

# **Bioinformatics**

## **CS300**

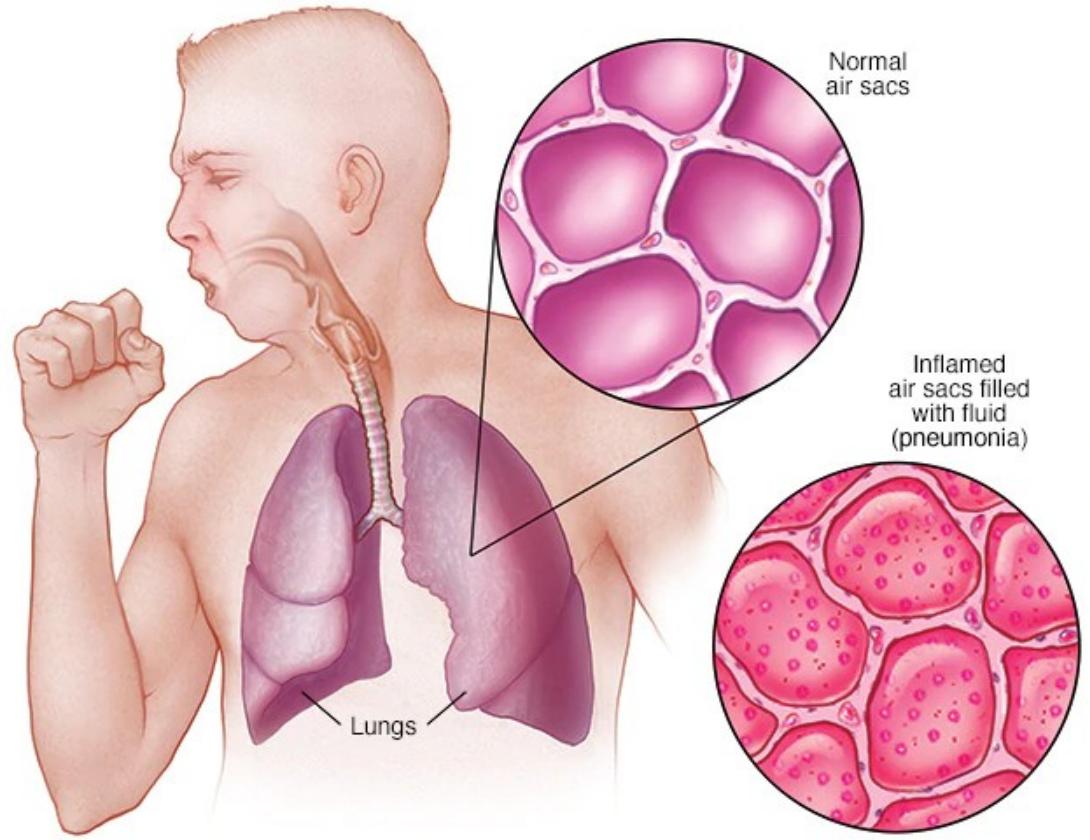
**Blast, Substitution Matrices and  
Protein Alignments  
(Chap 4 and 5 in textbook)**

**Spring 2021**  
**Oliver BONHAM-CARTER**



# Pneumonia

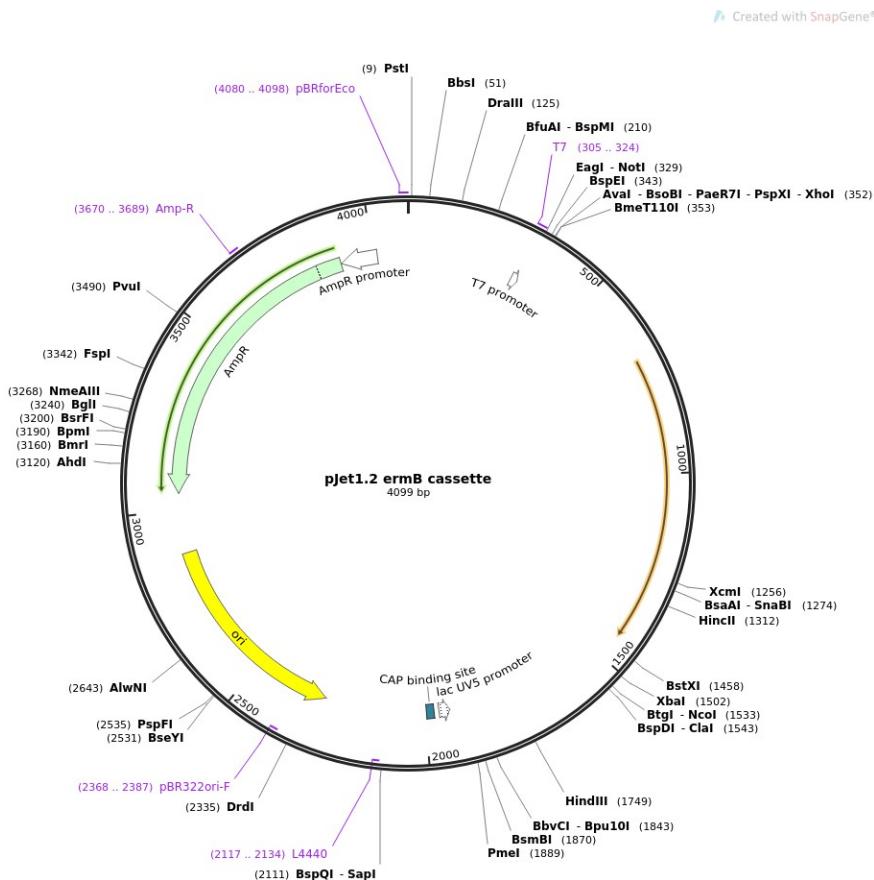
- Pneumonia is an infection that inflames the air sacs in one or both lungs. The air sacs may fill with fluid or pus (purulent material), causing cough with phlegm or pus, fever, chills, and difficulty breathing. A variety of organisms, including bacteria, viruses and fungi, can cause pneumonia.
- A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





# 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)





# 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

Look for  
this record

Nucleotide

1

NCBI Resources ▾ How To ▾

**PubMed.gov**  
US National Library of Medicine  
National Institutes of Health

PubMed DQ355148.1  
Create alert Advanced

**Article types**  
Clinical Trial  
Review  
Customize ...

**Text availability**  
Abstract  
Free full text  
Full text

[Streptococcus agalactiae strain KMP104 transposon Tn917 rRNA methylase \(\*ermB\*\) gene, complete cds](#)  
738 bp genomic DNA.  
Strain: KMP104.  
Accession: **DQ355148.1** GI: 87042723  
[GenBank](#) [FASTA](#) [Graphics](#)

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



# How to get the Data?

A screenshot of a web browser displaying a protein sequence from NCBI. A context menu is open over the sequence, with the 'Send to:' option expanded. The 'Send to:' dropdown shows three options: 'Complete Record' (selected), 'Coding Sequences', and 'Gene Features'. Below this, under 'Choose Destination', there are two groups of options: 'File' (selected) and 'Clipboard' in the first group, and 'Collections' and 'Analysis Tool' in the second group. At the bottom of the menu, it says 'Download 1 item.' followed by a 'Format' dropdown set to 'FASTA' and a 'Show GI' checkbox. A 'Create File' button is also present.

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

A screenshot of a web browser displaying a protein sequence from NCBI. A context menu is open over the sequence, with the 'Send to:' option expanded. The 'Send to:' dropdown shows three options: 'Complete Record' (selected), 'Coding Sequences', and 'Gene Features'. Below this, under 'Choose Destination', there are two groups of options: 'File' (selected) and 'Clipboard' in the first group, and 'Collections' and 'Analysis Tool' in the second group. At the bottom of the menu, it says 'Download 1 item.' followed by a 'Format' dropdown set to 'FASTA' and a 'Show GI' checkbox. A 'Create File' button is also present. To the right of the menu, several other options are visible: 'Change region shown', 'Customize view', 'Analyze this sequence' (with sub-options 'Run BLAST', 'Pick Primers', 'Highlight Sequence Features', and 'Find Similar Sequence').

**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 ▾

## 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

Send to: ▾

- 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 (Weighted)

## Recent activity

[Streptococcus agalactiae](#) transposon Tn917[DQ355148.1](#)



# Ah, The Sequence in FASTA Format

```
>DQ355148.1 Streptococcus agalactiae strain KMP104 transposon Tn917 rRNA methylase (ermB) gene, complete cds
ATGAACAAAAATATAAAATATTCTCAAAACTTTAACGAGTGAAAAAGTACTCAACCAAATAATAAAAC
AATTGAATTAAAAGAAACCGATACCGTTACGAAATTGGAACAGGTAAAGGGCATTAAACGACGAAACT
GGCTAAAATAAGTAAACAGGTAACGTCATTGAATTAGACAGTCATCTATTCAACTTATCGTCAGAAAAA
TTAAAACGTAAACATTCGTGTCACTTTAATTACCAAGATATTCTACAGTTCAATTCCCTAACAAACAGA
GGTATAAAATTGTTGGGAATTCCCTTACCATTTAACGCACACAAATTATTAAAAAGTGGTTTGAAAG
CCATGCGTCTGACATCTATCTGATTGTTGAAGAAGGATTCTACAAGCGTACCTGGATATTACCGAAC
CTAGGGTTGCTCTGCACACTCAAGTCTCGATTCAAGCTTAAGCTGCCAGCGGAATGCTTCATC
CTAAACCAAAAGTAAACAGTGTCTAATAAAACTTACCCGCCATACCACAGATGTTCCAGATAAATATTG
GAAGCTATACGTACTTGTTCAAAATGGGTCAATCGAGAATATCGTCAACTGTTACTAAAAATCAG
TTTCATCAAGCAATGAAACACGCCAAGTAAACAATTAAAGTACCGTTACTATGAGCAAGTATTGTCTA
TTTTAATAGTTATCTATTATTAACGGGAGGAAATAA
```



# Blast Website

NIH U.S. National Library of Medicine NCBI Sign in to NCBI

**BLAST®** Home Recent Results Saved Strategies Help

## Basic Local Alignment Search Tool

**BLAST** finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

[Learn more](#)

**N E W S**

**End of updates for BLAST+ version 4 databases (dbV4)**  
Start moving to the new version 5 databases!  
Fri, 27 Sep 2019 16:00:00 EST [More BLAST news...](#)

### Web BLAST

**Nucleotide BLAST**  
nucleotide ► nucleotide

**blastx**  
translated nucleotide ► protein

**tblastn**  
protein ► translated nucleotide

**Protein BLAST**  
protein ► protein

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



# Run The Query

Standard Nucleotide BLAST

blastn    blastp    blastx    tblastn    tblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

**Enter Query Sequence**

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

```
>DQ355148.1 Streptococcus agalactiae strain KMP104 transposon Tn917 rRNA methylase (ermB) gene, complete cds
ATGAACAAAAATATAAAATTCTCAAAACTTTAACGAGTGAAAAAGTACTCAACCAAATAATAAAAC
AATTGAATTAAAAGAAACCGATACCGTTACGAAATTGGAACAGGTAAAGGGCATTAAACGACGAAACT
GGCTAAAAATAAGTAAACAGGTAAACGTCTATTGAATTAGACAGTCATCTATTCAACTTATCGTCAGAAAAA
TTAAAACGTAAACATTCTGTCACTTAATTCCAAGATATTCTACAGTTCAATTCCCTAACAAACAGA
GGTATAAAATTGTTGGGAATTCTTACCATTAAGCACACAAATTATTAAAAAAGTGGTTTGAAAG
CCATGCCGTCTGACATCTATCTGATTGTTGAAGAAGGATTCTACAAGCGTACCTTGGATATTCAACCGAACAA
```

**Query subrange** [?](#)

From

To

Or, upload file  Choose File No file chosen [?](#)

Job Title  DQ355148.1 Streptococcus agalactiae strain...  
Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

**BLAST results will be displayed in a new format by default**  
You can always switch back to the Traditional Results page.

**Choose Search Set**

Database  Human genomic + transcript  Mouse genomic + transcript  Others (nr/nt etc.) [?](#)

Nucleotide collection (nr/nt) [?](#)

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



# Results

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



# 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.



ALLEGHENY  
COLLEGE

# Results

Descriptions

**Graphic Summary**

Alignments

Taxonomy

👉 hover to see the title

👉 click to show alignments

Alignment Scores

■ < 40

■ 40 - 50

■ 50 - 80

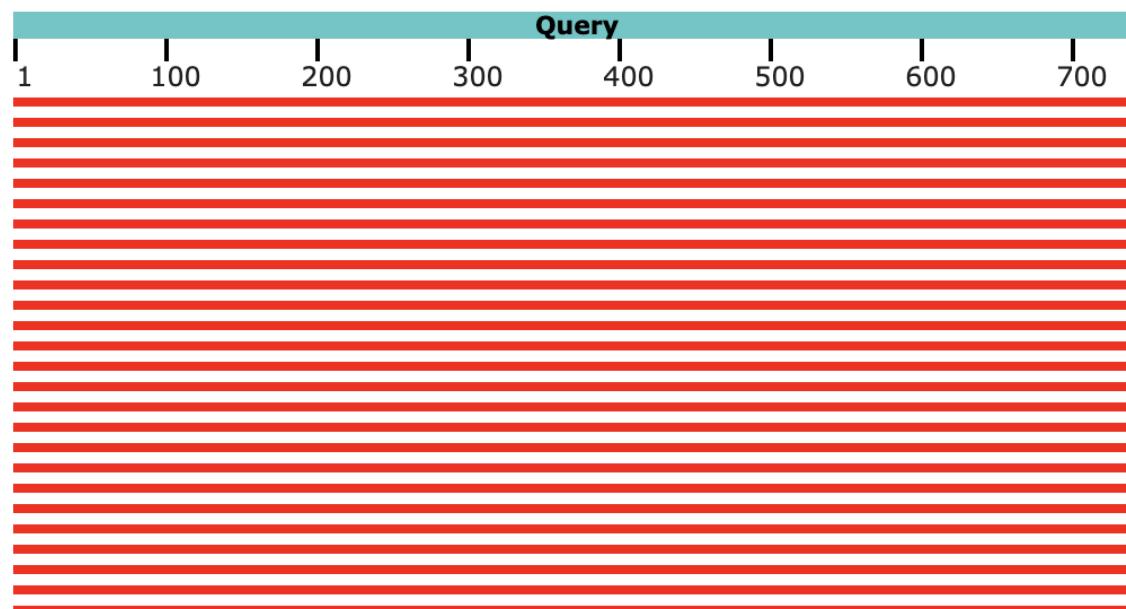
■ 80 - 200

■ >= 200

100 sequences selected



## Distribution of the top 111 Blast Hits on 100 subject sequences





ALLEGHENY  
COLLEGE

# Results

Descriptions

**Graphic Summary**

Alignments

Taxonomy

👉 hover to see the title

👉 click to show alignments

Alignment Scores

■ < 40

■ 40 - 50

■ 50 - 80

■ 80 - 200

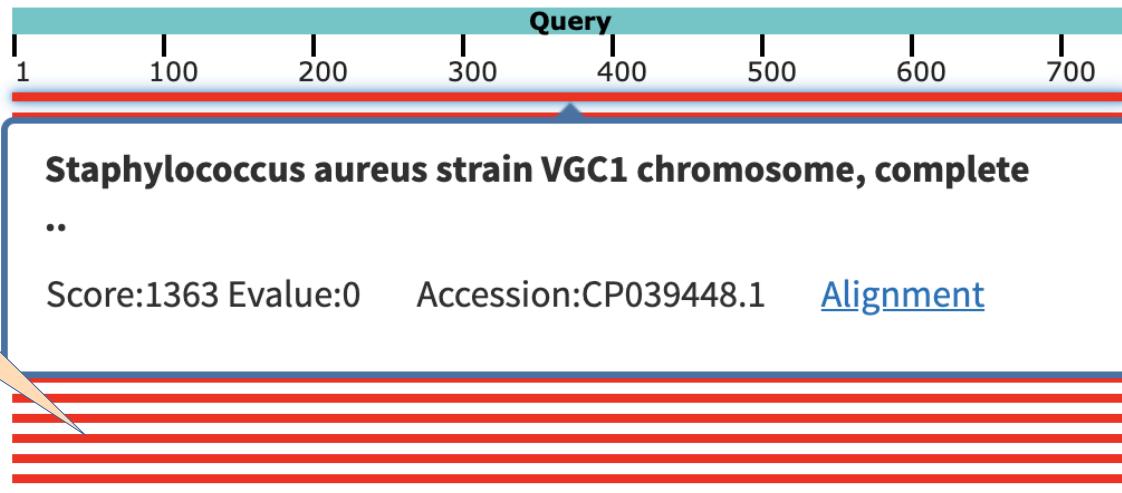
■ >= 200

100 sequences selected



Sequences  
Producing  
Significant  
alignments.

## Distribution of the top 111 Blast Hits on 100 subject sequences





# 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

AACAGGTAAACGTCATTGAATTAGACAGTCATCTATTCAACTTATCGTCAGAAAAA

An Identical sequence in another's genome

99 AACAGGTAAACGTCATTGAATTAGACAGTCATCTATTCAACTTATCGTCAGAAAAA

58 AACTGAATACTCGTGTCACTTAATTCAACCAAGATATTCTACAGTTCAATTCCCT

58 AACTGAATACTCGTGTCACTTAATTCAACCAAGATATTCTACAGTTCAATTCCCT

58 AACAGAGGTATAAAATTGTTGGGAATATTCTTACCATTTAACGACACAAATTATT

Sbjct 16118 AACAGAGGTATAAAATTGTTGGGAATATTCTTACCATTTAACGACACAAATTATT

Query 181 AAGTGGTTTTGAAAGCCGTGCGTCTGACATCTATCTGATTGTTGAAGAAGGATT

Sbjct 16178 AAGTGGTTTTGAAAGCCGTGCGTCTGACATCTATCTGATTGTTGAAGAAGGATT

Query 241 AGCGTACCTTGGATATTCAACCGAACACTAGGGTTGCTCTGCACACTCAAGTCTC

Sbjct 16238 AGCGTACCTTGGATATTCAACCGAACACTAGGGTTGCTCTGCACACTCAAGTCTC

Query 301 AGCAATTGCTTAAGCTGCCAGCGGAATGCTTCATCCTAAACCAAAAGTAAACAG

Sbjct 16298 AGCAATTGCTTAAGCTGCCAGCGGAATGCTTCATCCTAAACCAAAAGTAAACAG



# 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?



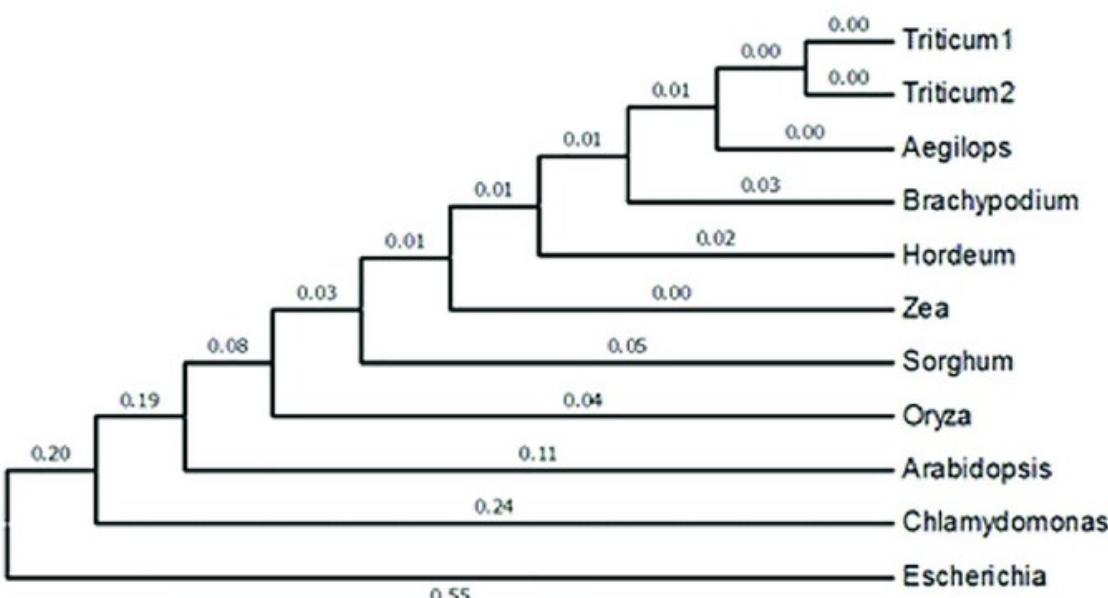
**THINK**

**Activity 07:**

<https://forms.gle/mE55miv68ShnsmPE8>

# Blast Also Works With Proteins!!

	1	10	20	30	40	50	60	70	80	90		
TuVipp1	M	H	N	L	F	O	R	L	T	R	V	V
TuVipp2	M	H	N	L	F	O	R	L	T	R	V	V
AtVipp1	M	H	N	L	F	E	R	F	S	R	V	V
CrVipp1	M	A	N	L	F	S	R	A	R	I	V	V
PspA	.	h	n	l	f	s	r	a	r	i	v	v
Consensus	.	M	H	N	L	F	O	R	T	R	V	V
	91	100	110	120	130	140	150	160	170	180		
TuVipp1	A	L	K	R	R	K	S	Y	A	N	R	S
TuVipp2	A	L	K	R	R	K	S	Y	A	N	R	S
AtVipp1	A	L	K	R	R	K	S	Y	A	N	R	S
CrVipp1	A	L	K	R	R	K	S	Y	A	N	R	S
PspA	A	L	K	R	R	K	S	Y	A	N	R	S
Consensus	A	L	K	R	R	K	S	Y	A	N	R	S
	181	190	200	210	220	230	240	250	26862			
TuVipp1	S	Q	A	R	E	A	L	G	Q	L	G	Q
TuVipp2	S	Q	A	R	E	A	L	G	Q	L	G	Q
AtVipp1	S	E	A	D	A	L	T	Q	D	Q	L	G
CrVipp1	R	Q	A	E	S	T	K	M	L	Y	G	D
PspA	R	A	E	S	-	-	-	-	-	-	-	-
Consensus	a	#	A	E	s	.....	g	.d	l	#	k	...





# 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			
Sbjct	1	MSVKHAISRGLELGRSFLQIGLLKSGGRVAAKLRADRFRVGPSVRTV			
Query	61	RTSLRGLAAQLQSAGFRRRTGASPRNRAVFLAFLGLGVGLIEQQLE+			
Sbjct	61	RTSLRGLAAQLQSAGFRRRTGASPRNRAVFLAFLGLGVGLIEQQLE+			
Query	121	VFKKKKIQSTLRPFTSGFKLEDYVIGNQIGKGSNAAVYEAAAQFAHP			
Sbjct	121	VFKKKKIQSTLRPFTSGFKLEDYVIGNQIGKGSNAAVYEAAAQF+HP			
Query	181	DNEVEVQNVRSAACCSLRNFPLAIKMLWNFGAGSSSEAILKSMSQEI			
Sbjct	181	DNEEVQNVRSPSCCSLRNFPLAIKMLWNFGAGSSSEAILKSMSQEI			
Query	241	HITLDGHGVLPKRVSAHPNVIRVYRAFTADVPLLPGAEEYPDVLF			
Sbjct	241	ITLDG FGVL+RVSAHPNVIRVYRAFTADVPLLPGA+EEYPDVLF			
Query	301	LFLVMKNYPYTLRQYLQVSTPNRRQGSLMVLQLL	334		
Sbjct	301	LFLVMKNYP TLRQYLQVSTPNRRQGSLMVLQLL	334		
Query	301	LFLVMKNYPCTLRQYLQVSTPNRRQGSLMVLQLL	334		

# The central dogma of molecular biology



Transcription



Translation

Protein  
N-terminus

C-terminus

H<sub>2</sub>N — Met Ser Tyr Lys Cys Val — COOH



# More About Silent Mutations

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

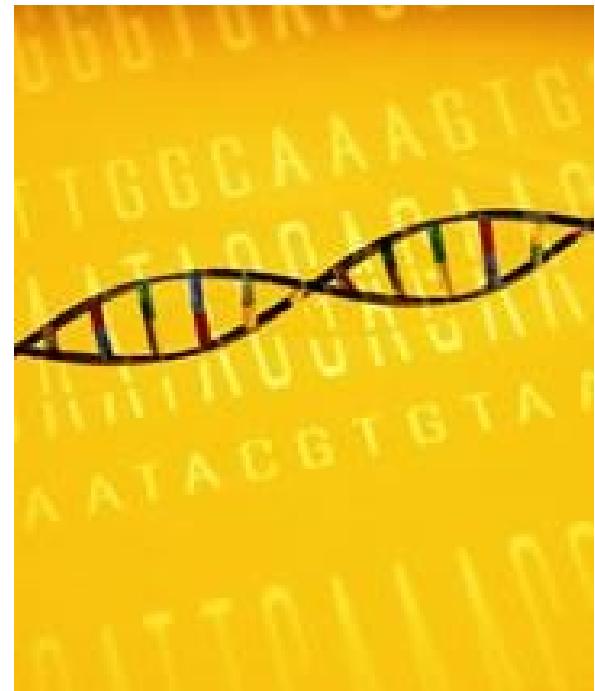
No mutation		Point mutations		
		Silent	Nonsense	Missense
				conservative non-conservative
DNA level	TTC	TTT	ATC	TCC TGC
mRNA level	AAG	AAA	UAG	AGG ACG
protein level	Lys	Lys	STOP	Arg Thr

[https://en.wikipedia.org/wiki/Silent\\_mutation](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>



ALLEGHENY  
COLLEGE

# Bring the Tool!



**Up Next!**

# *dbDSM*

*Database of Deleterious Synonymous Mutation*

Home      Search      Download      About      Submit      Contact us

Search for “cystic fibrosis”

Disease ▾

eg. : Early-Onset Epilepsy; PCDH19; X:100342042; c.954G>C; p.Thr318=

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	<a href="#">DSM000732</a>

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	<a href="#">DOID:1485</a>
Gene	CFTR
GenID	<a href="#">1080</a>
MIM	MIM:602421
Map_Location	7q31.2
VariantType	germline
Protein	p.Lys163=
c.DNA	c.489G>A
SNPID	<a href="#">rs397508733</a>
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 acagcaggc agagaaaaag ggttgagcgg caggcaccca
61 gagtagtagg tctttggcat taggagcttgc agcccagacg gccctagcag ggaccccagc
121 gcccggagaga ccatgcagag gtcgcctctg gaaaaggcca gcgttgctc caaactttt
181 ttcaagcttggccat ccagaccaat tttgaggaaa ggatacagac agcgcccttggaa attgtcagac
241 atataccaaa tcccttctgt tgattctgct gacaatctat ctgaaaaatt ggaaagagaaa
301 tgggatagag agctggcttc aaagaaaaat cctaaactca ttaatgccct tcggcgatgt
361 ttttcttggaa gatttatgtt ctatggaaatc tttttatatt tagggaaatg caccggaaatc
421 gtacagcctc tcttacttggg aagaatcata gtttcctatg acccgatcaa caaggaggaa
481 cgctctatcg cgatttatct aggcataggc ttatgccttc tctttattgtt gaggacactg
541 ctcctacacc cagccatttt tggccttcat cacattggaa tgcagatgag aatagctatg
601 ttttagttt a tttataagaa gactttaaag ctgtcaagcc gtgttctaga taaaataagt
661 attggaaatc ttgttagtct ctttccaaac 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](https://www.ncbi.nlm.nih.gov/nuccore/NM_000492.3)



ALLEGHENY  
COLLEGE

# Awesome!

*(But not as awesome as this!)*





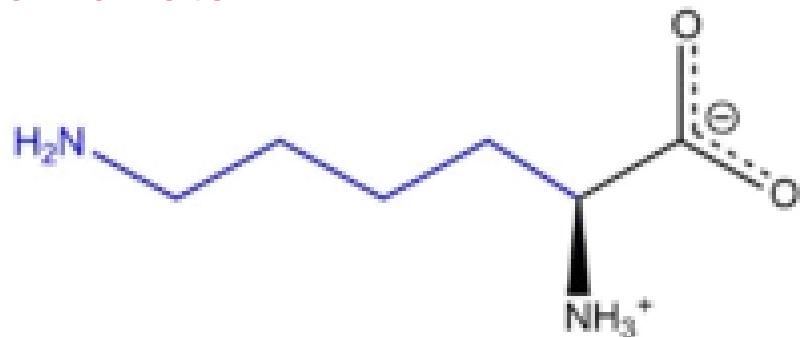
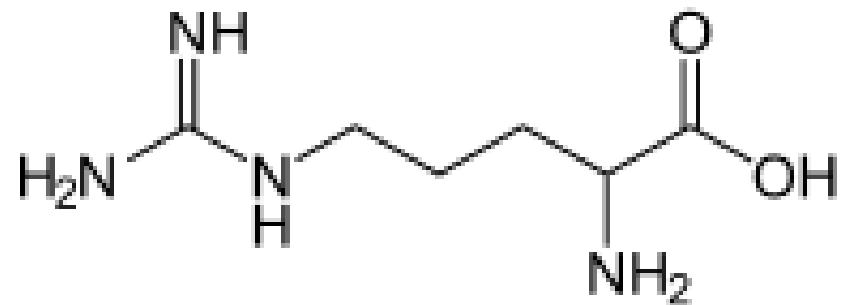
## Second letter

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



# 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

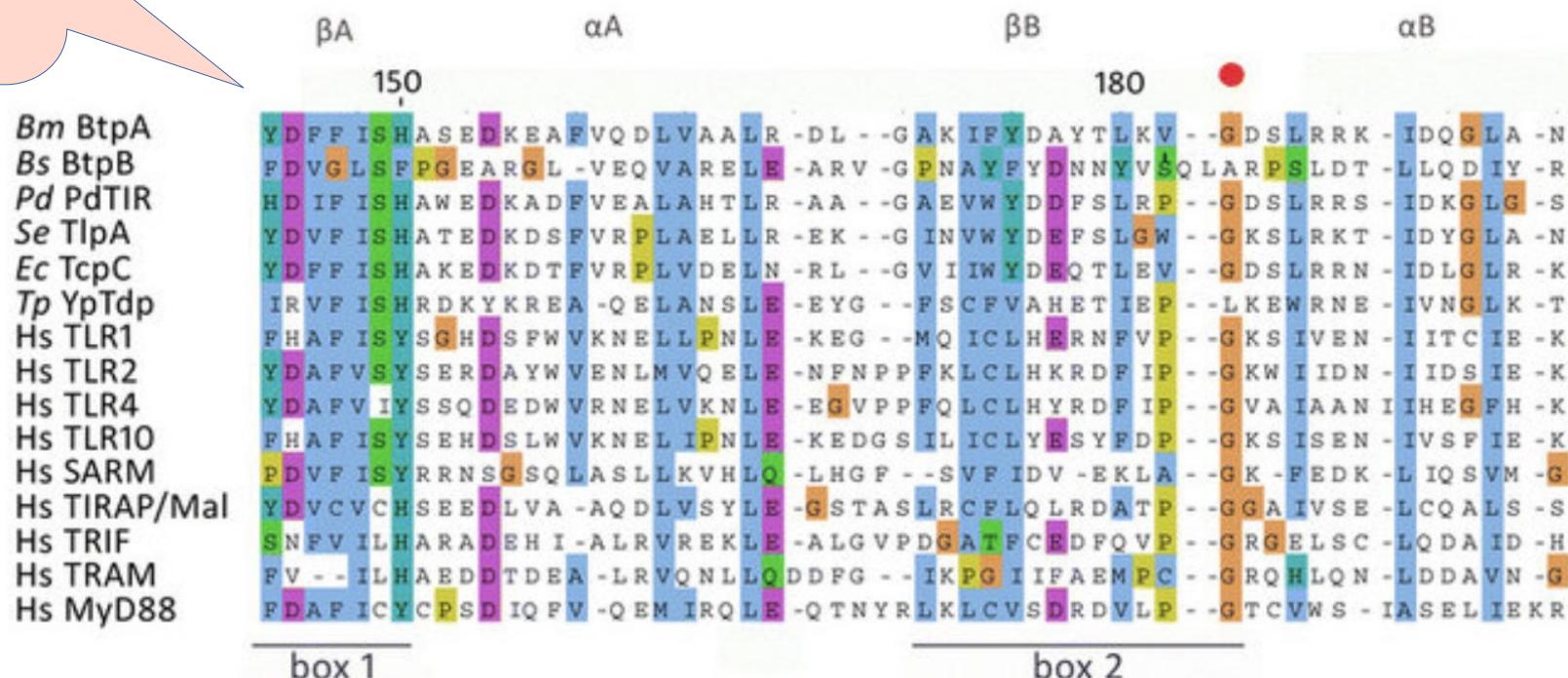
Wait! How did we know where silent mutations came from?!

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

These domains have individual functions



Change domain composition may change function



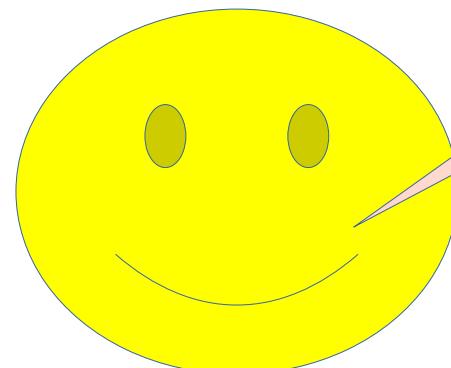
# When Comparing Proteins...

How are proteins different?

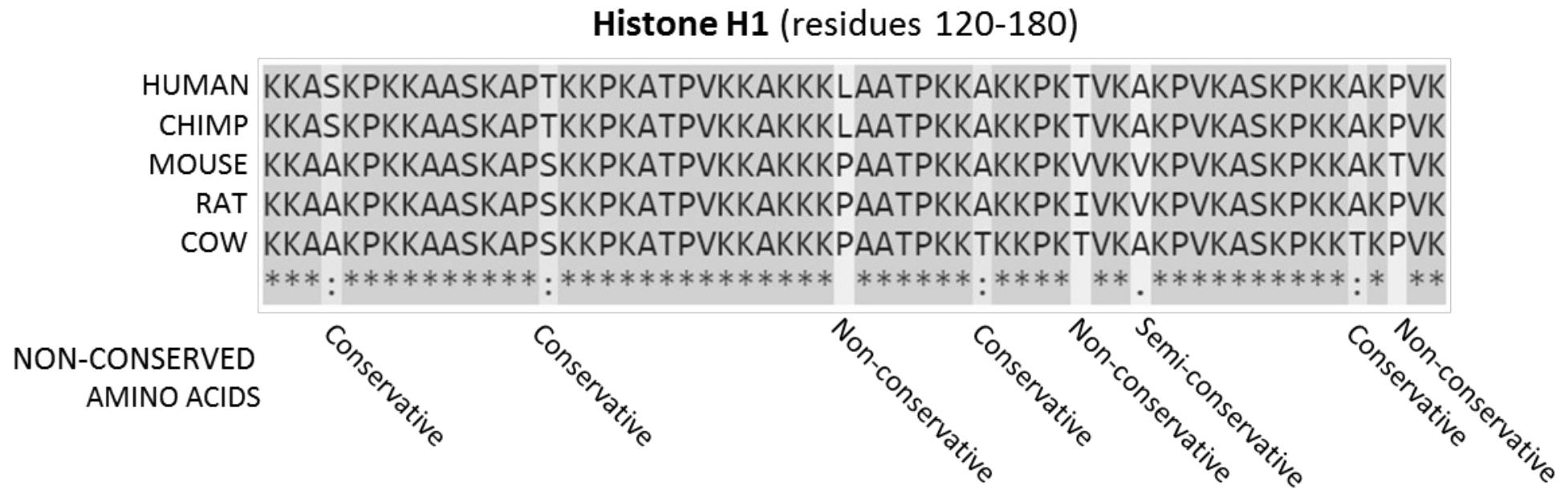
How much difference is there?

Was this difference due to chance?

Could the altered protein have a new and different function?



# Protein Amino Acid Replacements



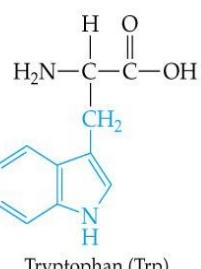
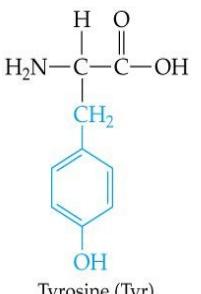
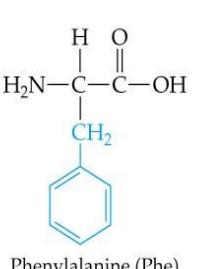
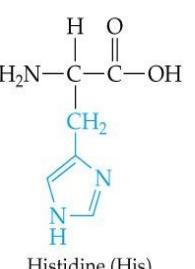
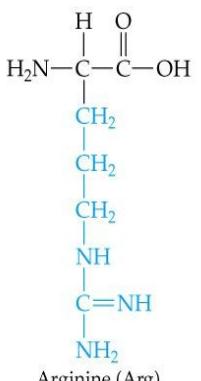
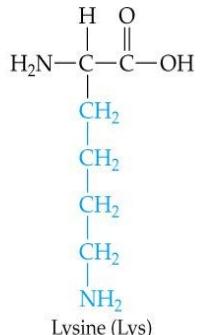
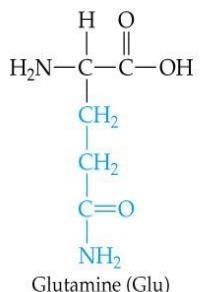
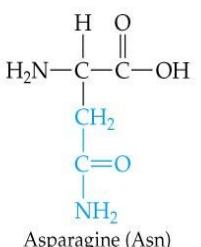
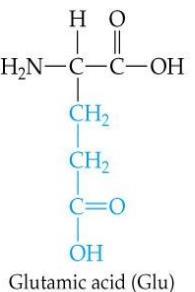
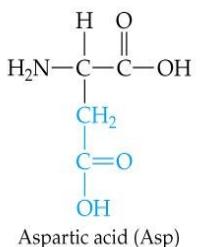
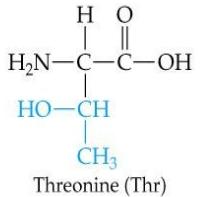
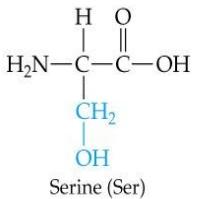
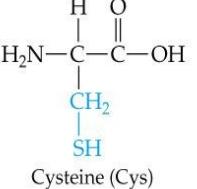
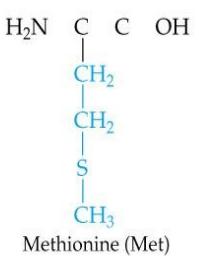
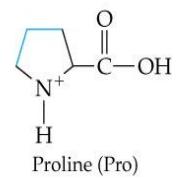
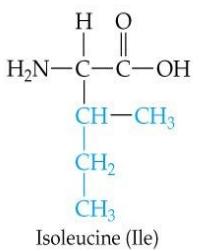
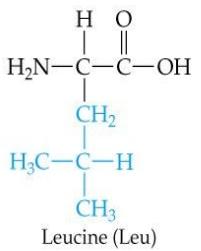
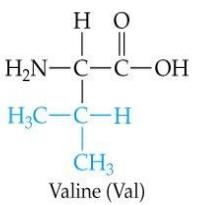
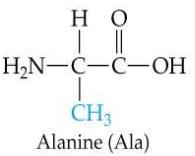
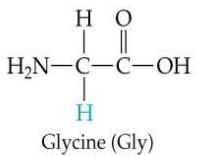
Generally, replacements are ...

- **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



ALLEGHENY  
COLLEGE

# Amino Acids





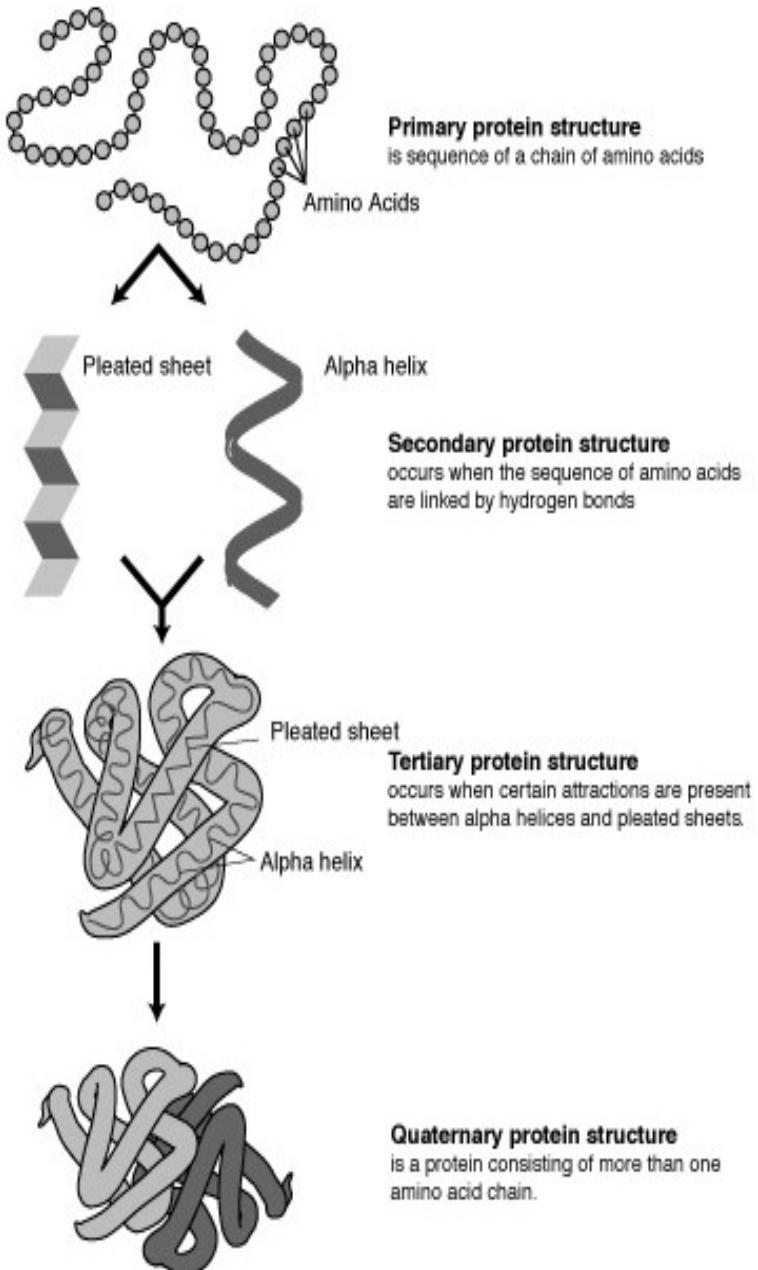
# Amino Acid Components

- **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>



ALLEGHENY  
COLLEGE

# 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](https://en.wikipedia.org/wiki/Point_accepted_mutation)

# The PAM Matrix

The probability calculations for Substitutions have been done for you!

		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	



# PAM Matrices

- Point Accepted Mutation
- 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

BLOSUM 80

PAM 1

*Less divergent*

BLOSUM 62

PAM 120



BLOSUM 45

PAM 250

*More divergent*

More similarity



Less similarity



# 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)



# BLOSUM matrix

## Heinkoff and Heinkoff, 1992

- **BLOcks SUbstitution 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
- $\lambda$  - 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 1<sup>st</sup> 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

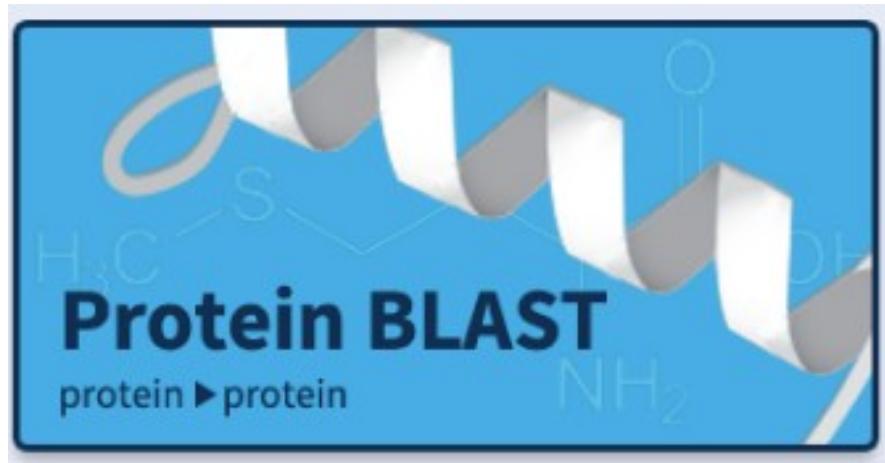
- Create N x M matrix
- Place each sequence along one axis
- Place score 0 at the up-left corner
- Fill in 1<sup>st</sup> row & column with gap penalty multiples
- Fill in the matrix with max value of 3 possible moves:
  - Vertical move: Score + gap penalty
  - Horizontal move: Score + gap penalty
  - 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.



ALLEGHENY  
COLLEGE

# Blast-Off!!

- Let's blast some protein sequences
- [https://blast.ncbi.nlm.nih.gov/Blast.cgi#dtr\\_Quer\\_y\\_98931](https://blast.ncbi.nlm.nih.gov/Blast.cgi#dtr_Quer_y_98931)



THINK



ALLEGHENY  
COLLEGE

# Blasting Proteins



**National Library of Medicine**  
*National Center for Biotechnology Information*



**COVID-19 is an emerging, rapidly evolving situation.**

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data](#)

Search NCBI

A search bar containing the text "sirt1", which is highlighted with a red rectangular border.A small blue button with a white "x" symbol.A blue button with the word "Search" in white.

Results found in 32 databases



# Blasting Proteins

Select Protein

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

## Proteins

Conserved Domains 9

Identical Protein Groups 570

Protein 6,651

Protein Family Models 45

Structure 129



[SIRT1 \[Mytilus coruscus\]](#)

2. 188 aa protein

Accession: CAC5409616.1 GI: 1866842361

[BioProject](#) [Nucleotide](#) [Taxonomy](#)

[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)



# Blasting Options

**Enter Query Sequence**

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

CAC5409616.1

Query subrange [?](#)

From  To

Or, upload file  No file chosen [?](#)

Job Title   
Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

**New columns added to the Description Table**   
Click 'Select Columns' or 'Manage Columns'.

**Choose Search Set**

Database: Non-redundant protein sequences (nr) [?](#)

Organism Optional: Enter organism name or id—completions will be suggested   exclude [Add organism](#)  
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Exclude Optional:  Models (XM/XP)  Non-redundant RefSeq proteins (WP)  Uncultured/environmental sample sequences

**Program Selection**

Algorithm:

- Quick BLASTP (Accelerated protein-protein BLAST)
- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm [?](#)



# Blasting Results

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



# 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](#) ▲ [Previous](#)

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		MESAEELPRKMAAQPLKNEPPVKRQKLNEEDDNDSD EQGCSNISKDNEAGNTE			
Sbjct 1		MESAEELPRKMAAQPLKNEPPVKRQKLNEEDDNDSD EQGCSNISKDNEAGNTE			
Query 61		IDNSDNCSEISNL SGLSEEAWKPTSGAMSWIHKQIMNGVNPRPILNGLIPDD			
Sbjct 61		IDNSDNCSEISNL SGLSEEAWKPTSGAMSWIHKQIMNGVNPRPILNGLIPDD			
Query 121		DFTLWKIVINIMSEPPP RKKL SHINTLQDV IQLLQNCKNIMVLTGAGVS VSC			
Sbjct 121		DFTLWKIVINIMSEPPP RKKL SHINTLQDV IQLLQNCKNIMVLTGAGVS VSC			
Query 181	MESMLALL	188			
Sbjct 181	MESMLALL	188			