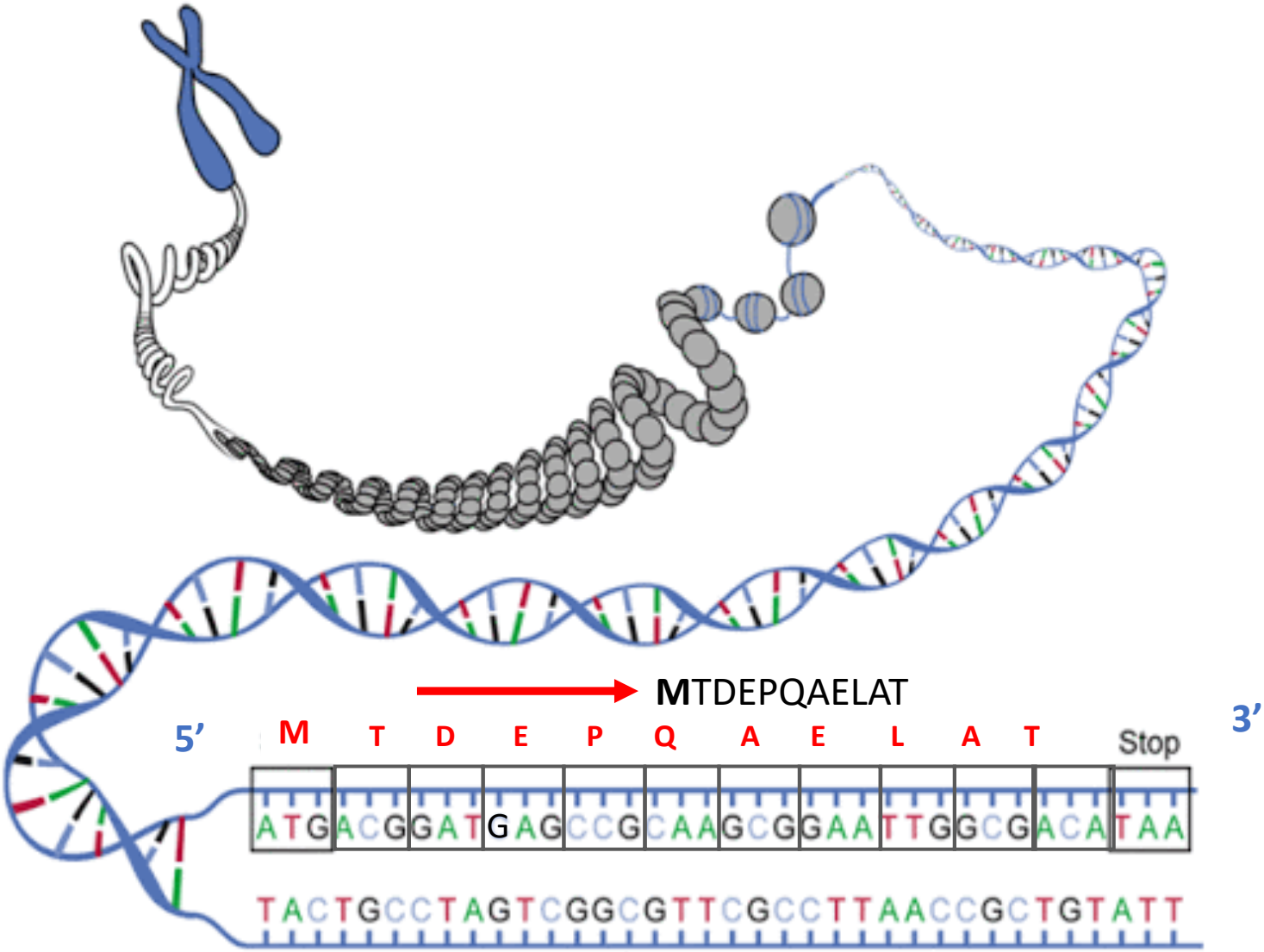
A bit of review: DNA -> RNA ->Protein



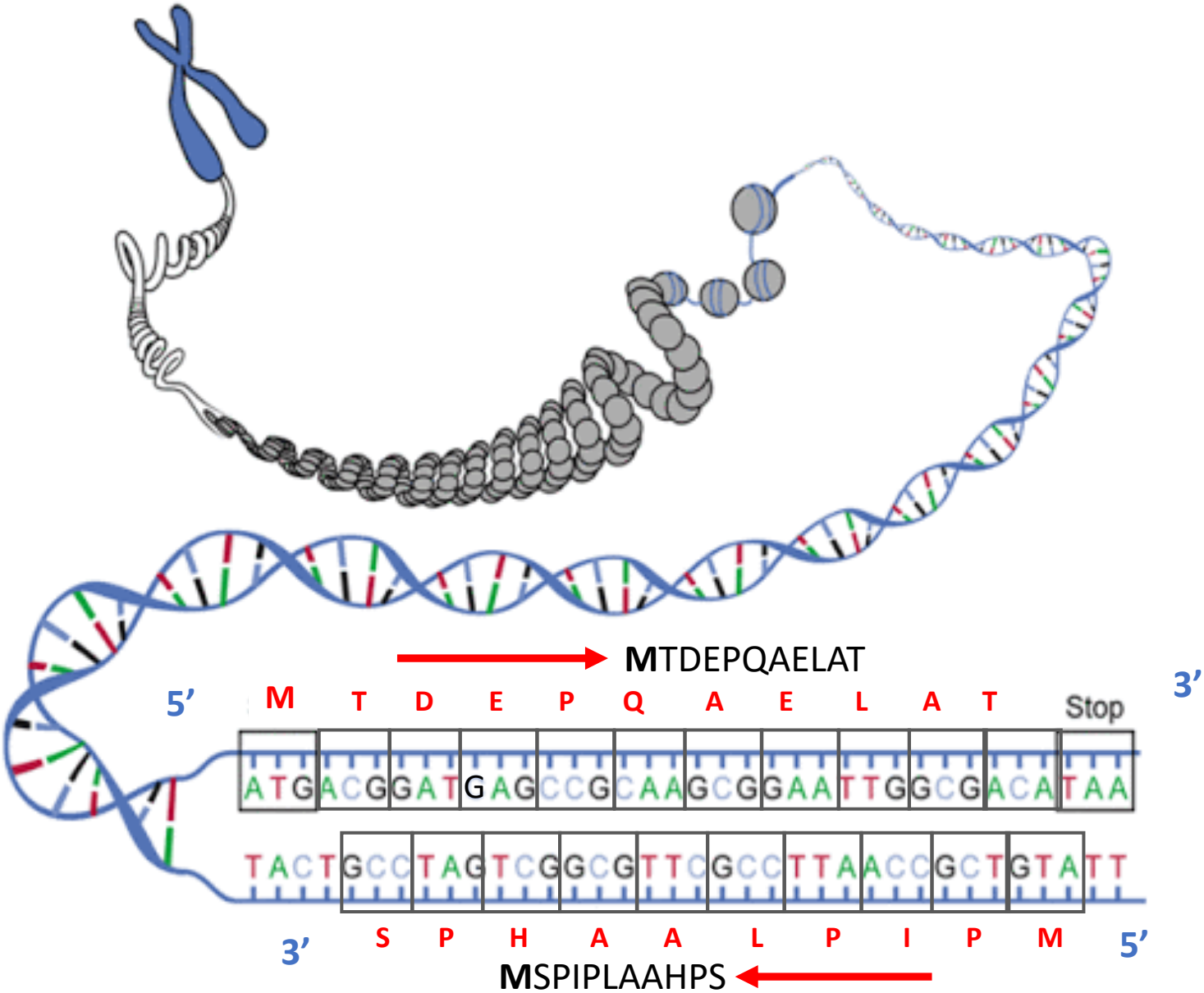
A bit of review: DNA -> RNA ->Protein



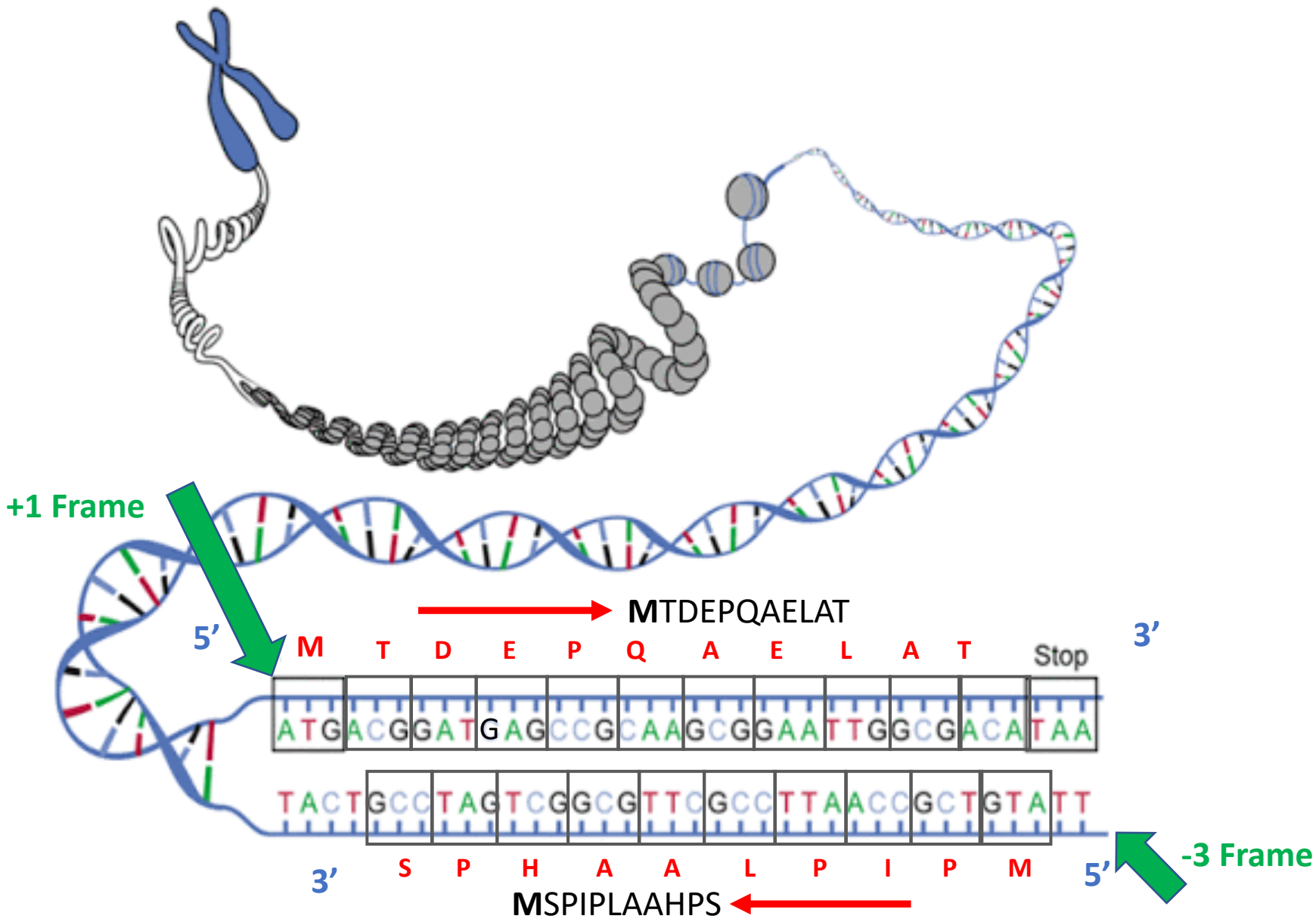
A bit of review: DNA -> RNA ->Protein



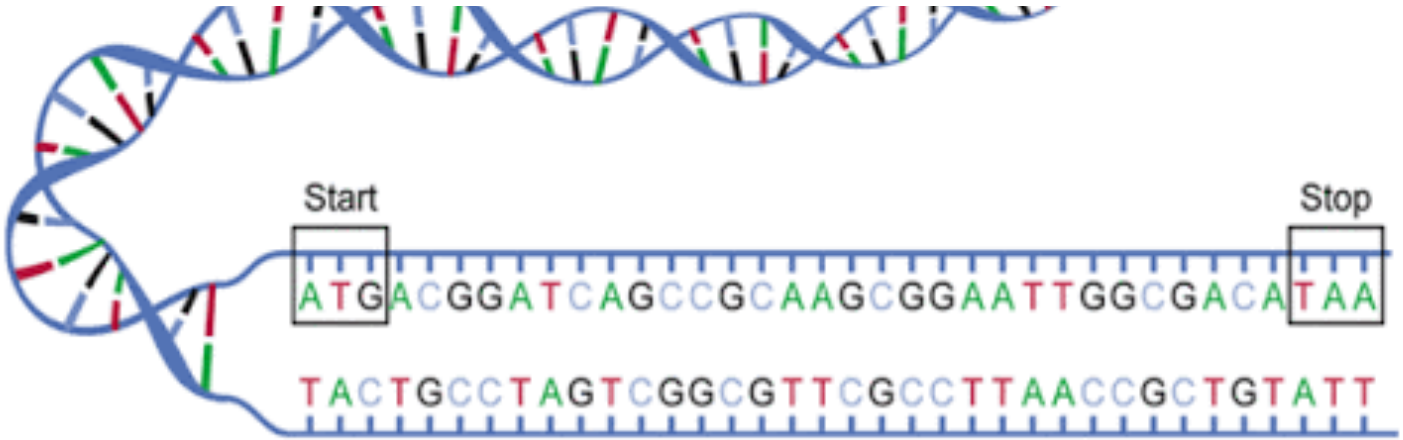
A bit of review: DNA -> RNA ->Protein

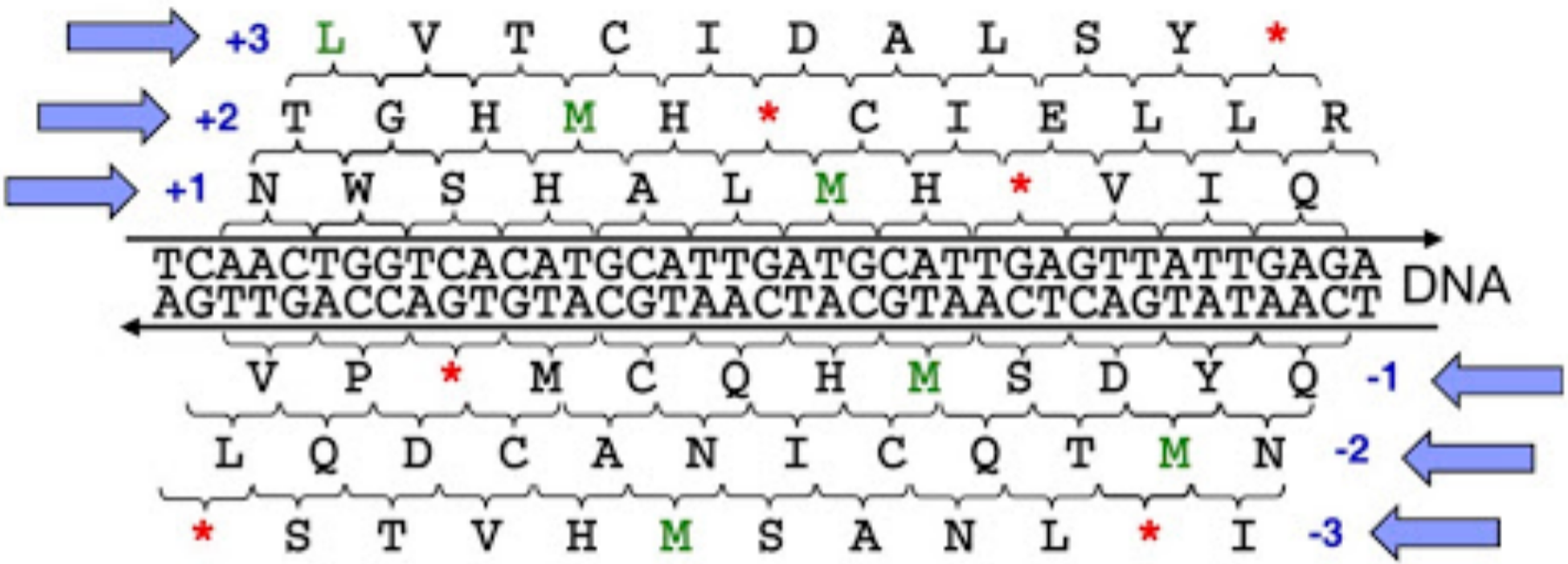


A bit of review: DNA -> RNA ->Protein



A bit of review: DNA -> RNA ->Protein





A bit of review: DNA -> RNA ->Protein

Possible Amino Acid Sequences (Forward)

```
R S R A F W S P M S A A D S S * K A A P F T N R A S N R Q P R T A K
 D L G R S G R R C R R P T H L E R L H R S R T G R R T G N R G R R
  I S G V L V A D V G G R L I L K G C T V H E P G V E P A T A D G E
```

Nucleotide Sequence

```
CGATCTCGGGCGTTCTGGTCGCCGATGTCGGCGGCCGACTCATCTTGAAAGGCTGCACCGTTCACGAACCGGGCGTCGAACCGGCAACCGCGGACGGCGA
....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|
GCTAGAGCCCGCAAGACCAGCGGCTACAGCCGCCGGCTGAGTAGAACTTTCCGACGTGGCAAGTGCTTGGCCCGCAGCTTGGCCGTTGGCGCCTGCCGCT
```

Possible Amino Acid Sequences (Reverse)

```
 R D R A N Q D G I D A A S E D Q F A A G N V F R A D F R C G R V A
  S R P R E P R R H R R G V * R S L S C R E R V P R R V P L R P R R
A I E P T R T A S T P P R S M K F P Q V T * S G P T S G A V A S P S
```

Gene 1

```
  S * S T K Q M W T T C R F P E R R C R * V A F V A S S G T V R G L
N R D R Q S K C G L H A D S L R G G V G K W L L S P R R E P F A G C
 I V I D K A N V D Y M Q I P * E A V S V S G F C R L V G N R S R V
AATCGTGATCGACAAAGCAAATGTGGACTACATGCAGATTCCCTGAGAGGCGGTGTCGGTAAGTGGCTTTTGTCGCCTCGTCGGGAACCGTTCGCGGGTT
....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|
TTAGCACTAGCTGTTTCGTTTACACCTGATGTACGTCTAAGGGACTCTCCGCCACAGCCATTCACCGAAAACAGCGGAGCAGCCCTTGGCAAGCGCCCAA
F D H D V F C I H V V H L N G S L R H R Y T A K T A E D P V T R P N
 F R S R C L L H P S C A S E R L P P T P L H S K D G R R S G N A P
  I T I S L A F T S * M C I G Q S A T D T L P K Q R R T P F R E R T
```

Gene 2

```
 S R P N I P A P Q R I A L S P W V S M E Y Y E I W R R Q P A V R R
  R A P T F Q R R N A S R C R R G C R W S I T R F G A V S P R C G A
V A P Q H S S A A T H R A V A V G V D G V L R D L A P S A R G A A R
GTCGCGCCCCAACATTCCAGCGCCGCAACGCATCGCGCTGTCGCCGTGGGTGTCGATGGAGTATTACGAGATTTGGCGCCGTCAGCCCGCGGTGCGGCGC
....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|
CAGCGCGGGGTTGTAAGGTCGCGGCGTTGCGTAGCGCGACAGCGGCACCCACAGCTACCTCATAATGCTCTAAACCGCGGCAGTCGGGCGCCACGCCGCG
  D R G L M G A G C R M A S D G H T D I S Y * S I Q R R * G A T R R
Q R A G V N W R R L A D R Q R R P H R H L I V L N P A T L G R H P A
 T A G W C E L A A V C R A T A T P T S P T N R S K A G D A R P A A

A V S F L A R N I A Q L G L H L F E R K D D A D R K R L T D H P L A
 R C R S W R A T S R N S V C T C S S A R T M P T A S G * P T T R S
  G V V P G A Q H R A T R S A P V R A Q G R C R P Q A V D R P P A R
GCGGTGTCGTTCCTGGCGCGCAACATCGCGCAACTCGGTCTGCACCTGTTCGAGCGCAAGGACGATGCCGACCGCAAGCGGTTGACCGACCACCCGCTCG
....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|
CGCCACAGCAAGGACCGCGCGTTGTAGCGCGTTGAGCCAGACGTGGACAAGCTCGCGTTCCTGCTACGGCTGGCGTTCGCCAACTGGCTGGTGGGCGAGC
 A T D N R A R L M A C S P R C R N S R L S S A S R L R N V S W G S
  R H R E Q R A V D R L E T Q V Q E L A L V I G V A L P Q G V V R E
R P T T G P A C C R A V R D A G T R A C P R H R G C A T S R G G A R
```

A bit of review: Translating Proteins

ExPASy - Translate tool

web.expasy.org/translate/

ExPASy Bioinformatics Resource Portal

Translate

Home | Contact

Programmatic access

Translate is a tool which allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence.

DNA or RNA sequence

Please enter a DNA or RNA sequence - numbers and blanks are ignored

Output format

- (•) Verbose: Met, Stop, spaces between residues
- ( ) Compact: M, -, no spaces
- ( ) Includes nucleotide sequence
- ( ) Includes nucleotide sequence, no spaces

DNA strands

[x] forward [x] reverse

Genetic codes - See NCBI's genetic codes

Standard

reset TRANSLATE!

BLAST – Programs in Detail

Program	Query	Number of database searches	Database
BLASTP	protein	1	protein

Use BLASTP to compare a protein query to a database of proteins.

BLASTN	DNA	1	DNA

Use BLASTN to compare both strands of a DNA query against a DNA database.

BLASTX	DNA	6	protein

BLASTX translates a DNA sequence into six protein sequences using all six possible reading frames, and then compares each of these proteins to a protein database.

TBLASTN	protein	6	DNA

TBLASTN is used to translate every DNA sequence in a database into six potential proteins, and then to compare your protein query against each of those translated proteins.

TBLASTX	DNA	36	DNA

TBLASTX is the most computationally intensive BLAST algorithm. It translates DNA from both a query and a database into six potential proteins, then performs 36 protein-protein database searches.

Query: FASTA format, accession number, paste sequence, or upload file

Database "Subject"

Algorithm

parameters

blastn | blastp | blastx | tblastn | tblastx

BLASTN programs search nucleotide databases using a nucleotide que

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query s

From

To

```
>NP_002102.4 huntingtin [Homo sapiens]
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPPQLPQPPPQAQPLLPQPQPP
PPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPEFQKLLGIAMELFLLCSDD
AESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAPRSLRAALWRFAELAHLVRPQKCRPYLVN
LLPCLTRTSKRPEESVQETLAAAVPKIMASFGNFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSIC
```

Or, upload file: Choose File No file chosen

Job Title: NP_002102.4 huntingtin [Homo sapiens]

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database: Human genomic + transcript | Mouse genomic + transcript | Others (nr etc.):

Nucleotide collection (nr/nt)

Organism (Optional): Enter organism name or id--completions will be suggested — exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude (Optional): Models (XM/XP) Uncultured/environmental sample sequences

Limit to (Optional): Sequences from type material

Entrez Query (Optional): YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for:

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm

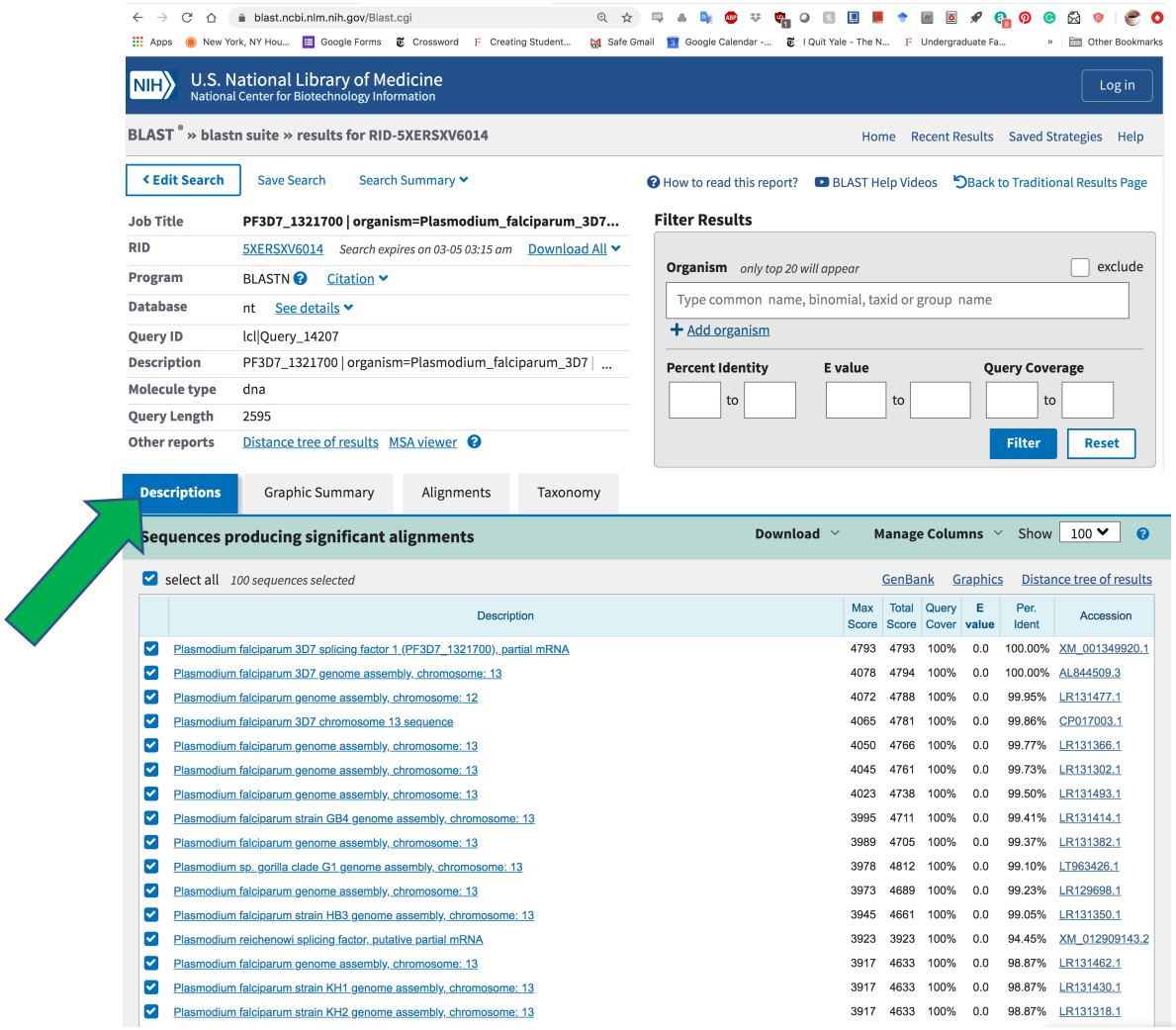
BLAST

Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequen

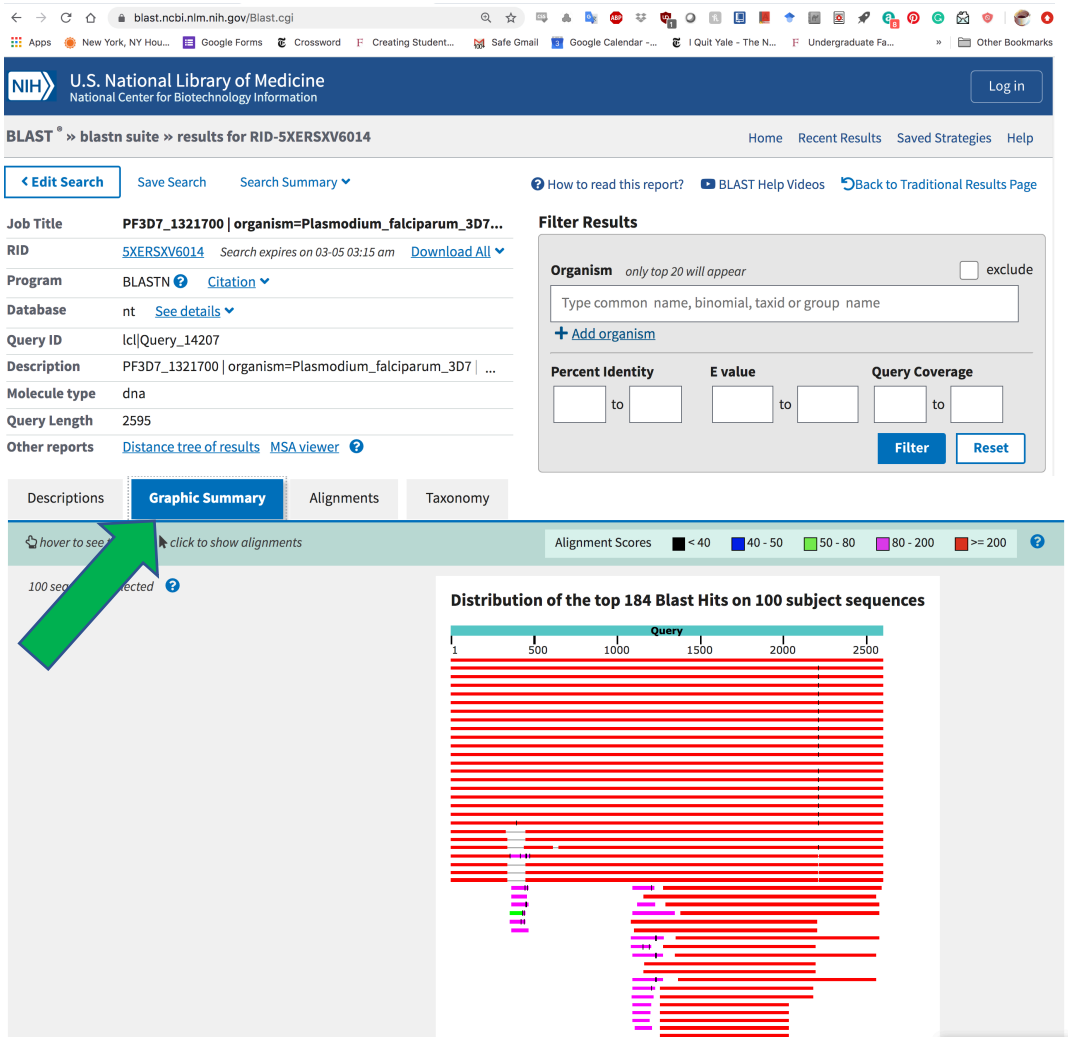
Show results in a new window

Algorithm parameters

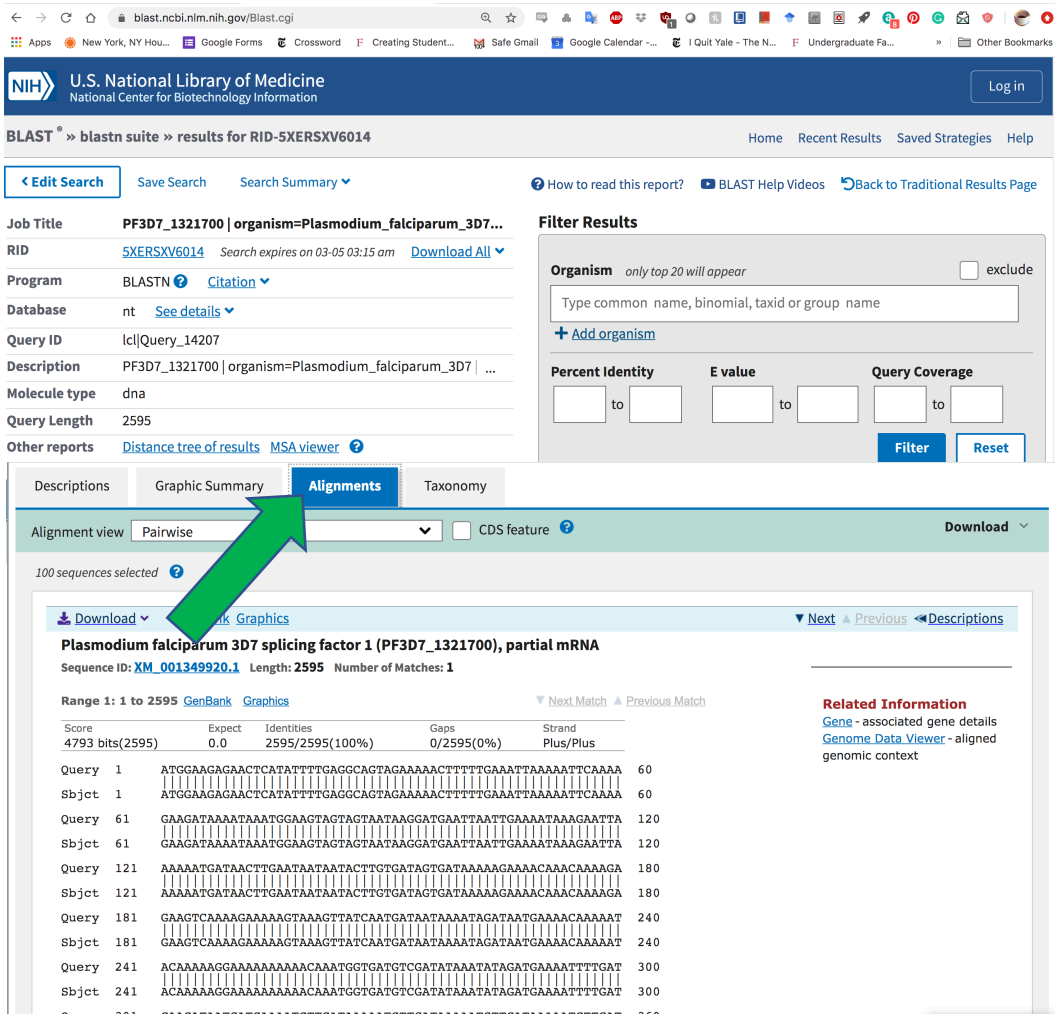
Web BLAST


blast.ncbi.nlm.nih.gov/Blast.cgi
U.S. National Library of Medicine
National Center for Biotechnology Information
Log in
BLAST » blastn suite » results for RID-5XERSXV6014
Home Recent Results Saved Strategies Help
Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page
Job Title PF3D7_1321700 | organism=Plasmodium_falciparum_3D7...
RID 5XERSXV6014 Search expires on 03-05 03:15 am Download All
Program BLASTN Citation
Database nt See details
Query ID lcl|Query_14207
Description PF3D7_1321700 | organism=Plasmodium_falciparum_3D7 | ...
Molecule type dna
Query Length 2595
Other reports Distance tree of results MSA viewer
Filter Results
Organism only top 20 will appear
exclude
Type common name, binomial, taxid or group name
Add organism
Percent Identity
E value
Query Coverage
Filter
Reset
Descriptions
Graphic Summary
Alignments
Taxonomy
Sequences producing significant alignments
Download
Manage Columns
Show 100
select all 100 sequences selected
GenBank Graphics Distance tree of results
Description Max Score Total Score Query Cover E value Per. Ident Accession
Plasmodium falciparum 3D7 splicing factor 1 (PF3D7_1321700), partial mRNA 4793 4793 100% 0.0 100.00% XM_001349920.1
Plasmodium falciparum 3D7 genome assembly, chromosome: 13 4078 4794 100% 0.0 100.00% AL844509.3
Plasmodium falciparum genome assembly, chromosome: 12 4072 4788 100% 0.0 99.95% LR131477.1
Plasmodium falciparum 3D7 chromosome 13 sequence 4065 4781 100% 0.0 99.86% CP017003.1
Plasmodium falciparum genome assembly, chromosome: 13 4050 4766 100% 0.0 99.77% LR131366.1
Plasmodium falciparum genome assembly, chromosome: 13 4045 4761 100% 0.0 99.73% LR131302.1
Plasmodium falciparum genome assembly, chromosome: 13 4023 4738 100% 0.0 99.50% LR131493.1
Plasmodium falciparum strain GB4 genome assembly, chromosome: 13 3995 4711 100% 0.0 99.41% LR131414.1
Plasmodium falciparum genome assembly, chromosome: 13 3989 4705 100% 0.0 99.37% LR131382.1
Plasmodium sp. gorilla clade G1 genome assembly, chromosome: 13 3978 4812 100% 0.0 99.10% LT963426.1
Plasmodium falciparum genome assembly, chromosome: 13 3973 4689 100% 0.0 99.23% LR129698.1
Plasmodium falciparum strain HB3 genome assembly, chromosome: 13 3945 4661 100% 0.0 99.05% LR131350.1
Plasmodium reichenowi splicing factor, putative partial mRNA 3923 3923 100% 0.0 94.45% XM_012909143.2
Plasmodium falciparum genome assembly, chromosome: 13 3917 4633 100% 0.0 98.87% LR131462.1
Plasmodium falciparum strain KH1 genome assembly, chromosome: 13 3917 4633 100% 0.0 98.87% LR131430.1
Plasmodium falciparum strain KH2 genome assembly, chromosome: 13 3917 4633 100% 0.0 98.87% LR131318.1

Web BLAST


blast.ncbi.nlm.nih.gov/Blast.cgi
U.S. National Library of Medicine
National Center for Biotechnology Information
Log in
BLAST » blastn suite » results for RID-5XERSXV6014
Home
Recent Results
Saved Strategies
Help
Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page
Job Title
PF3D7_1321700 | organism=Plasmodium_falciparum_3D7...
RID
5XERSXV6014
Search expires on 03-05 03:15 am
Download All
Program
BLASTN
Citation
Database
nt
See details
Query ID
lcl|Query_14207
Description
PF3D7_1321700 | organism=Plasmodium_falciparum_3D7
Molecule type
dna
Query Length
2595
Other reports
Distance tree of results
MSA viewer
Filter Results
Organism
only top 20 will appear
exclude
Type common name, binomial, taxid or group name
Add organism
Percent Identity
E value
Query Coverage
Filter
Reset
Descriptions
Graphic Summary
Alignments
Taxonomy
hover to see
click to show alignments
Alignment Scores
< 40
40 - 50
50 - 80
80 - 200
>= 200
Distribution of the top 184 Blast Hits on 100 subject sequences
Query
1
500
1000
1500
2000
2500

Web BLAST


blast.ncbi.nlm.nih.gov/Blast.cgi
U.S. National Library of Medicine
National Center for Biotechnology Information
Log in
BLAST » blastn suite » results for RID-5XERSXV6014
Home Recent Results Saved Strategies Help
Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page
Job Title PF3D7_1321700 | organism=Plasmodium_falciparum_3D7...
RID 5XERSXV6014 Search expires on 03-05 03:15 am Download All
Program BLASTN Citation
Database nt See details
Query ID lcl|Query_14207
Description PF3D7_1321700 | organism=Plasmodium_falciparum_3D7 | ...
Molecule type dna
Query Length 2595
Other reports Distance tree of results MSA viewer
Filter Results
Organism only top 20 will appear
exclude
Type common name, binomial, taxid or group name
Add organism
Percent Identity
E value
Query Coverage
Filter
Reset
Descriptions
Graphic Summary
Alignments
Taxonomy
Alignment view Pairwise
CDS feature
Download
100 sequences selected
Download
Graphics
Next
Previous
Descriptions
Plasmodium falciparum 3D7 splicing factor 1 (PF3D7_1321700), partial mRNA
Sequence ID: XM_001349920.1 Length: 2595 Number of Matches: 1
Range 1: 1 to 2595 GenBank Graphics
Next Match
Previous Match
Score 4793 bits(2595)
Expect 0.0
Identities 2595/2595(100%)
Gaps 0/2595(0%)
Strand Plus/Plus
Query 1 ATGGAAGAGAACTCATATTTTGAGGCAGTAGAAAAACTTTTTGAAATTAAAAATTCAAAA 60
Sbjct 1 ATGGAAGAGAACTCATATTTTGAGGCAGTAGAAAAACTTTTTGAAATTAAAAATTCAAAA 60
Query 61 GAAGATAAAATAAATGGAAGTAGTAGTAATAAGGATGAATTAATTGAAAATAAAGAATTA 120
Sbjct 61 GAAGATAAAATAAATGGAAGTAGTAGTAATAAGGATGAATTAATTGAAAATAAAGAATTA 120
Query 121 AAAAATGATAACTTGAATAATAATACTTGTGATAGTGATAAAAAGAAAACAAACAAAAGA 180
Sbjct 121 AAAAATGATAACTTGAATAATAATACTTGTGATAGTGATAAAAAGAAAACAAACAAAAGA 180
Query 181 GAAGTCAAAAGAAAAAGTAAAGTTATCAATGATAATAAAATAGATAATGAAAACAAAAAT 240
Sbjct 181 GAAGTCAAAAGAAAAAGTAAAGTTATCAATGATAATAAAATAGATAATGAAAACAAAAAT 240
Query 241 ACAAAAAGGAAAAAAAAAACAAATGGTGATGTCGATATAAATATAGATGAAAATTTTGAT 300
Sbjct 241 ACAAAAAGGAAAAAAAAAACAAATGGTGATGTCGATATAAATATAGATGAAAATTTTGAT 300
Related Information
Gene - associated gene details
Genome Data Viewer - aligned genomic context

BLAST Exercises

(1) Use all algorithms for nucleotide sequences and use BLAST to explore the database using each locus from the Nucleotide_sequences_for_BLAST_13Oct21.txt file. (tblastx doesn't seem to be working, so skip that one.)

(2) What is the likely function of each gene? (Unknown is acceptable where it applies)

(3) In the translated queries, how do you know what reading frame for the results?

(4) Repeat this exercise with the protein sequences, Protein_sequences_for_BLAST_13Oct21.txt

(5) Identify the sequence you used for the translation exercise