Bioinformatics

CS300

Blast, Substitution Matrices and Protein Alignments

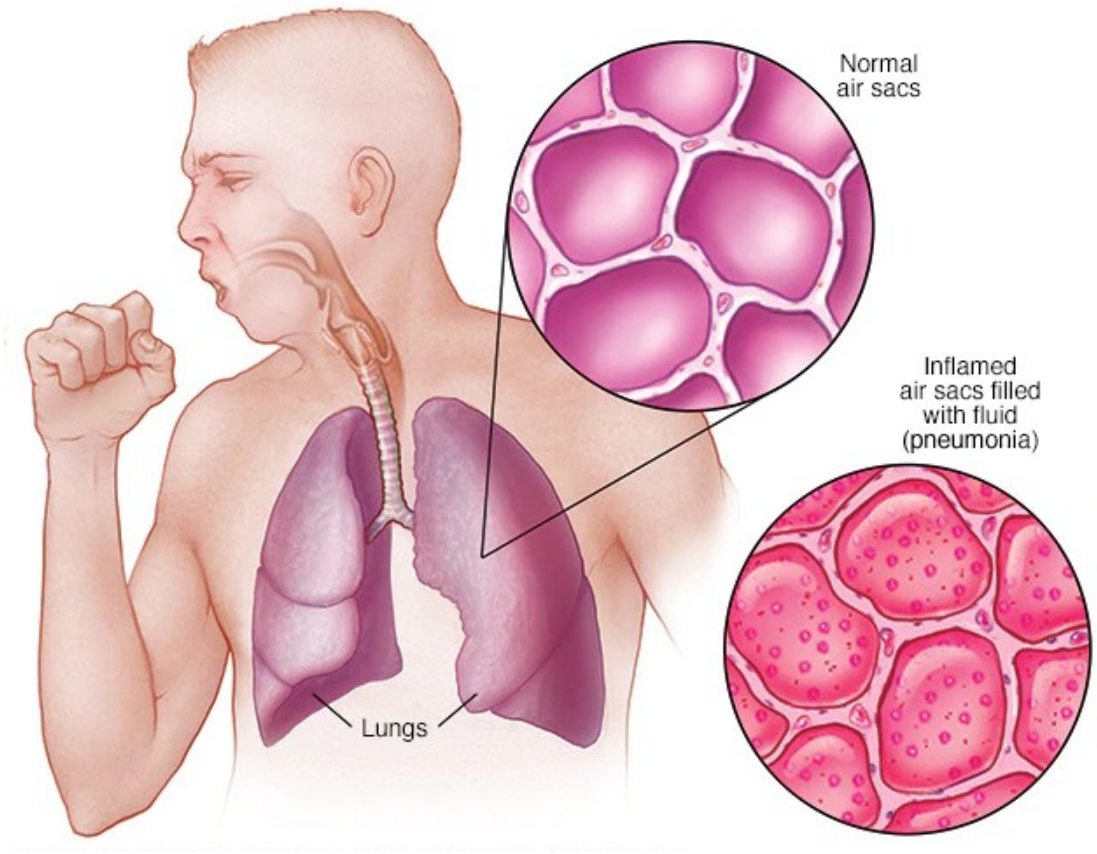
(Chap 4 and 5 in textbook)

Spring 2021

Oliver BONHAM-CARTER

Pneumonia

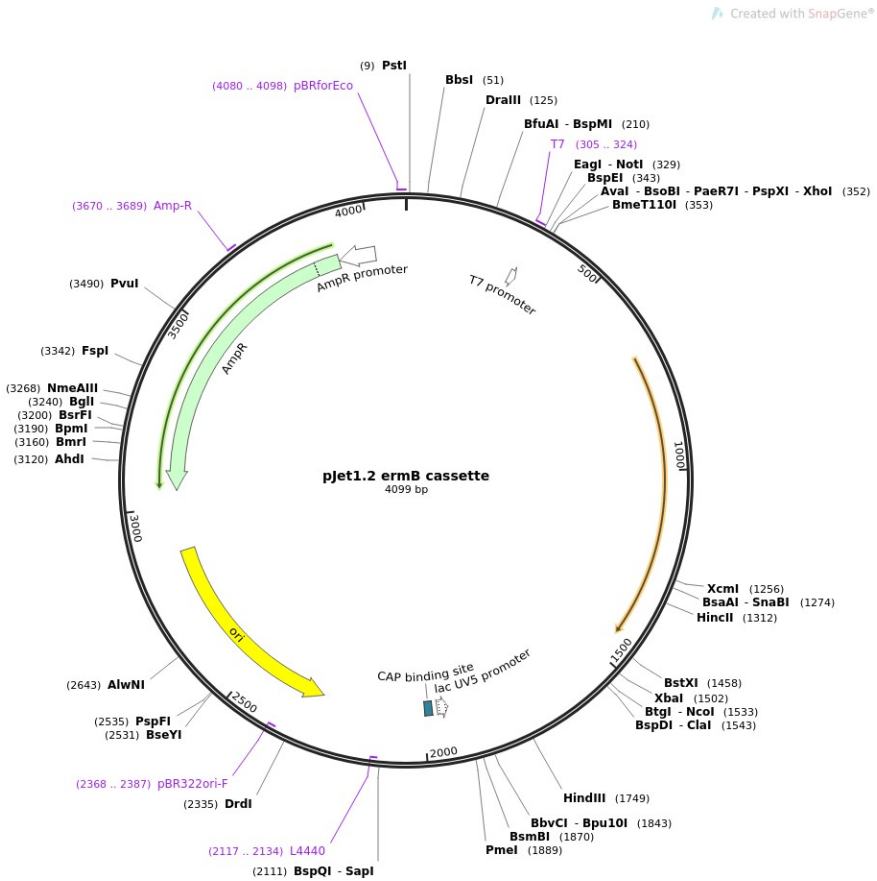
- Pneumonia is an infection that inflames the air sacs in one or both lungs. The air sacs may fill with fluid or pus (purulent material), causing cough with phlegm or pus, fever, chills, and difficulty breathing. A variety of organisms, including bacteria, viruses and fungi, can cause pneumonia.
- A classic sign of bacterial pneumonia is a cough that produces thick, blood-tinged or yellowish-greenish sputum with pus.



https://www.mayoclinic.org/diseases-conditions/pneumonia/symptoms-causes/syc-20354204

Human Pathogen Inquiry: The *ermB* gene

- An erythromycin-resistance gene from Streptococcus agalactiae, a gram-positive bacterial species commonly associated with the udders of cows, causing mastitis (i.e., inflammation of breast tissue that sometimes involves an infection and may cause fever)



https://www.addgene.org/117121/

Pneumonia and *ermB*

- Drug resistant: Erythromycin is a macrolide antibiotic used to treat bacterial infections
- Resistance is due to the *ermB* gene which has been noted in the bacteria, *Streptococcus pneumonia* – a common cause of bacterial **pneumonia**.

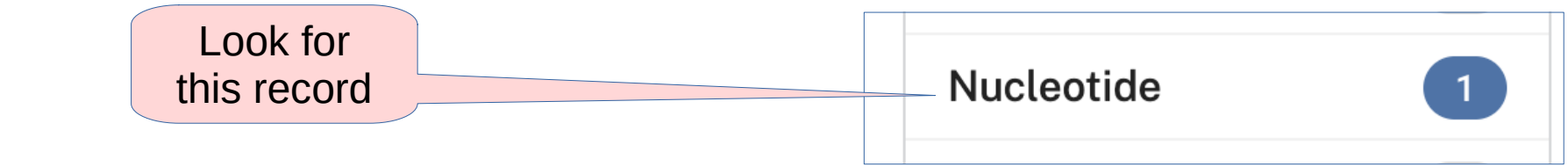
Horizontal Gene Transfer?

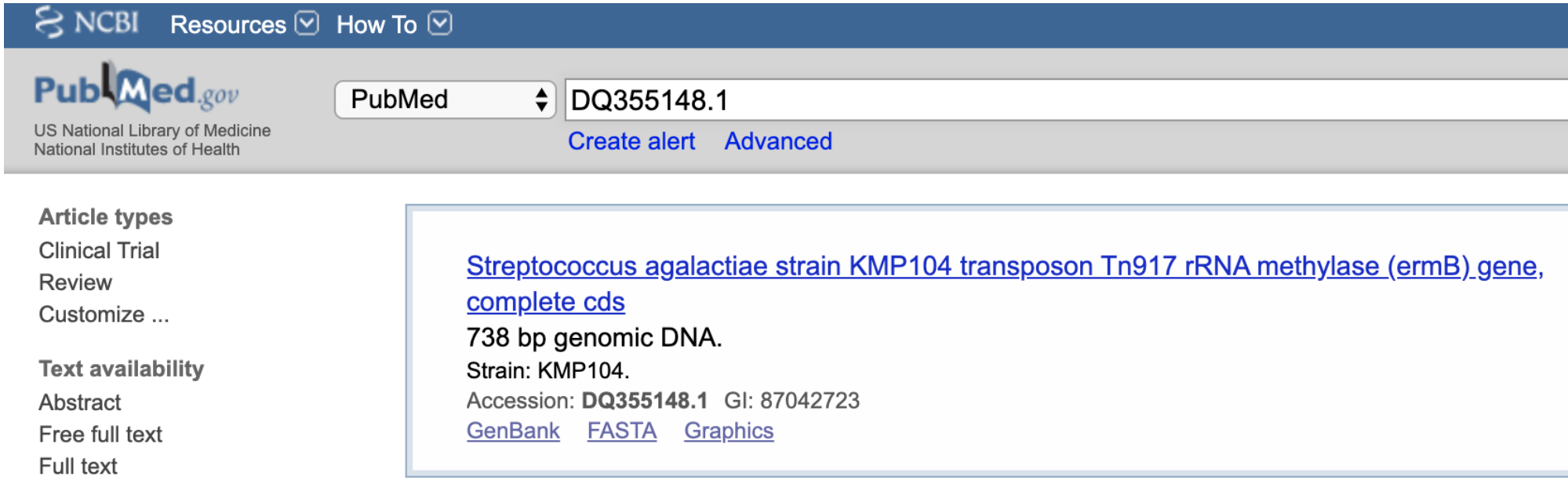
- This type of pneumonia is not believed to have always been resistant to drugs.
- Could the resistance gene have come from another bacteria via HGT?
- How could we check what other bacterial organisms have a specific allele for the gene that effectively resists drugs?
- We will use Blast for this task.

BLAST BLAST BLAST

Let's Study HGT

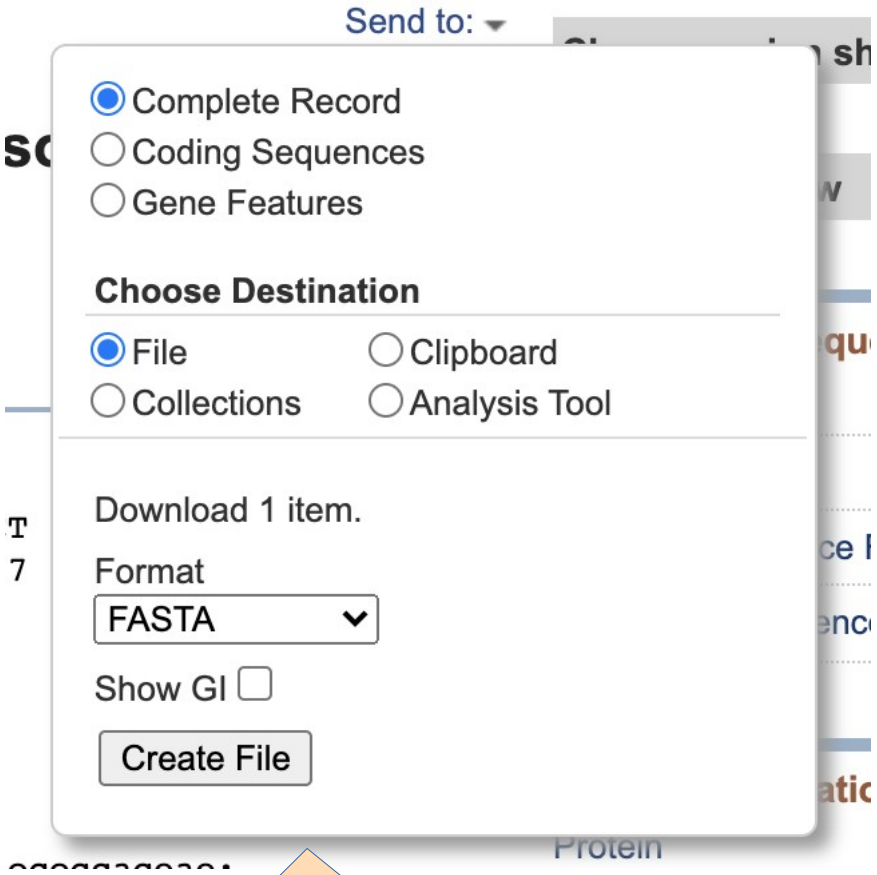
- Locate the Accession number, ***DQ355148.1***, on https://www.ncbi.nlm.nih.gov/
- *Streptococcus agalactiae* strain KMP104 transposon Tn917 rRNA methylase (ermB) gene, complete cds

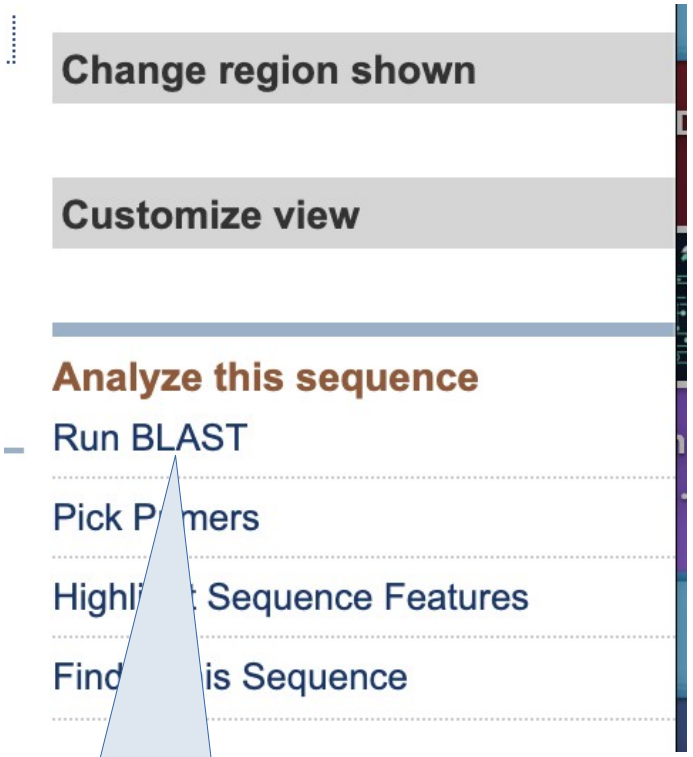




Quick link:
https://www.ncbi.nlm.nih.gov/search/all/?term=DQ355148.1

How to get the Data?





Method 1:
Get a text file of the gene to have the sequence or now and future work.

Method 2:
Locate a gene record on NCBI and click the Blast button.

Find the Nucleotide Sequence

GenBank ▾ Send to: ▾

Streptococcus agalactiae strain KMP104 transposon Tn917 rRNA methylase (ermB) gene, complete cds

GenBank: DQ355148.1

FASTA Graphics

Go to:

```
LOCUS       DQ355148                738 bp    DNA     linear   BCT 13-FEB-2006
DEFINITION  Streptococcus agalactiae strain KMP104 transposon Tn917 rRNA
            methylase (ermB) gene, complete cds.
ACCESSION   DQ355148
VERSION     DQ355148.1
KEYWORDS    .
SOURCE      Streptococcus agalactiae
  ORGANISM  Streptococcus agalactiae
            Bacteria; Firmicutes; Bacilli; Lactobacillales; Streptococcaceae;
            Streptococcus.
REFERENCE   1  (bases 1 to 738)
  AUTHORS   Puopolo,K.M., Klinzing,D.C., Lin,M.P., Yesucevitz,D.L. and
            Cieslewicz,M.J.
  TITLE     A Composite Transposon Responsible for ErmB-Mediated Erythromycin
            Resistance in Group B Streptococcus
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 738)
  AUTHORS   Puopolo,K.M., Klinzing,D.C., Lin,M.P., Yesucevitz,D.L. and
            Cieslewicz,M.J.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-JAN-2006) Channing Laboratory, Brigham and Women's
            Hospital, 181 Longwood Avenue, Boston, MA 02115, USA
```

Get the FASTA file: "send to" → "FASTA"

Save the Sequence

GenBank

Send to:

Streptococcus agalactiae strain KMP104 transposon Tn917 rRNA (ermB) gene, complete cds

GenBank: DQ355148.1

FASTA Graphics

Go to:

```
LOCUS       DQ355148                738 bp    DNA     linear   BCT 13-FEB-2006
DEFINITION  Streptococcus agalactiae strain KMP104 transposon Tn917 rRNA
            methylase (ermB) gene, complete cds.
ACCESSION   DQ355148
VERSION     DQ355148.1
KEYWORDS    .
SOURCE      Streptococcus agalactiae
  ORGANISM  Streptococcus agalactiae
            Bacteria; Firmicutes; Bacilli; Lactobacillales; Streptococcaceae;
            Streptococcus.
REFERENCE   1  (bases 1 to 738)
  AUTHORS   Puopolo,K.M., Klinzing,D.C., Lin,M.P., Yesucevitz,D.L. and
            Cieslewicz,M.J.
  TITLE     A Composite Transposon Responsible for ErmB-Mediated Erythromycin
            Resistance in Group B Streptococcus
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 738)
  AUTHORS   Puopolo,K.M., Klinzing,D.C., Lin,M.P., Yesucevitz,D.L. and
            Cieslewicz,M.J.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-JAN-2006) Channing Laboratory, Brigham and Women's
            Hospital, 181 Longwood Avenue, Boston, MA 02115, USA
```

Complete Record
Coding Sequences
Gene Features

Choose Destination

File Clipboard
Collections Analysis Tool

Download 1 item.

Format: FASTA

Show GI

Create File

Protein

Taxonomy

PubMed (Weighte

Recent activity

Streptococcus transposon Tr

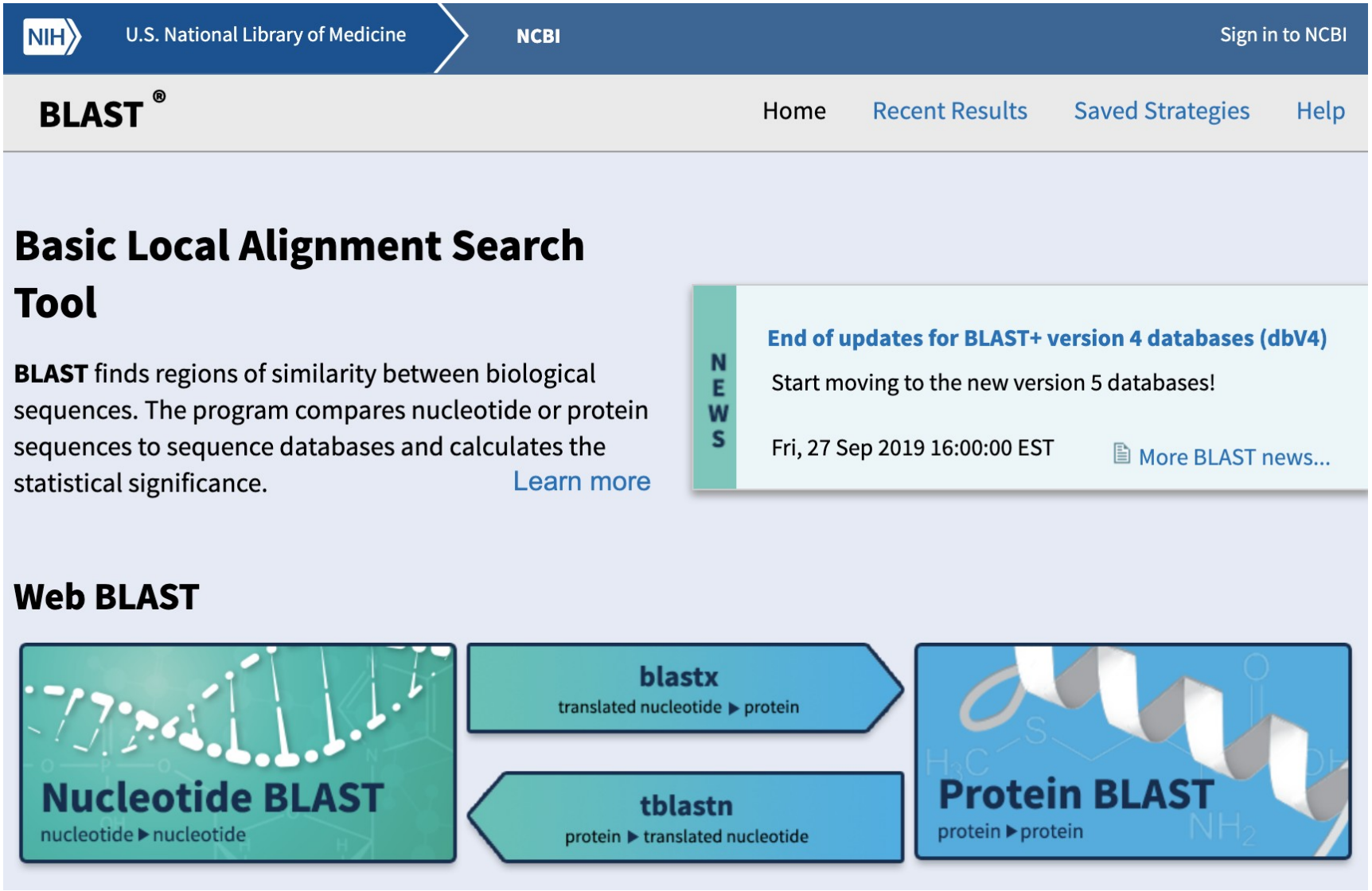
DQ355148.1

Ah, The Sequence in FASTA Format

>DQ355148.1 Streptococcus agalactiae strain KMP104 transposon Tn917 rRNA methylase (ermB) gene, complete cds

ATGAACAAAAATATAAAATATTCTCAAAACTTTTTAACGAGTGAAAAAGTACTCAACCAAATAATAAAAC
AATTGAATTTAAAAGAAACCGATACCGTTTACGAAATTGGAACAGGTAAAGGGCATTTAACGACGAAACT
GGCTAAAATAAGTAAACAGGTAACGTCTATTGAATTAGACAGTCATCTATTCAACTTATCGTCAGAAAAA
TTAAAACTGAACATTCGTGTCACTTTAATTCACCAAGATATTCTACAGTTTCAATTCCCTAACAAACAGA
GGTATAAAATTGTTGGGAATATTCCTTACCATTTAAGCACACAAATTATTAAAAAAGTGGTTTTTGAAAG
CCATGCGTCTGACATCTATCTGATTGTTGAAGAAGGATTCTACAAGCGTACCTTGGATATTCACCGAACA
CTAGGGTTGCTCTTGCACACTCAAGTCTCGATTCAGCAATTGCTTAAGCTGCCAGCGGAATGCTTTCATC
CTAAACCAAAAGTAAACAGTGTCTTAATAAAACTTACCCGCCATACCACAGATGTTCCAGATAAATATTG
GAAGCTATATACGTACTTTGTTTCAAAATGGGTCAATCGAGAATATCGTCAACTGTTTACTAAAAATCAG
TTTCATCAAGCAATGAAACACGCCAAAGTAAACAATTTAAGTACCGTTACTTATGAGCAAGTATTGTCTA
TTTTTAATAGTTATCTATTATTTAACGGGAGGAAATAA

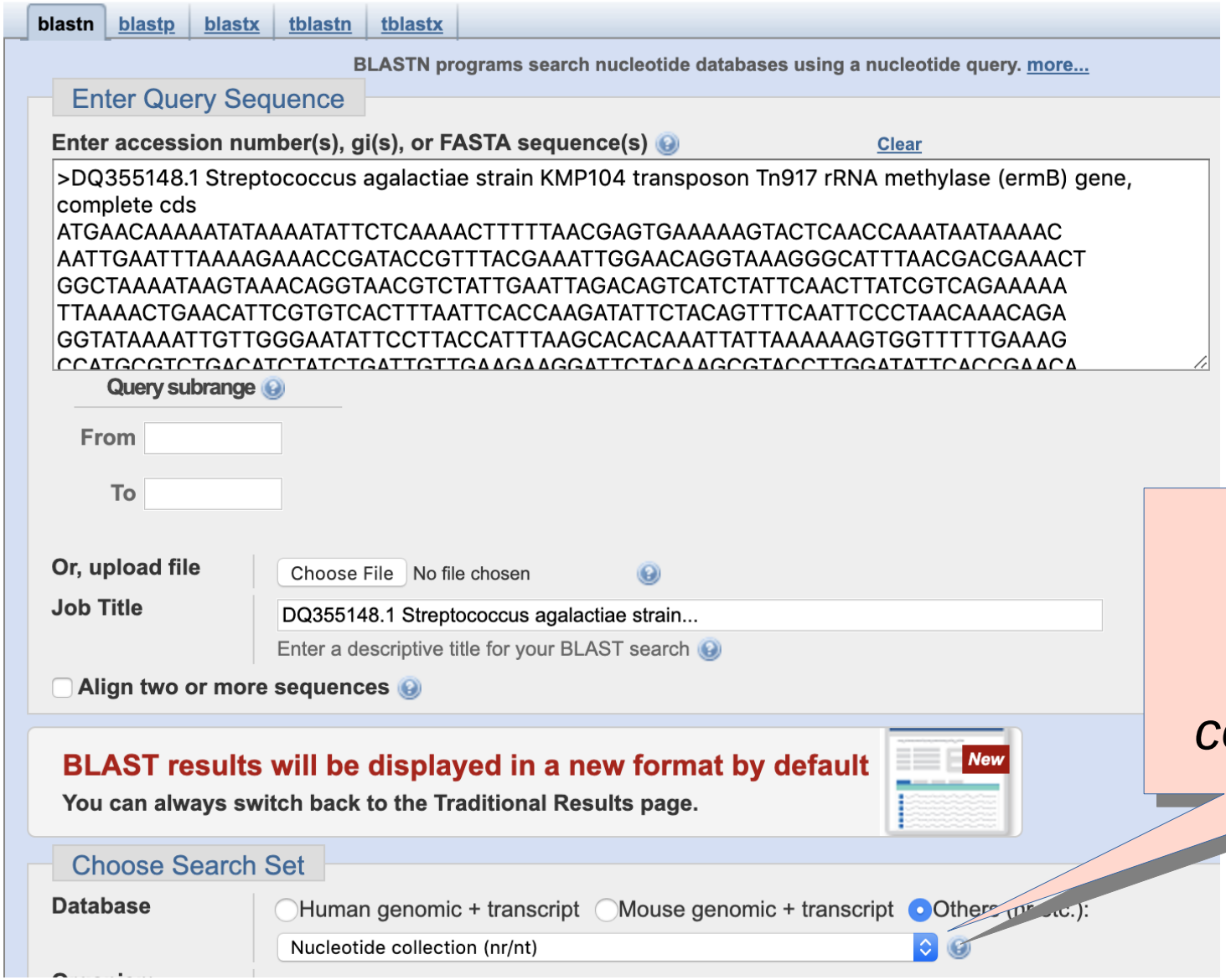
Blast Website



- https://blast.ncbi.nlm.nih.gov/Blast.cgi

Run The Query

Standard Nucleotide BLAST



Use database: *Nucleotide collection (nr/nt)*

Results

ALLEGHENY COLLEGE

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
Staphylococcus aureus strain VGC1 chromosome, complete genome	1363	1363	100%	0.0	100.00%	CP039448.1
Enterococcus durans strain VREdu plasmid pSULI, complete sequence	1363	1363	100%	0.0	100.00%	CP043327.1
Enterococcus durans strain VREdu chromosome	1363	1363	100%	0.0	100.00%	CP042597.1
Enterococcus faecalis EnGen0107 strain B594 plasmid p2, complete sequence	1363	1363	100%	0.0	100.00%	CP041740.1
Enterococcus faecalis strain 4928STDY7071263 genome assembly, chromosome: 1	1363	1363	100%	0.0	100.00%	LR607346.1
Enterococcus faecium strain N56454 plasmid unnamed, complete sequence	1363	1363	100%	0.0	100.00%	CP040905.1
Enterococcus avium strain 352 plasmid unnamed, complete sequence	1363	1363	100%	0.0	100.00%	CP034168.1
Listeria monocytogenes hypothetical protein, IS1216 transposase, 3-aminoglycoside o-phosp	1363	1363	100%	0.0	100.00%	MK490828.1
Enterococcus faecium isolate E8407 genome assembly, plasmid: 2	1363	1363	100%	0.0	100.00%	LR536659.1
Enterococcus faecium SMVRE20 plasmid pSMVRE20S DNA, complete genome	1363	1363	100%	0.0	100.00%	AP019410.1
Enterococcus faecium strain 37BA plasmid pEf37BA, complete sequence	1363	1363	100%	0.0	100.00%	MG957432.1
Enterococcus faecium strain FSIS1608820 plasmid pFSIS1608820, complete sequence	1363	2668	100%	0.0	100.00%	CP028728.1
Streptococcus pneumoniae isolate GPS_HK_21-sc-2296565 genome assembly, chromosome	1363	1363	100%	0.0	100.00%	LR216058.1
Synthetic construct clone pEP1237, complete sequence	1363	1363	100%	0.0	100.00%	MH626525.1

Scores

- **Max Score**
  - The score of the best matching segment for local alignment, not global
- **Total Score**
  - The total scores of all matching segments found (same as max score if there is only one matching segment)
- **Query Coverage**
  - The percentage of the query sequence that aligned to some part of the match.
- **E-Value**
  - A statistical measure evaluating how likely it is that a match this good could occur by chance. Lower e-scores indicate that both sequences are truly similar and are not similar by chance alone. Identical sequences have e-scores of zero.
- **Max Indent**
  - The percentage of nucleotides that are identical between the query and the target sequences within the matching regions.

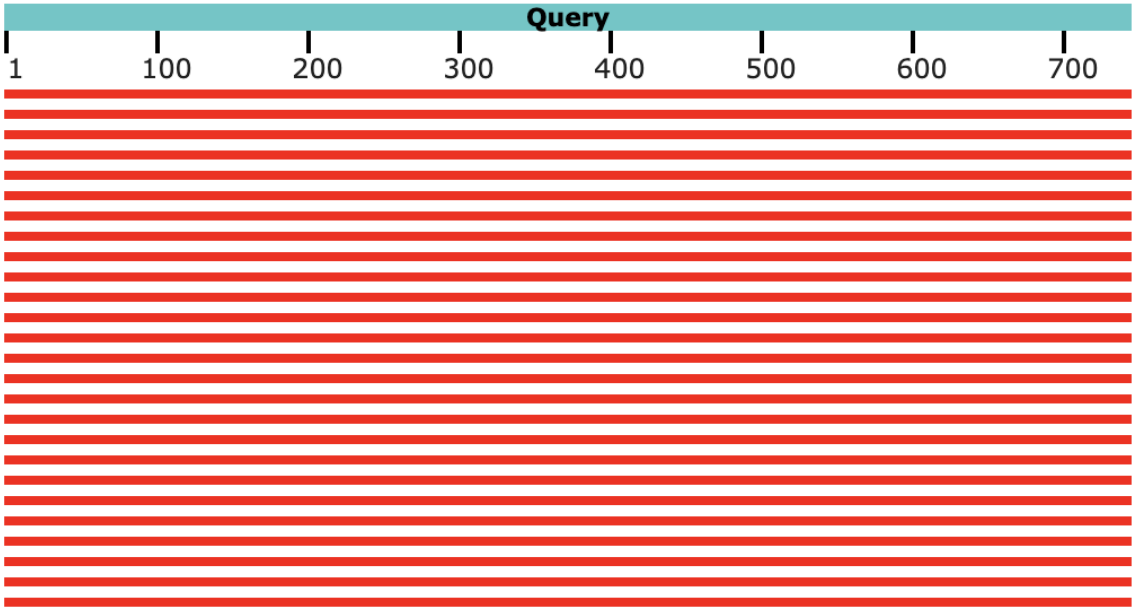
Results

Descriptions | **Graphic Summary** | Alignments | Taxonomy

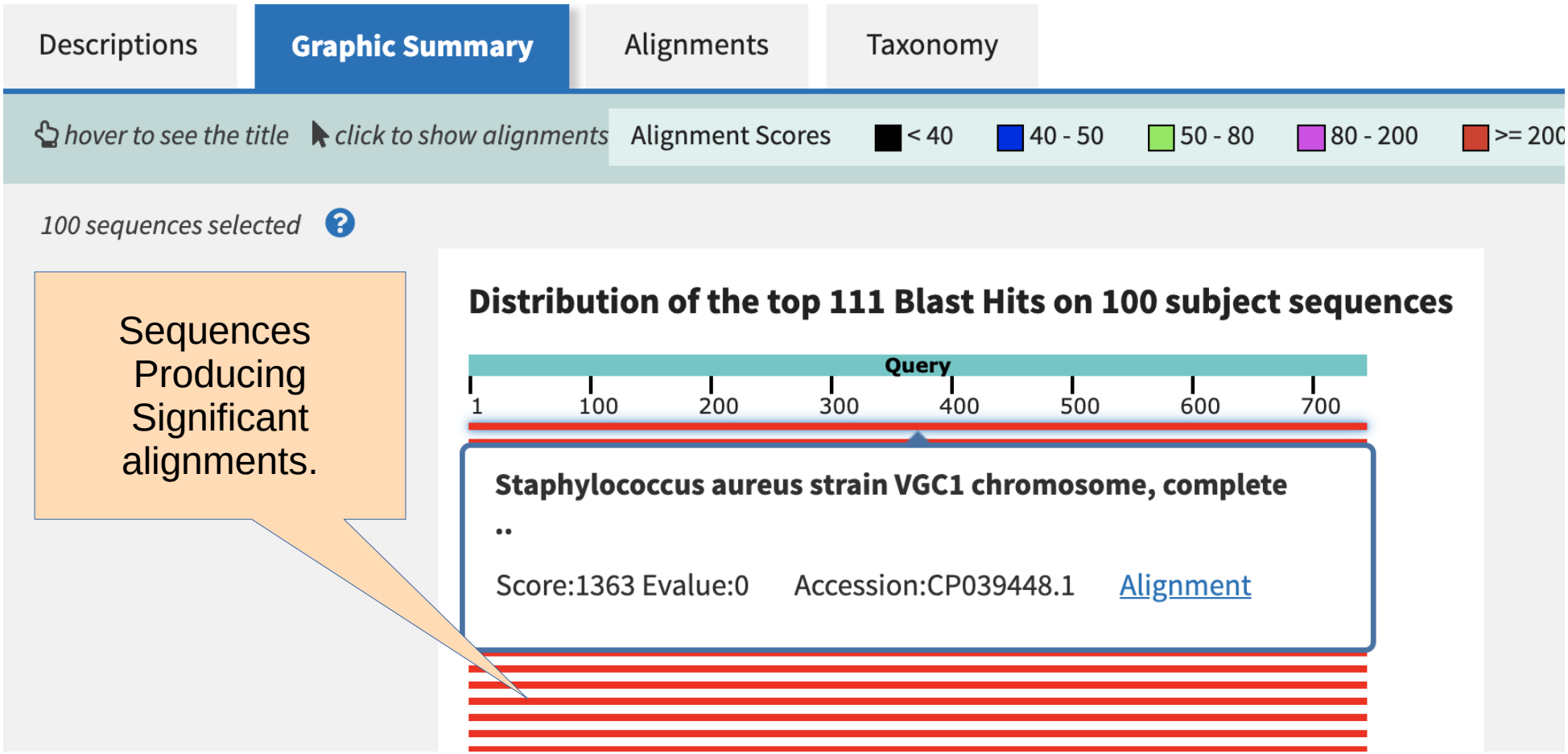
hover to see the title *click to show alignments* Alignment Scores < 40 | 40 - 50 | 50 - 80 | 80 - 200 | >= 200

100 sequences selected

Distribution of the top 111 Blast Hits on 100 subject sequences



Results


Descriptions
Graphic Summary
Alignments
Taxonomy
hover to see the title
click to show alignments
Alignment Scores
< 40
40 - 50
50 - 80
80 - 200
>= 200
100 sequences selected
Sequences Producing Significant alignments.
Distribution of the top 111 Blast Hits on 100 subject sequences
Query
1
100
200
300
400
500
600
700
Staphylococcus aureus strain VGC1 chromosome, complete
..
Score:1363 Evalue:0
Accession:CP039448.1
Alignment

Results

Streptococcus suis strain SC216 ICESsuSC216 sequence

Sequence ID: MK359991.1 **Length:** 54396 **Number of Matches:** 2

Range 1: 15998 to 16451 GenBank Graphics ▼ Next Match ▲

Score	Expect	Identities	Gaps	Strand
839 bits(454)	0.0	454/454(100%)	0/454(0%)	Plus/Plus

```
Query  1      AACAGGTAACGTCTATTGAATTAGACAGTCATCTATTCAACTTATCGTCAGAAAAA
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  15998  AACGTCTATTGAATTAGACAGTCATCTATTCAACTTATCGTCAGAAAAA

              AACTGAATACTCGTGTCACTTTAATTCACCAAGATATTCTACAGTTTCAATTCCCT
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16058  AACTGAATACTCGTGTCACTTTAATTCACCAAGATATTCTACAGTTTCAATTCCCT

              AACAGAGGTATAAAATTGTTGGGAATATTCCTTACCATTTAAGCACACAAATTATT
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16118  AACAGAGGTATAAAATTGTTGGGAATATTCCTTACCATTTAAGCACACAAATTATT

Query  181    AAGTGGTTTTTGAAAGCCGTGCGTCTGACATCTATCTGATTGTTGAAGAAGGATTC
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16178  AAGTGGTTTTTGAAAGCCGTGCGTCTGACATCTATCTGATTGTTGAAGAAGGATTC

Query  241    AGCGTACCTTGGATATTCACCGAACACTAGGGTTGCTCTTGCACACTCAAGTCTCC
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16238  AGCGTACCTTGGATATTCACCGAACACTAGGGTTGCTCTTGCACACTCAAGTCTCC

Query  301    AGCAATTGCTTAAGCTGCCAGCGGAATGCTTTCATCCTAAACCAAAAGTAAACAGT
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16298  AGCAATTGCTTAAGCTGCCAGCGGAATGCTTTCATCCTAAACCAAAAGTAAACAGT
```

An Identical sequence in another's genome

Conclusions on HGT?

- Typically, researchers allow for a 95% similarity between genes found between *unrelated* organisms.
- Here, **we may conclude that HGT is a good hypothesis** but more research must be done to determine whether there was a chance for two organisms to be close enough to each other to share genetic material.

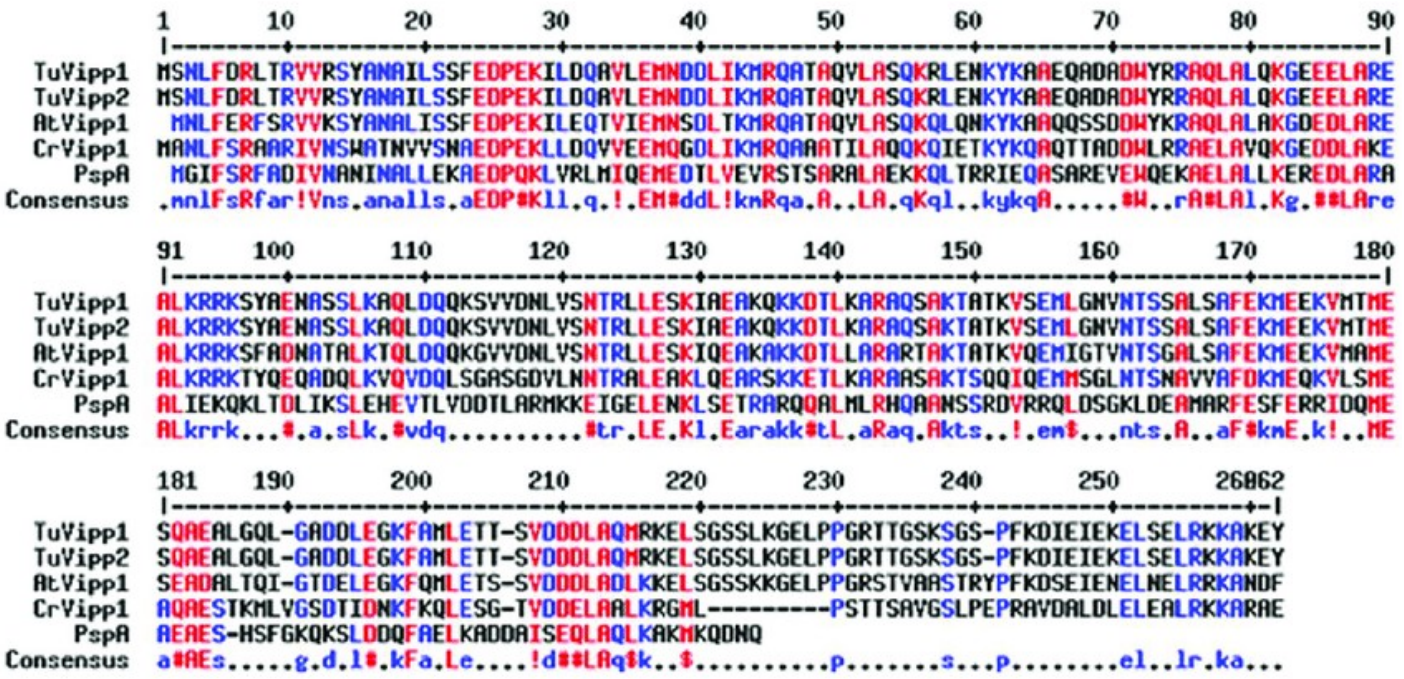
Your Turn to Investigate!!!

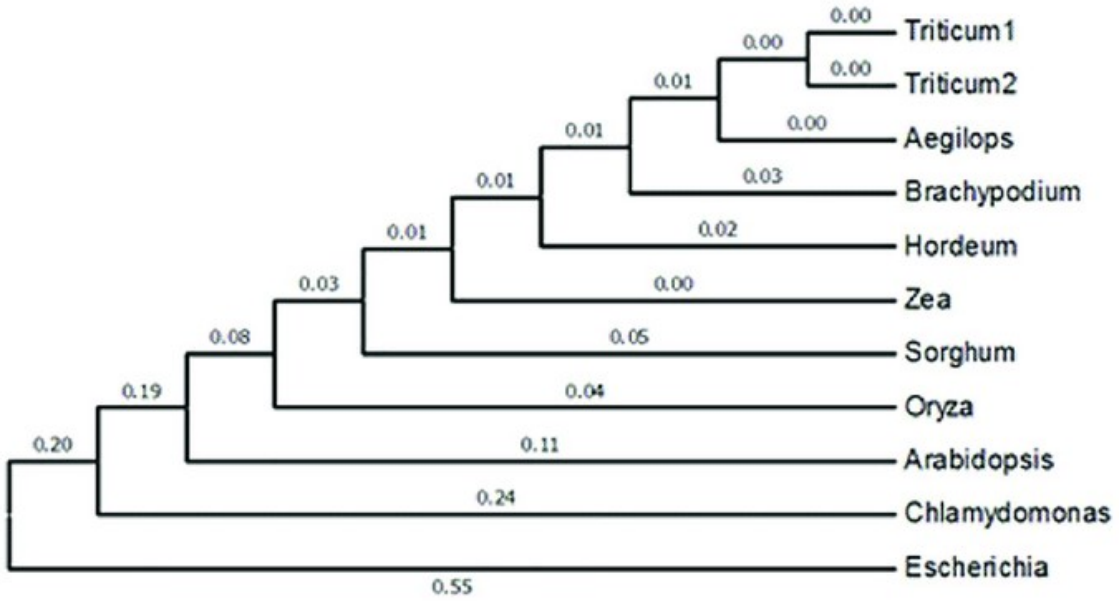
- Investigate a gene of resistance: *ermA* (Accession number: LT549456)
- Questions:
  - What is the description of this gene? (hint: see Genbank record)
  - About how many other organisms appear to have traces of the same gene sequence?
  - What is the closest match? Which organism? What *e*-score? Conclusions?

Activity 07:

https://forms.gle/mE55miv68ShnsmPE8

Blast Also Works With Proteins!!





Gao, Fei, et al. "Two novel vesicle-inducing proteins in plastids 1 genes cloned and characterized in Triticum urartu." PloS one 12.1 (2017): e0170439.

Proteins Can Also Be Blasted

A difference: results may not have been experimentally observed, DNA can be translated to produce this protein.

The reading frame of the DNA might produce a different protein than this one

Download GenPept Graphics

PREDICTED: serine/threonine-protein kinase PINK1, mitochondria

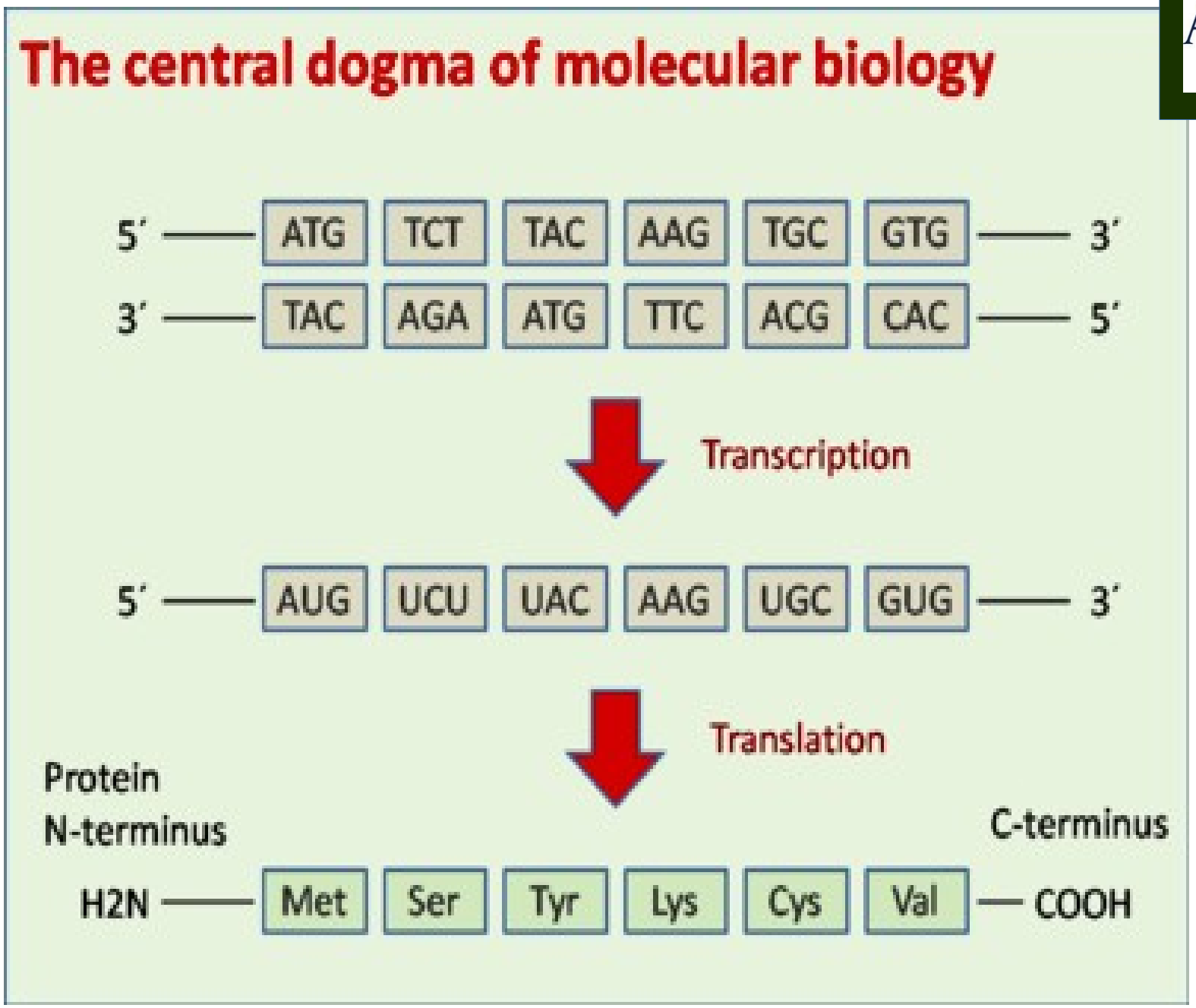
Sequence ID: XP_014893419.1 **Length:** 575 **Number of Matches:** 1

Range 1: 1 to 334 GenPept Graphics

Score	Expect	Method	Identities	Positives
653 bits(1684)	0.0	Compositional matrix adjust.	319/334(96%)	328/334

```
Query  1    MSVKHAISRGLELGRSFLQIGLLKSGGRVAAKLRADRFRVGPSVRTV
            MSVKHAISRGLELGRSFLQIGLLKSGGRVAAKLRADRFRVGPSVRTV
Sbjct  1    MSVKHAISRGLELGRSFLQIGLLKSGGRVAAKLRADRFRVGPSVRTV

Query  61   RTSLRGLAAQLQSAGFRRRFTGASPRNRAVFLAFGLGVGLIEQQLEE
            RTSL+GLAAQLQSAGFRRRFTGASPRNRAVFLAFGLGVGLIEQQLE+
Sbjct  61   RTSLKGLAAQLQSAGFRRRFTGASPRNRAVFLAFGLGVGLIEQQLED

Query  121  VFKKKKIQSTLRPFTSGFKLEDYVIGNQIGKGSNAAVYEAAAQFAHE
            VFKKKKIQSTLRPFTSGFKLEDYVIGNQIGKGSNAAVYEAAAQF+HE
Sbjct  121  VFKKKKIQSTLRPFTSGFKLEDYVIGNQIGKGSNAAVYEAAAQFSHE

Query  181  DNEVEVQNVRSAACCSLRNFPLAIKMLWNFGAGSSSEAILKSMSQEI
            DNE EVQNVRS +CCSLRNFPLAIKMLWNFGAGSSSEAILKSMSQEI
Sbjct  181  DNEEEVQNVRSPSCCSLRNFPLAIKMLWNFGAGSSSEAILKSMSQEI

Query  241  HITLDGHFGVLPKRVSAHPNVIRVYRAFTADVPLLPGAEEEYPDVLF
             ITLDG FGVLP+RVSAHPNVIRVYRAFTADVPLLPGA+EEYPDVLF
Sbjct  241  QITLDGRFGVLPRRVSAHPNVIRVYRAFTADVPLLPGAQEEYPDVLF

Query  301  LFLVMKNYPYTLRQYLQVSTPNRRQGSLMVLQLL   334
            LFLVMKNYP TLRQYLQVSTPNRRQGSLMVLQLL
Sbjct  301  LFLVMKNYPCTLRQYLQVSTPNRRQGSLMVLQLL   334
```

The central dogma of molecular biology

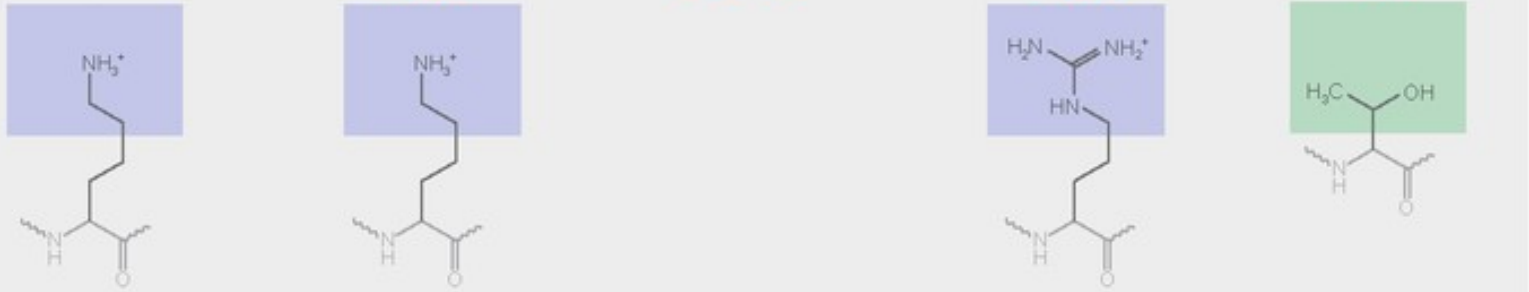


https://en.wikipedia.org/wiki/DNA_codon_table

More About Silent Mutations

- Redundant codons mean ~1/3 of DNA mutations often do not alter protein sequence

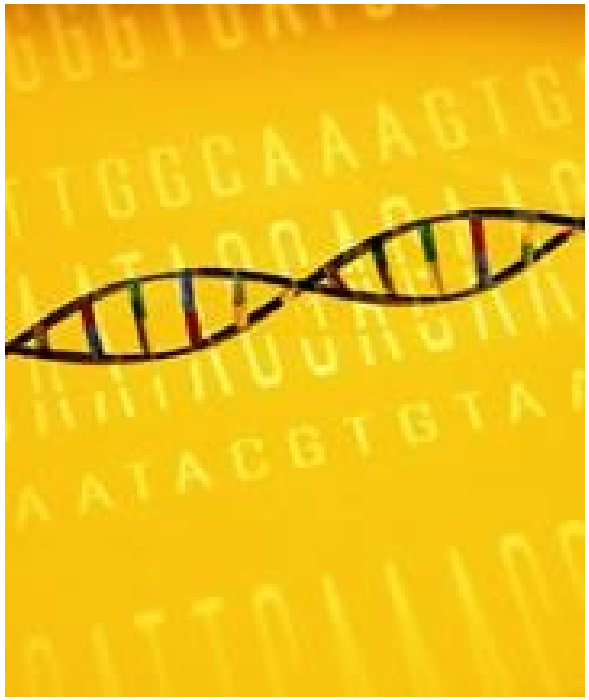
	No mutation	Point mutations			
		Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	Lys	Lys	STOP	Arg	Thr



https://en.wikipedia.org/wiki/Silent_mutation

Silent Mutations

- Are these mutations really so subtle?
- Are there dangers involved?
  - While the protein may be “fine,” the RNA may still have dangerous folding issues
- Nature: *Silent Mutations Speak Up: Overlooked genetic changes could impact on disease*
  - http://www.nature.com/news/2006/061221/full/news061218-12.html



nature International weekly journal of science

Is there a database to keep track of these silent mutations in genes?!

- **Article;** *dbDSM: a manually curated database for deleterious synonymous mutations*
  - *https://academic.oup.com/bioinformatics/article/32/12/1914/1744313*
- **DataBase Tool**
  - http://bioinfo.ahu.edu.cn:8080/dbDSM/index.jsp

Bring the Tool!

Up Next!

Quick link to search:
http://bioinfo.ahu.edu.cn:8080/dbDSM/search.jsp

Results:

Which genes have been altered by silent mutations?

10 records per page

Disease	Gene	SNPID	GRCh38_Position	c.DNA	Protein
Cystic fibrosis	CFTR	rs397508733	7:117531114	c.489G>A	p.Lys163=
Cystic fibrosis	CFTR	rs397508419	7:117603553	c.2679G>T	p.Gly893=
Cystic fibrosis	CFTR	rs397508419	7:117603553	c.2679G>T	p.Gly893
Cystic fibrosis	STX1A	rs2228607	7:73708593	c.204T>C	p.Asp68=

Genes that have been affected by silent mutations

Results:

Where is substitution in the gene sequence? (which base position)?

Gene	SNPID	GRCh38_Position	c.DNA	Protein	dbDSMscore	dbDSM_AccNum
CFTR	rs397508733	7:117531114	c.489G>A	p.Lys163=	5	DSM000732

Which gene is affected?

Which base has substitution?

The silent mutation substitution

Click to search NCBI for protein: NM_000492.3

dbDSMAccNum	DSM000732
Disease	Cystic fibrosis
DOID	DOID:1485
Gene	CFTR
GeneID	1080
MIM	MIM:602421
Map_Location	7q31.2
VariantType	germline
Protein	p.Lys163=
c.DNA	c.489G>A
SNPID	rs397508733
CodonChange	aaG/aaA
RefseqTranscript	NM_000492.3:c.489G>A,NP_000483.3:p.Lys163=

Results:

NCBI record for the affected gene

```
  1 aattggaagc aaatgacatc acagcaggtc agagaaaaag ggttgagcgg caggcaccca
 61 gagtagtagg tctttggcat taggagcttg agcccagacg gccctagcag ggaccccagc
121 gcccgagaga ccatgcagag gtcgcctctg gaaaaggcca gcgttgtctc caaacttttt
181 ttcagctgga ccagaccaat tttgaggaaa ggatacagac agcgcctgga attgtcagac
241 atataccaaa tcccttctgt tgattctgct gacaatctat ctgaaaaatt ggaaagagaa
301 tgggatagag agctggcttc aaagaaaaat cctaaactca ttaatgccct tcggcgatgt
361 tttttctgga gatttatgtt ctatggaatc tttttatatt taggggaagt caccaaagca
421 gtacagcctc tcttactggg aagaatcata gcttcctatg acccggataa caaggaggaa
481 cgctctatcg cgatttatct aggcataggc ttatgccttc tctttattgt gaggacactg
541 ctcctacagc cagccatttt tggccttcat cacattggaa tgcagatgag aatagctatg
601 tttagttta tttataagaa gactttaaag ctgtcaagcc gtgttctaga taaaataagt
661 attggacac ttgttagtct cctttccaac aacctgaaca aatttgatga aggacttgca
```

Gene code from NCBI record, we find;

Position: 489
Substitution: G -> A

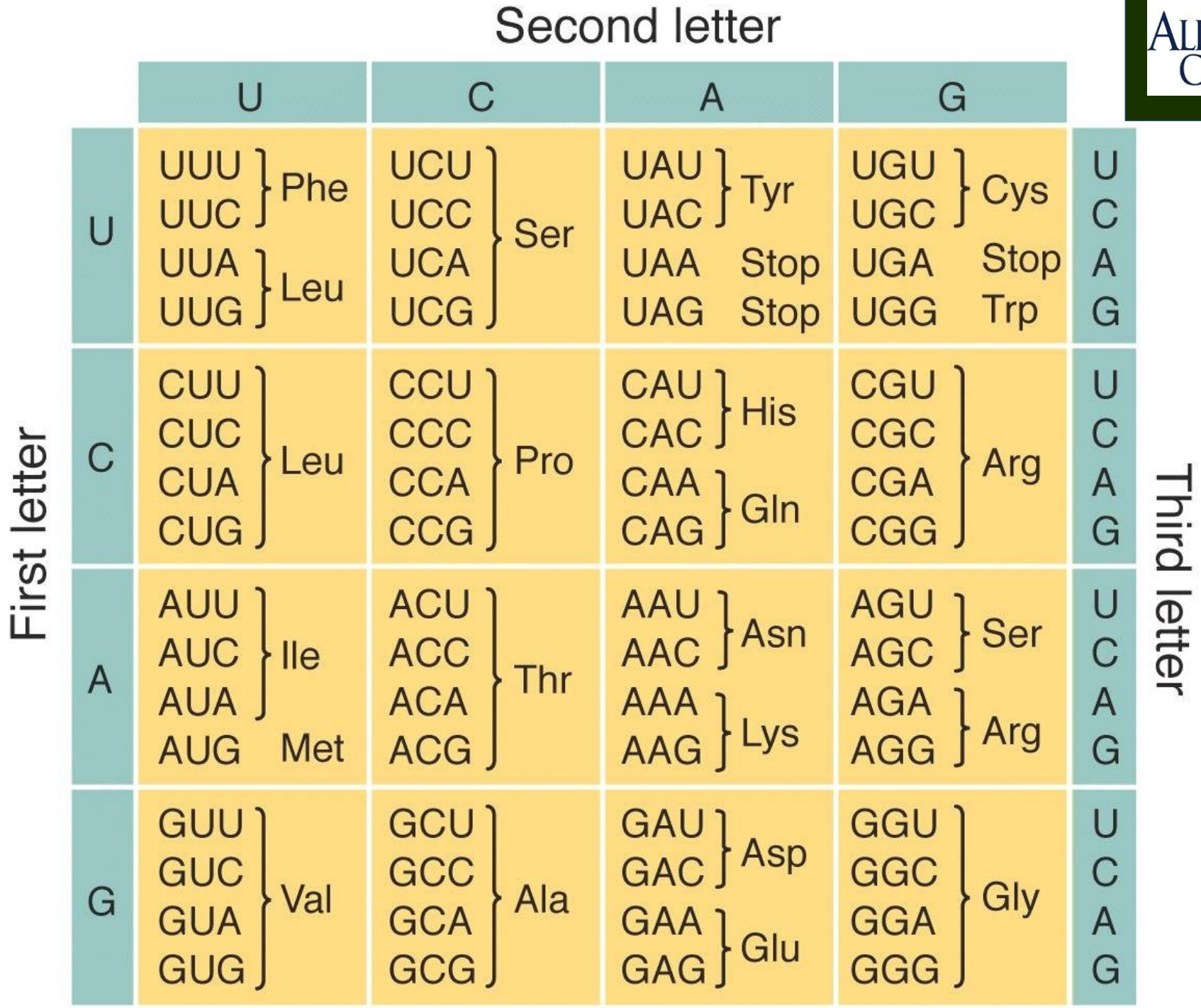
```
406..621
/gene="CFTR"
/gene_synonym="ABC35; ABCC7; CF; CFTR/MRP; dJ760C5.1;
MRP7; TNR-CFTR"
/inference="alignment:Splign:2.1.0"
```

Details Display: FASTA GenBank Help

Quick link:
https://www.ncbi.nlm.nih.gov/nuccore/NM_000492.3

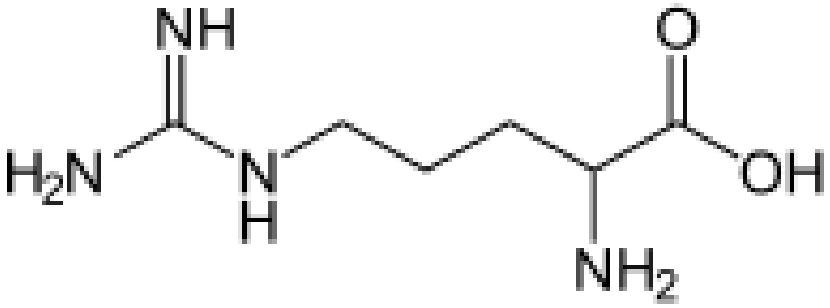
Awesome!

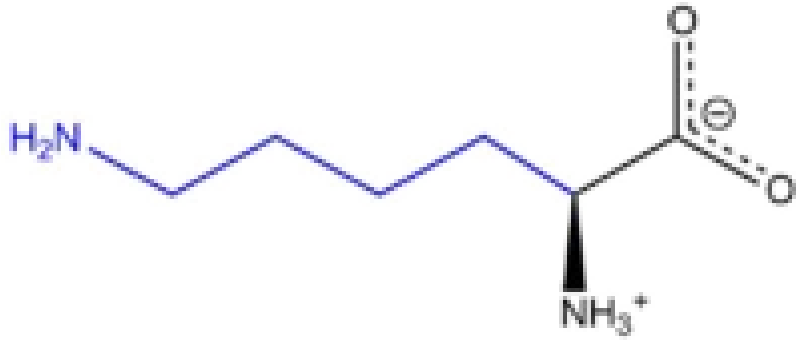
(*But not as awesome as this!*)



Protein: Alphabetical Interests

- With a larger protein “alphabet” (20 amino acids), it is much less likely to get matches by chance.
- Matches are likely to be statistically significant
- Amino acid changes are not equally harmful to protein structure
  - Chemical complexes being replaced by similar chemical complex.
  - Ex: Arginine (Arg) and Lysine (Lys)
  - Can this substitution cause harm, now or later?!





Amino Acid Substitutions

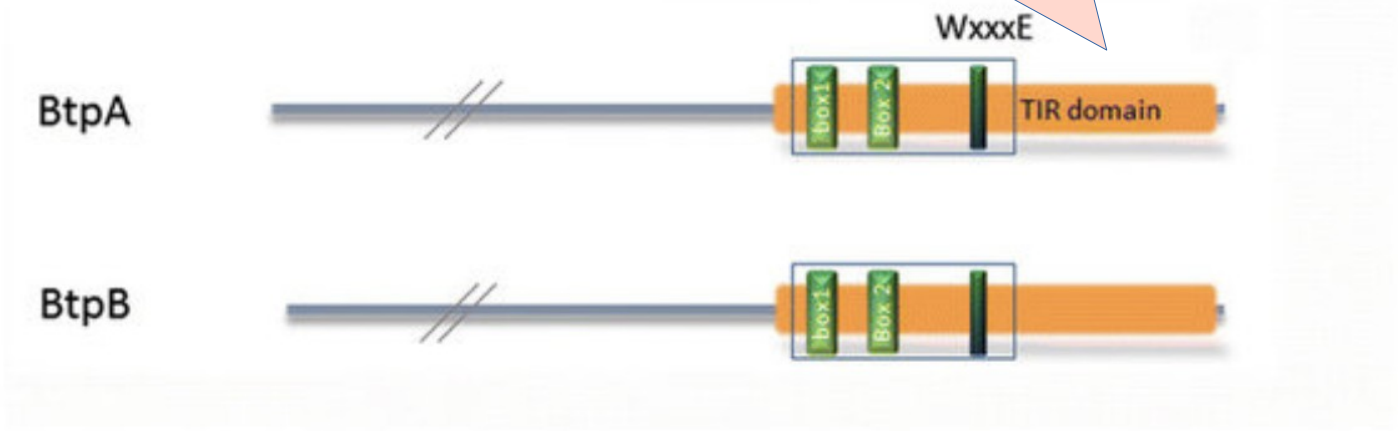
- Nucleotides – any (base) substitution makes the genetics "different" *in some way*

- Amino Acids
  - Substituting similar ones is likely to retain protein structure and function
  - Substituting dissimilar ones is likely to change protein structure and disrupt function

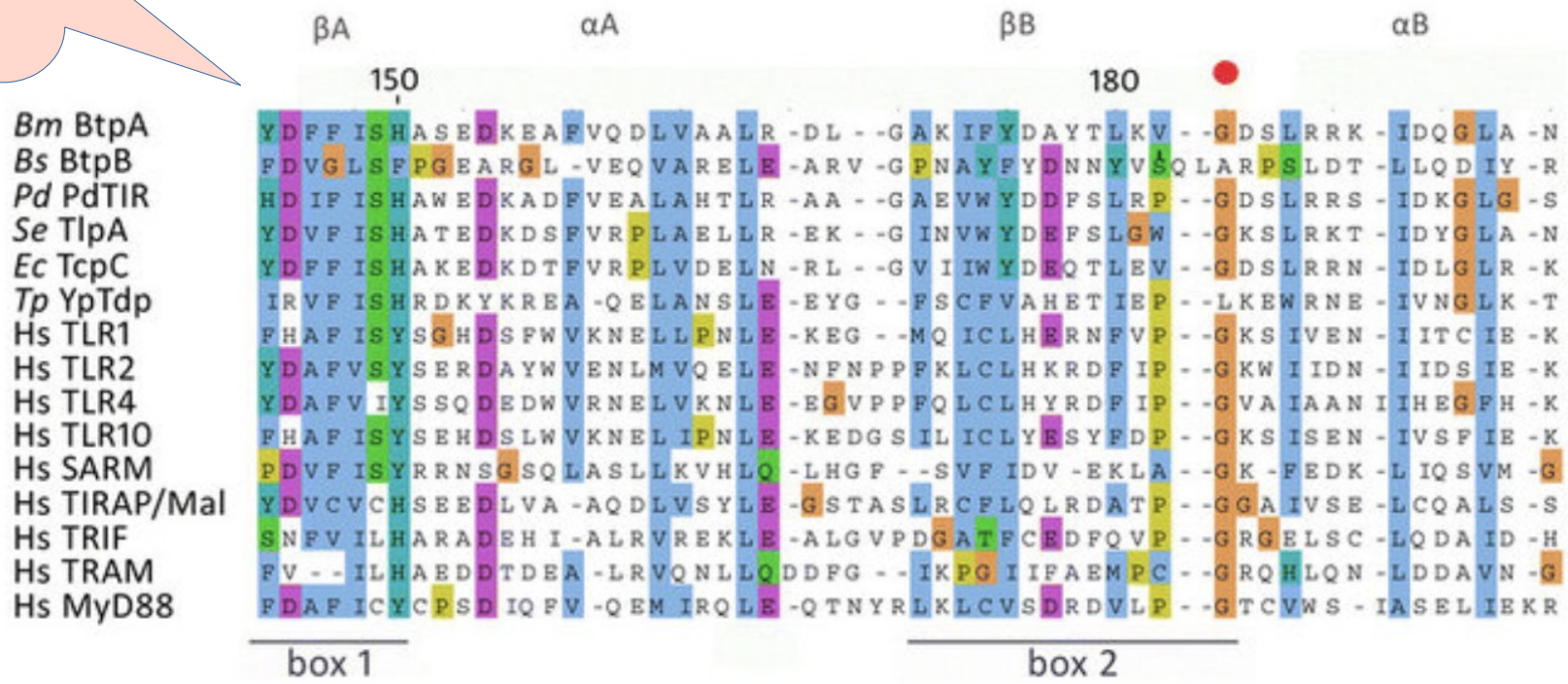
Wait! How did we know where silent mutations came from?!

Alignment of Protein *Domains*: the “Functional” Parts of Protein

These domains have individual functions



Change domain composition may change function



Felix, Christine, et al. "The Brucella TIR domain containing proteins BtpA and BtpB have a structural WxxxE motif important for protection against microtubule depolymerisation."Â Cell Communication and SignalingÂ 12.1 (2014): 1-15.

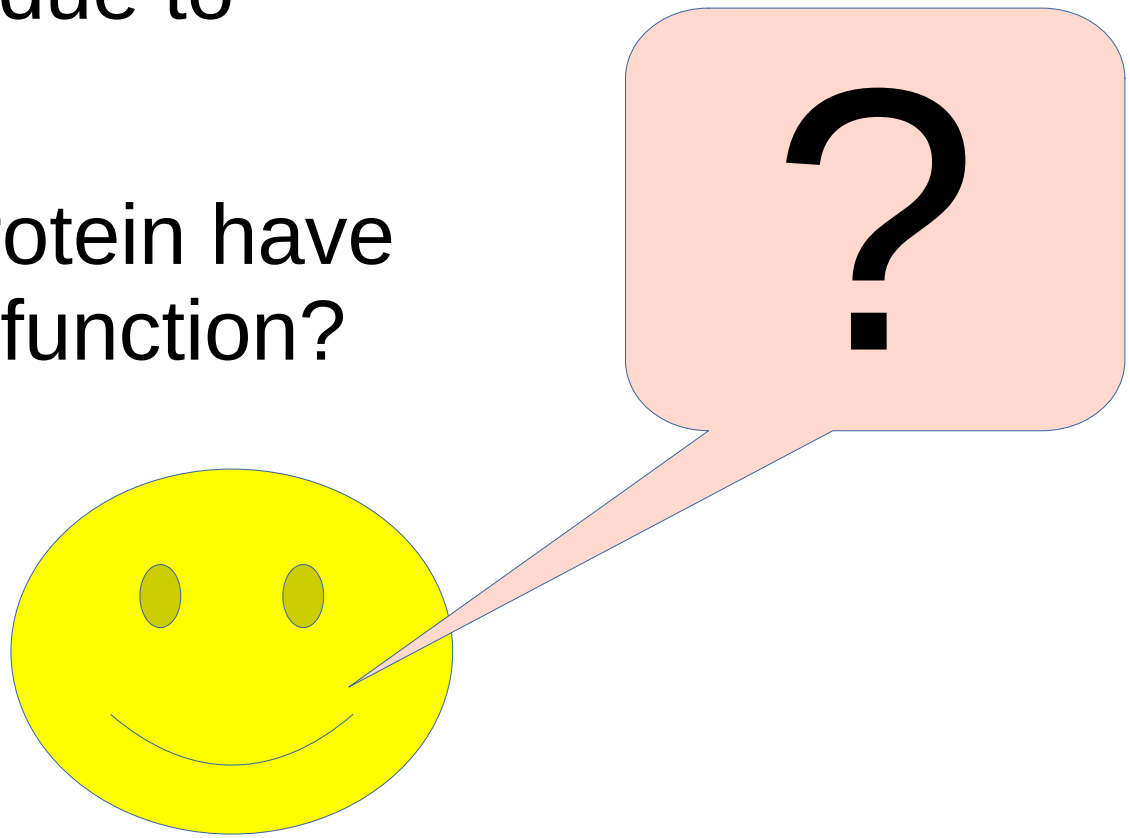
When Comparing Proteins...

How are proteins different?

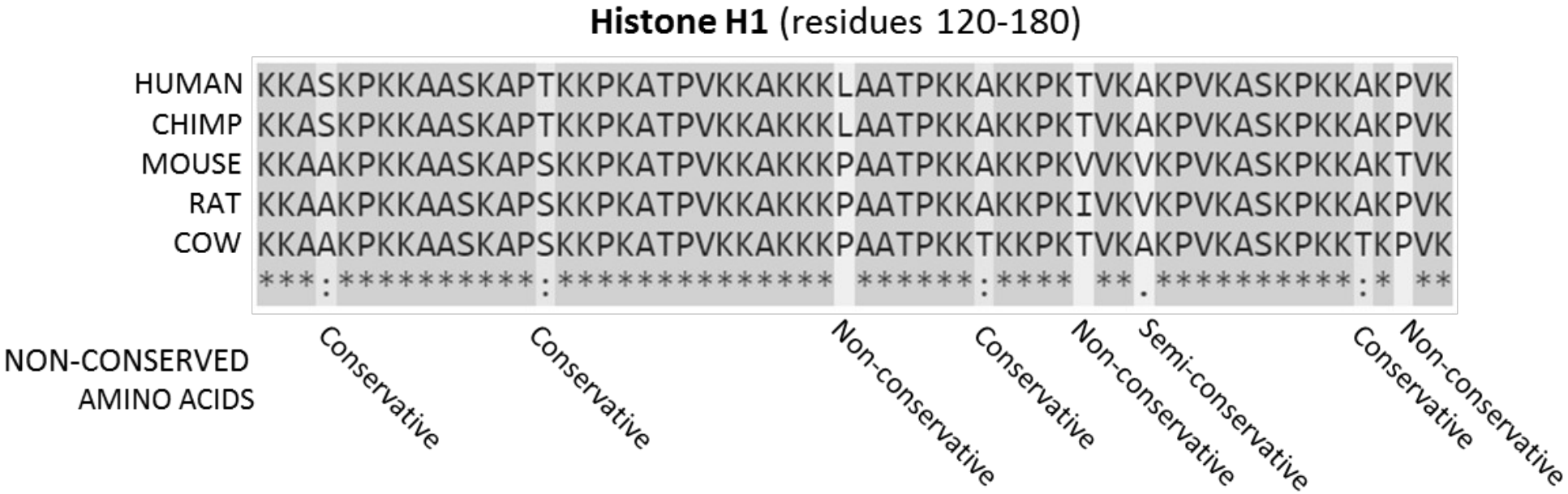
How much difference is there?

Was this difference due to chance?

Could the altered protein have a new and different function?



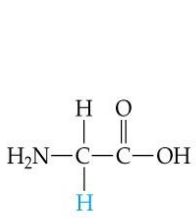
Protein Amino Acid Replacements

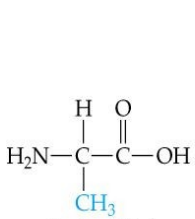


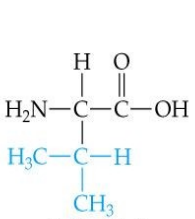
Generally, replacements are ...

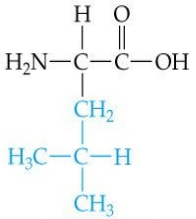
- **Conservative**: a change to an amino acid with similar physio-chemical properties; a smaller effect on function than non-conservative replacements.
- **Semi-conservative**: Minor changes that persist, depending on evolutionary conditions
- **Non-conservative:** Changes that are likely to be edited out by evolutionary pressures due to their deleterious effects

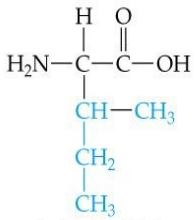
Amino Acids

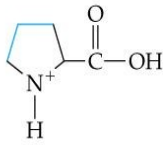

Glycine (Gly)

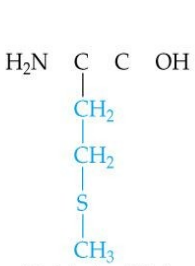

Alanine (Ala)

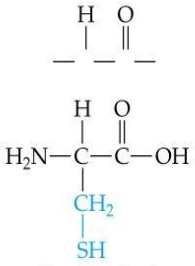

Valine (Val)

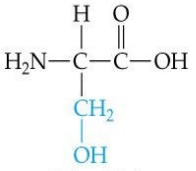

Leucine (Leu)

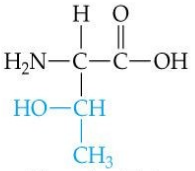

Isoleucine (Ile)

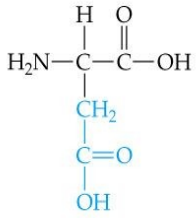

Proline (Pro)

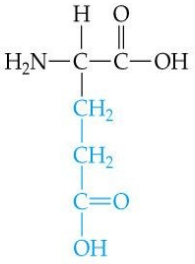

Methionine (Met)

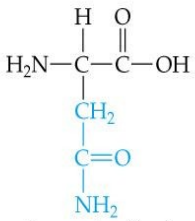

Cysteine (Cys)

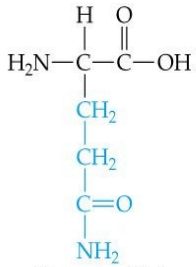

Serine (Ser)

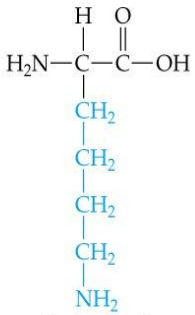

Threonine (Thr)

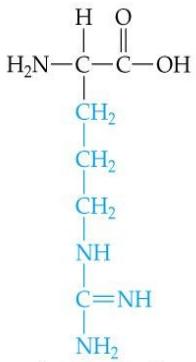

Aspartic acid (Asp)

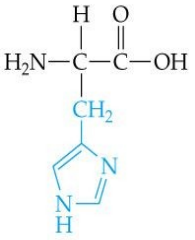

Glutamic acid (Glu)

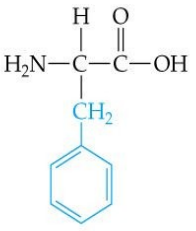

Asparagine (Asn)

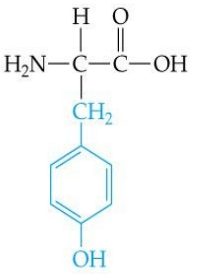

Glutamine (Glu)

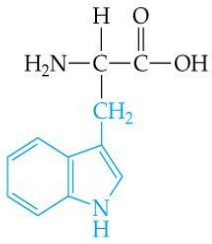

Lysine (Lys)


Arginine (Arg)


Histidine (His)


Phenylalanine (Phe)

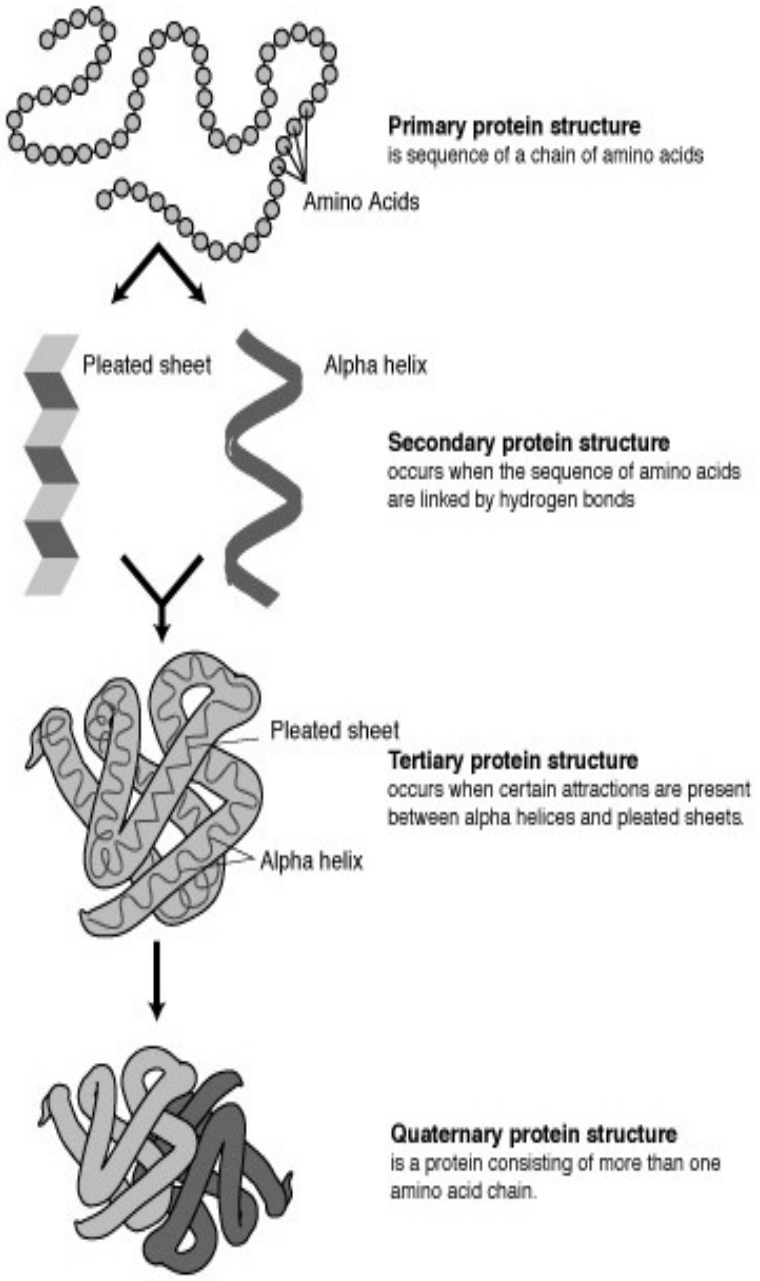

Tyrosine (Tyr)


Tryptophan (Trp)

Amino Acid Components

- **Similarity** of amino acids means
  - Similar *physicochemical properties (*Physics + chemistry)
    - Polar vs nonpolar
    - Hydrophobic vs hydrophilic
    - Positive electric charge vs negative electric charge
    - Basic vs Acidic
- Amino Acids and Chemistry Table: http://www.bio.davidson.edu/courses/genomics/jmol/aatable.html
- Roles in Protein Structures
- http://www.proteinstructures.com/Structure/Structure/amino-acids.html

Amino Acids Determine Protein's Shape and Function



The hierarchy of protein structure. Public domain image from The National Genome Research Institute

Scoring Amino Acid Substitutions

- Could we quantify sequence by physicochemical properties? (yes!)

Table 5.1 Hydrophobicity values for the 20 amino acids. A more positive value represents a more hydrophobic amino acid.

Amino Acid	Hydrophobicity	Amino Acid	Hydrophobicity	Amino Acid	Hydrophobicity
D	−3.5	Y	−1.3	I	4.5
K	−3.9	N	−3.5	C	2.5
H	−3.2	L	3.8	A	1.8
T	−0.7	E	−3.5	S	−0.8
V	4.2	R	−4.5	G	−0.4
F	2.8	W	−0.9	P	−1.6
M	1.9	Q	−3.5		

Scoring Amino Acid Substitutions

Better to study evolution of real proteins from closely related organisms

Minimizes likelihood that an observed difference represents a series of more than one individual mutation

Species A – Ala Species B – Ile

No intermediate mutations?

Ala --> Ile : 1 mutation

Ala --> Pro --> Ser --> Ile : 3 mutations

A few intermediate mutations?

A Model of Evolutionary Change in Proteins, Dayhoff et al., 1978

Global Pairwise Alignment

Observed frequency of each possible amino acid substitution:

$$10 \log_{10} (M_{ij}/f_j)$$

- M_{ij} - the probability of a mutation replacing amino *i* with *j*
- f_j - the frequency of amino acid *j* in a large set of sequences

A Model of Evolutionary Change in Proteins, Dayhoff et al., 1978

Global Pairwise Alignment

Observed frequency of each possible amino acid substitution:

$$10 \log_{10} (M_{ij}/f_j)$$

Odds ratio

= 1 - substitution of *j* for *i* is no more likely than the chance of finding *j* randomly

> 1 - substitution is evolutionarily conserved

< 1 – substitution is selected against

A Model of Evolutionary Change in Proteins, Dayhoff et al., 1978

Global Pairwise Alignment

Observed frequency of each possible amino acid substitution:

$$10 \log_{10} (M_{ij}/f_j)$$

log-odds ratio – easier for scoring

Greater positive for likely (conservative) substitutions

Greater negative for unlikely (non-conservative) substitutions

Multiplied by 10 and rounded to nearest integer

The PAM Matrix

- PAM matrices are used as substitution matrices to score sequence alignments for proteins.
- Each entry in a PAM matrix indicates the likelihood of the amino acid of that row being replaced with the amino acid of that column through a series of one or more point accepted mutations during a specified evolutionary interval, rather than these two amino acids being aligned due to chance.
- Different PAM matrices correspond to different lengths of time in the evolution of the protein sequence.

Ref:
https://en.wikipedia.org/wiki/Point_accepted_mutation

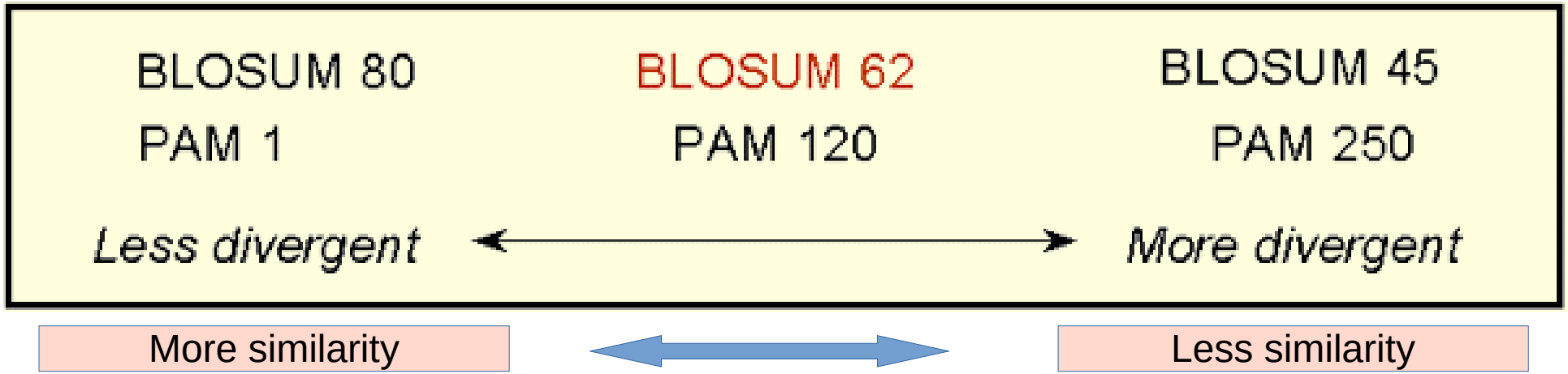
The PAM Matrix

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	2																			
Arg	R	–1	5																		
Asn	N	0	0	3																	
Asp	D	0	–1	2	5																
Cys	C	–1	–1	–1	–3	11															
Gln	Q	–1	2	0	1	–3	5														
Glu	E	–1	0	1	4	–4	2	5													
Gly	G	1	0	0	1	–1	–1	0	5												
His	H	–2	2	1	0	0	2	0	–2	6											
Ile	I	0	–3	–2	–3	–2	–3	–3	–3	–3	4										
Leu	L	–1	–3	–3	–4	–3	–2	–4	–4	–2	2	5									
Lys	K	–1	4	1	0	–3	2	1	–1	1	–3	–3	5								
Met	M	–1	–2	–2	–3	–2	–2	3	3	–2	3	3	–2	6							
Phe	F	–3	–4	–3	–5	0	–4	–5	–5	0	0	2	–5	0	8						
Pro	P	1	–1	–1	–2	–2	0	–2	–1	0	–2	0	–2	–2	–3	6					
Ser	S	1	–1	1	0	1	–1	–1	1	–1	–1	–2	–1	–1	–2	1	2				
Thr	T	2	–1	1	–1	–1	–1	–1	–1	–1	1	–1	–1	0	–2	1	1	2			
Trp	W	–4	0	–5	–5	1	–3	–5	–2	–3	–4	–2	–3	–3	–1	–4	–3	–4	15		
Tyr	Y	–3	–2	–1	–2	2	–2	–4	–4	4	–2	–1	–3	–2	5	–3	–1	–3	0	9	
Val	V	1	–3	–2	–2	–2	–3	–2	–2	–3	4	2	–3	2	0	–1	–1	0	–3	–3	4

The probability calculations for Substitutions have been done for you!

PAM Matrices

- **P**oint **A**ccepted **M**utation
- Family of matrices PAM 1, PAM 80, PAM 120, PAM 250
- The number in the name of a PAM matrix (i.e., the '*n*' in PAM *n*) represents the evolutionary distance between the sequences on which the matrix is based



PAM vs BLOSUM

- General Use
  - PAM 120
  - BLOSUM 62*
- Closely Related Species
  - PAM 60
  - BLOSUM 80
- Distantly Related Species
  - PAM 250
  - BLOSUM 45

PAM	BLOSUM
PAM100	BLOSUM90
PAM120	BLOSUM80
PAM160	BLOSUM60
PAM200	BLOSUM52
PAM250	BLOSUM45

*BLOSUM 62 – used by BLAST – computed by choosing blocks of local alignments more than 62% identical
** BLOSUM matrices are gradually replacing PAM matrices thanks to advanced data analysis for calculating probabilities of substitutions

Blast Subst Matrices

- Scoring for possible residue pair alignment
- Different substitution matrices are for detecting similarities according to degrees of divergence.
- BLOSUM-62 matrix good for detecting most weak protein similarities
- Provisional table of recommended substitution matrices and gap costs for various query lengths is

Query Length	Substitution Matrix	Gap Costs
<35	PAM-30	(9,1)
35-50	PAM-70	(10,1)
50-85	BLOSUM-80	(10,1)
85	BLOSUM-62	(10,1)

https://www.ncbi.nlm.nih.gov/blast/html/sub_matrix.html

BLOSUM matrix
Heinkoff and Heinkoff, 1992

- **BLOcks SUbstition Matrix** - Blocks of local alignments

$$S_{ij} = \left(\frac{1}{\lambda}\right) \log\left(\frac{p_{ij}}{q_i * q_j}\right)$$

- p_{ij} - probability j replacing i
- q_i and q_j - probabilities of finding the amino acids i and j in any protein sequence
- λ - scaling factor, set such that the matrix contains easily computable integer values.

- BLOSUM # - # = minimum % similarity of sequences compared

Needleman-Wunsch Algorithm: Nucleotide Alignment – Chap 3

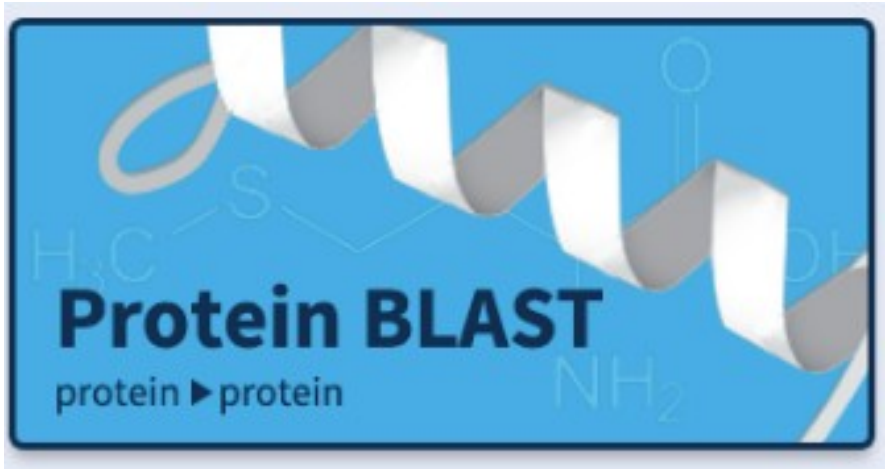
- Create N x M matrix
- Place each sequence along one axis
- Place score 0 at the up-left corner
- Fill in 1st row & column with gap penalty multiples
- Fill in the matrix with max value of 3 possible moves:
  - Vertical move: Score + gap penalty
  - Horizontal move: Score + gap penalty
  - Diagonal move: Score + match/mismatch score
- The optimal alignment score is in the lower-right corner
- To reconstruct the optimal alignment, trace back where the max at each step came from, stop when hit the origin.

Needleman-Wunsch Algorithm:
Protein Alignment – Chap 5

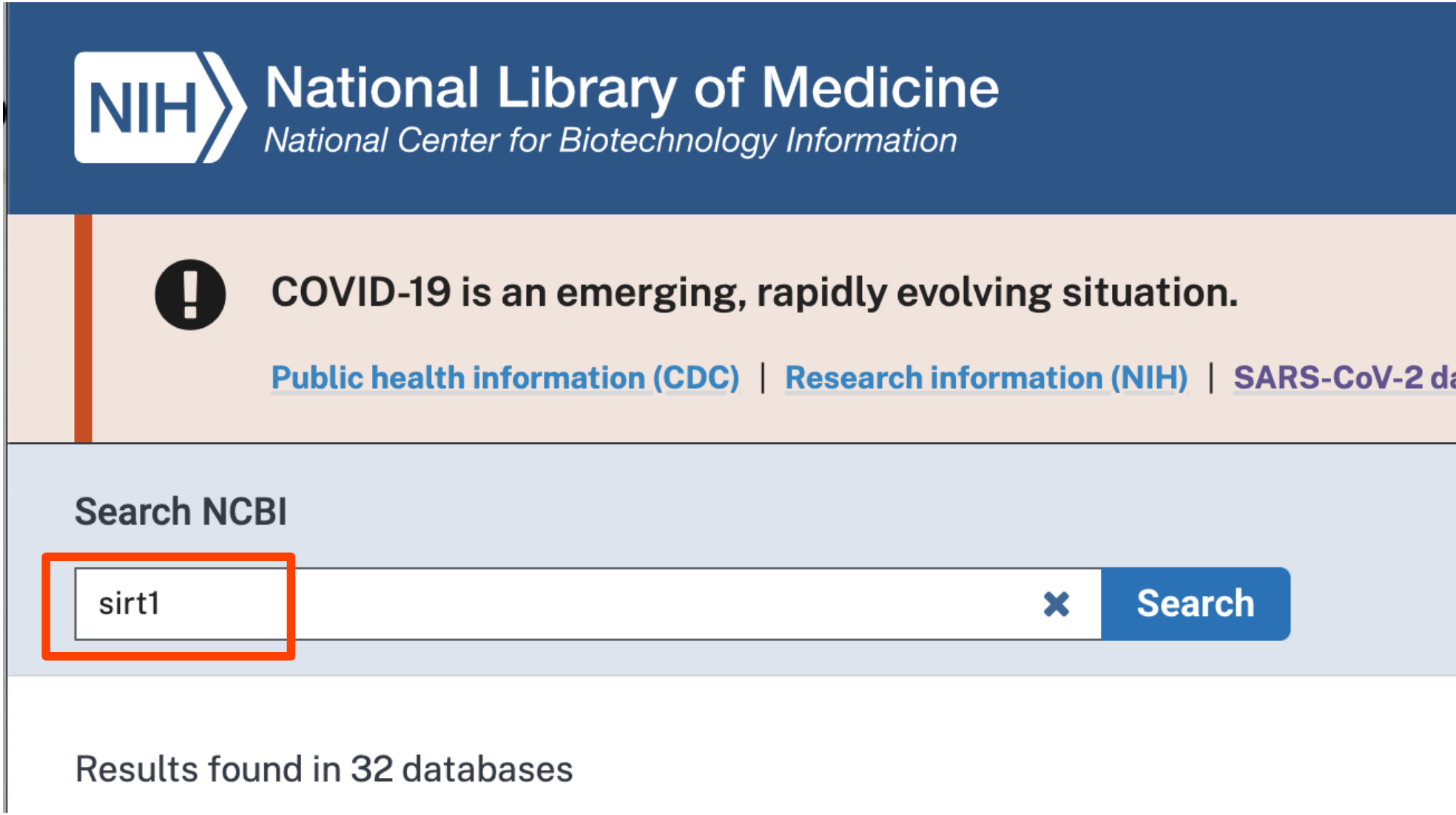
- Create N x M matrix
- Place each sequence along one axis
- Place score 0 at the up-left corner
- Fill in 1st row & column with gap penalty multiples
- Fill in the matrix with max value of 3 possible moves:
  - Vertical move: Score + gap penalty
  - Horizontal move: Score + gap penalty
  - Diagonal move: Score + **match/mismatch score from sub. matrix**
- The optimal alignment score is in the lower-right corner
- To reconstruct the optimal alignment, trace back where the max at each step came from, stop when hit the origin.

Blast-Off!!

- Let's blast some protein sequences
- https://blast.ncbi.nlm.nih.gov/Blast.cgi#dtr_Query_98931



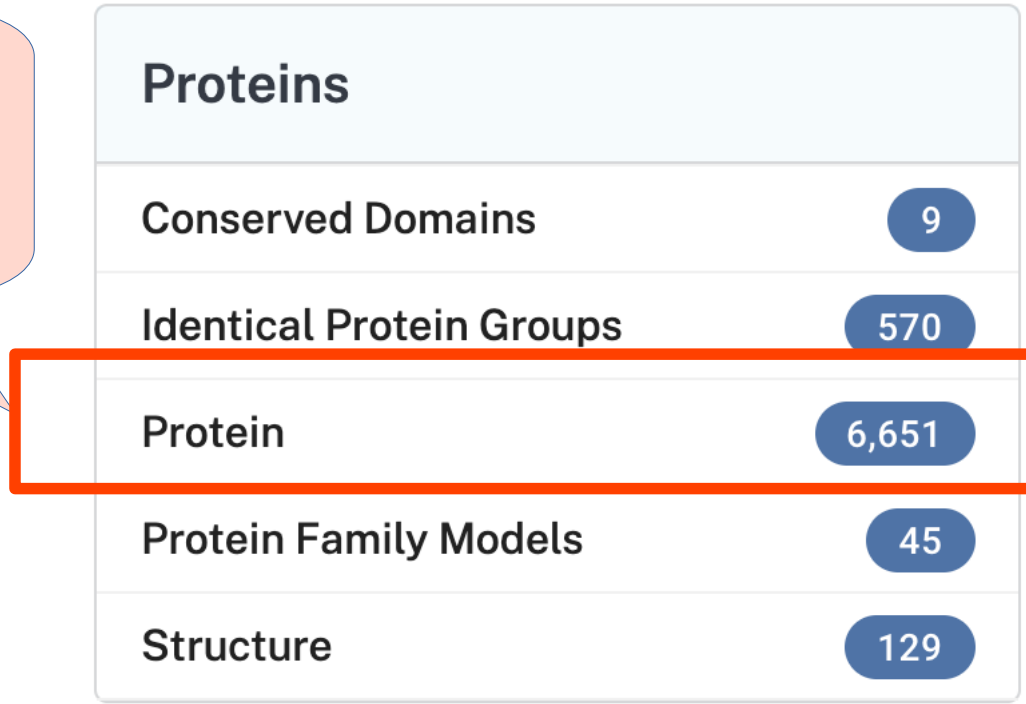
Blasting Proteins



Blasting Proteins

Select Protein

For example, choose this one and use Blast link In the record



2. **SIRT1** [Mytilus coruscus]

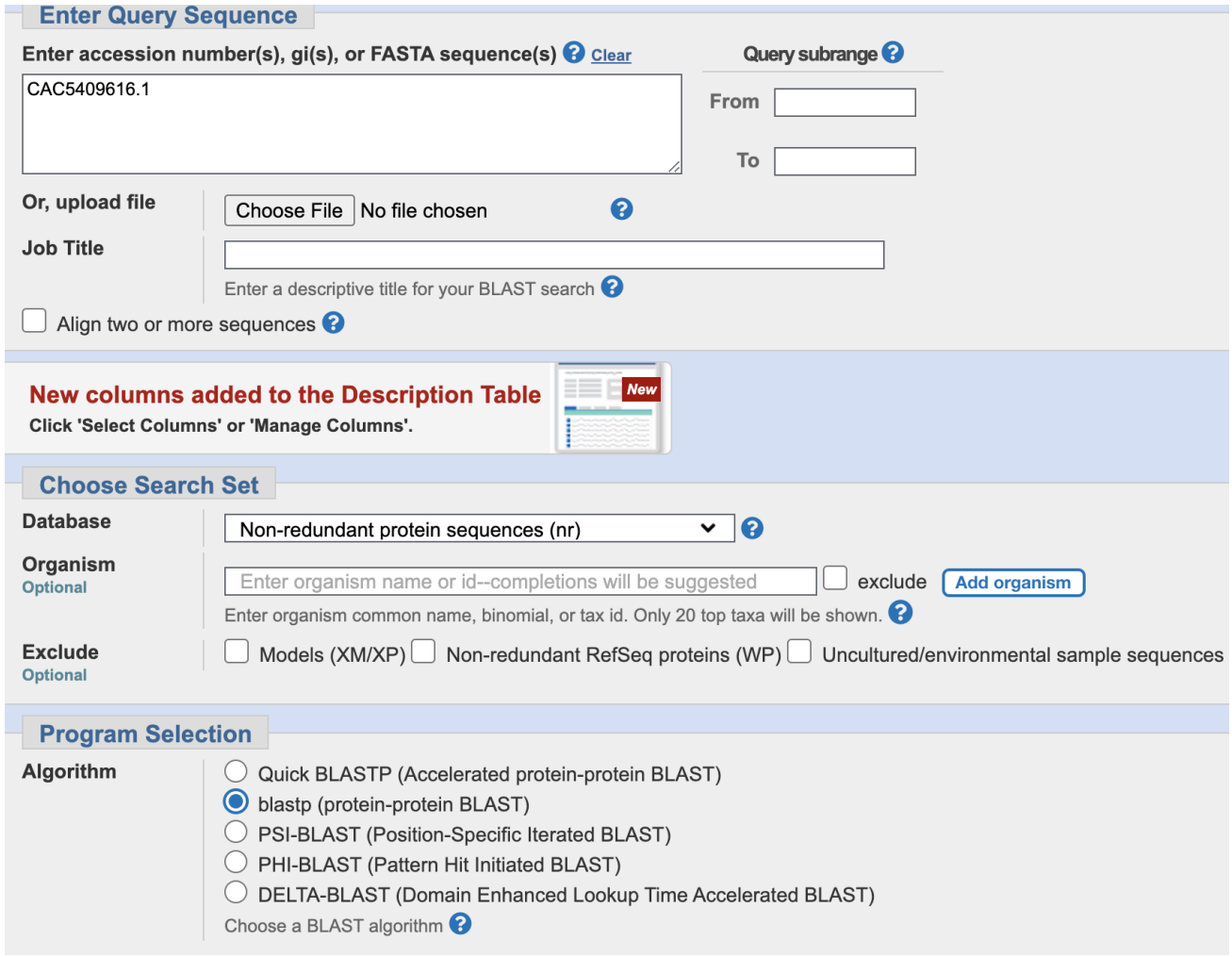
188 aa protein

Accession: CAC5409616.1 GI: 1866842361

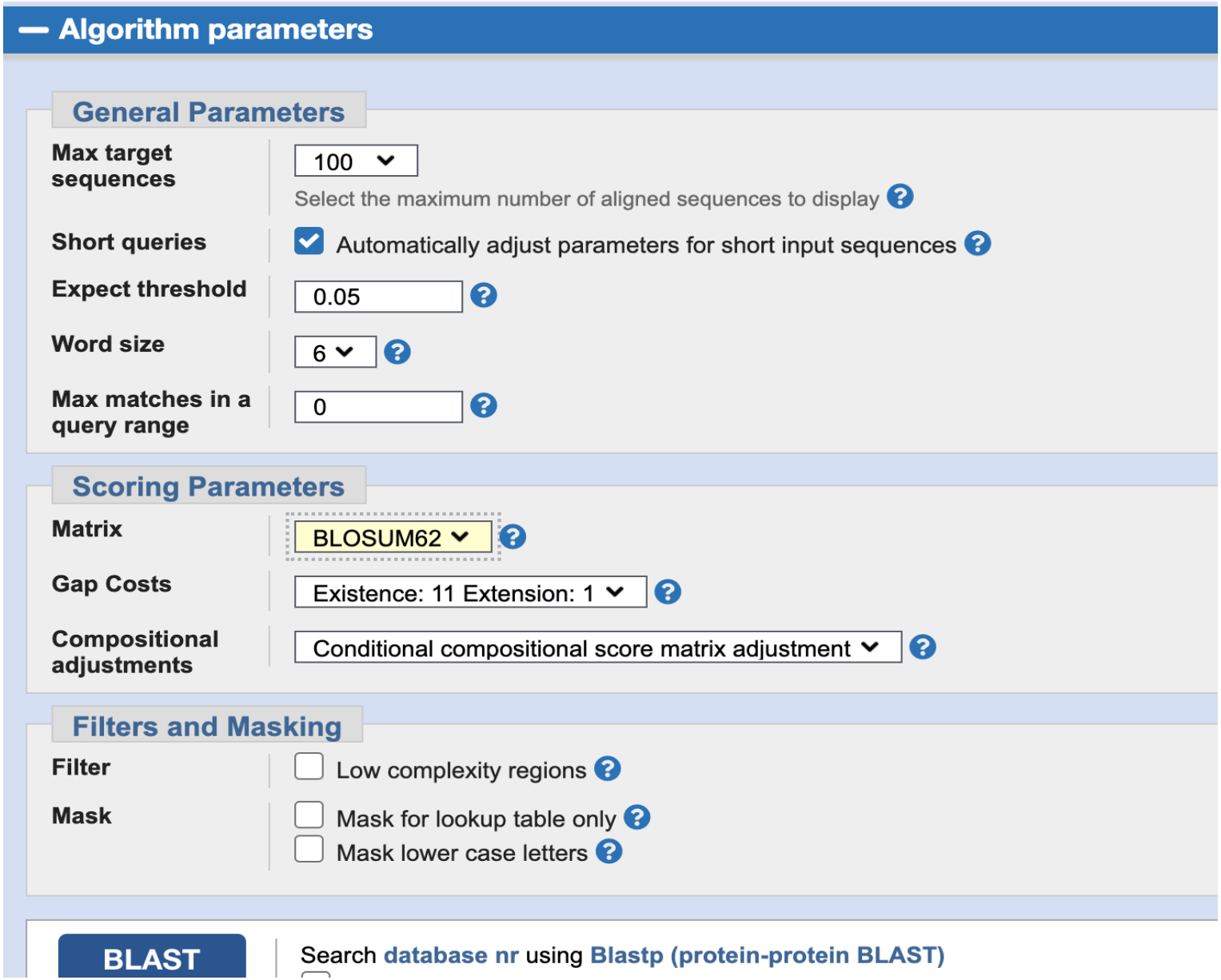
BioProject Nucleotide Taxonomy

GenPept Identical Proteins FASTA Graphics

Blasting Options



Blasting Options



Blasting Results

GenPept Graphics Distance tree of results Multiple alignment New MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
SIRT1 [Mytilus coruscus]	Mytilus c…	382	382	100%	9e-134	100.00%	188	CAC5409616.1
NAD-dependent deacetylase sirtuin 1 [Mytilus galloprovincialis]	Mytilus g…	323	323	89%	1e-102	93.45%	826	VDI49146.1
NAD-dependent protein deacetylase sirtuin-1-like isoform X3…	Mizuhop…	178	178	71%	3e-48	60.14%	869	XP_021368443.1
LOW QUALITY PROTEIN: NAD-dependent protein deacetyl…	Pecten m…	178	178	63%	4e-48	65.29%	834	XP_033738334.1
NAD-dependent protein deacetylase sirtuin-1 [Mizuhopecten …	Mizuhop…	176	176	71%	2e-47	60.42%	850	OWF43165.1
NAD-dependent protein deacetylase sirtuin-1-like [Pomacea …	Pomacea…	165	165	60%	1e-43	66.37%	848	XP_025089963.1
NAD-dependent protein deacetylase sirtuin-1-like isoform X2…	Mizuhop…	162	162	57%	9e-43	67.89%	749	XP_021368442.1
NAD-dependent protein deacetylase sirtuin-1-like isoform X1…	Mizuhop…	162	162	57%	1e-42	67.89%	765	XP_021368441.1
NAD-dependent protein deacetylase sirtuin-1 isoform X1 [Lin…	Lingula a…	157	157	60%	5e-41	60.53%	737	XP_013411402.1
NAD-dependent protein deacetylase sirtuin-1 isoform X2 [Lin…	Lingula a…	157	157	60%	5e-41	60.53%	736	XP_013411403.1
hypothetical protein Cfor_04474 [Coptotermes formosanus]	Coptoter…	155	155	76%	4e-40	52.78%	880	GFG40552.1
unnamed protein product [Timema poppensis]	Timema …	149	149	59%	5e-40	60.71%	377	CAD7411610.1
hypothetical protein C0J52_01051 [Blattella germanica]	Blattella …	155	155	71%	7e-40	53.62%	866	PSN54664.1

Blasting Results

Download GenPept Graphics Next Previous

SIRT1 [Mytilus coruscus]

Sequence ID: CAC5409616.1 **Length:** 188 **Number of Matches:** 1

Range 1: 1 to 188 GenPept Graphics Next Match Previou

Score	Expect	Method	Identities	Positives	Gaps
382 bits(980)	9e-134	Compositional matrix adjust.	188/188(100%)	188/188(100%)	0/188(0%)

```
Query   1     MESAELPRKMAAQPLKNEPPVKRQKLNEEDDNDSDEQGCSNISKDNEAGNTEQ
              MESAELPRKMAAQPLKNEPPVKRQKLNEEDDNDSDEQGCSNISKDNEAGNTEQ
Sbjct   1     MESAELPRKMAAQPLKNEPPVKRQKLNEEDDNDSDEQGCSNISKDNEAGNTEQ

Query   61    IDNSDNCSEISNLSGLSEEAWKPTSGAMSWIHKQIMNGVNPRPILNGLIPDDT
              IDNSDNCSEISNLSGLSEEAWKPTSGAMSWIHKQIMNGVNPRPILNGLIPDDT
Sbjct   61    IDNSDNCSEISNLSGLSEEAWKPTSGAMSWIHKQIMNGVNPRPILNGLIPDDT

Query   121   DFTLWKIVINIMSEPPPRKKLSHINTLQDVIQLLQNCKNIMVLTGAGVSVSCQ
              DFTLWKIVINIMSEPPPRKKLSHINTLQDVIQLLQNCKNIMVLTGAGVSVSCQ
Sbjct   121   DFTLWKIVINIMSEPPPRKKLSHINTLQDVIQLLQNCKNIMVLTGAGVSVSCQ

Query   181   MESMLALL  188
              MESMLALL
Sbjct   181   MESMLALL  188
```