

Bioinformatics

CS300

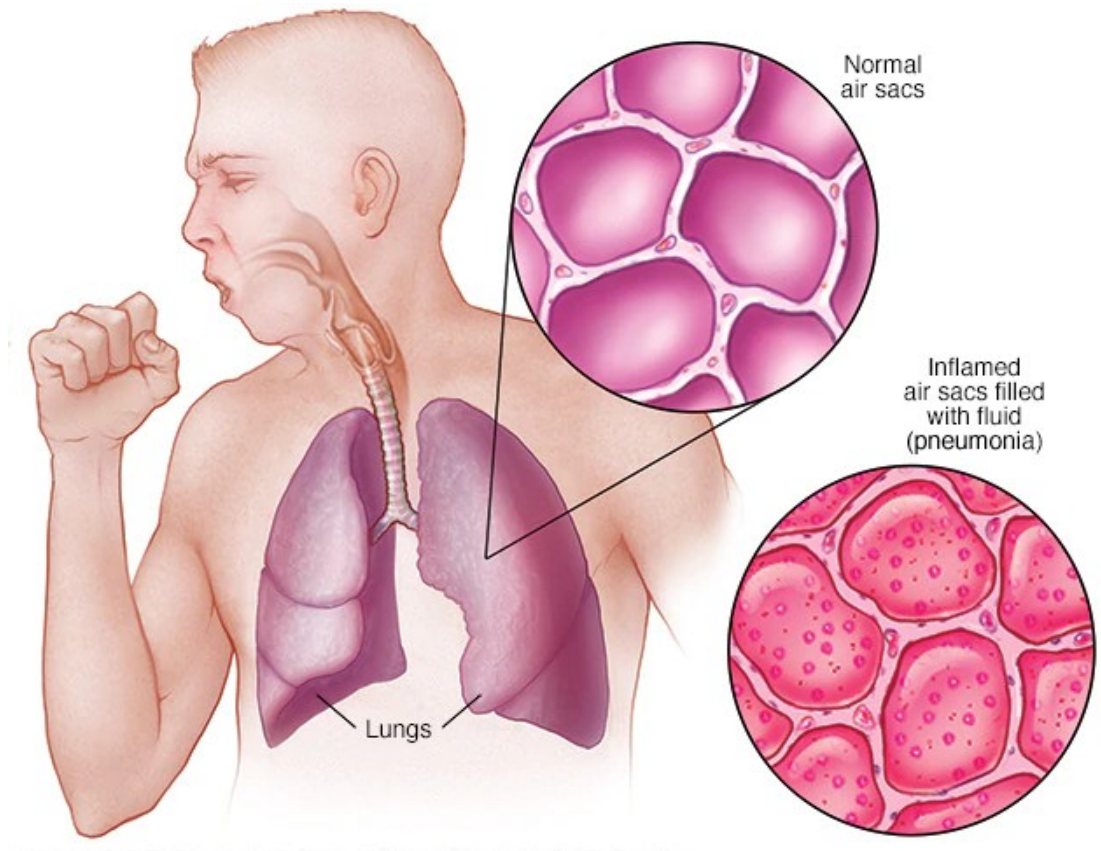
**Blast, Substitution Matrices and
Protein Alignments**
(Chap 4 and 5 in textbook)

Fall 2019

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.



- 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



Blast to Study HGT

- Locate the Accession number, **DQ355148.1**, on <http://www.pubmed.gov>
- *Streptococcus agalactiae* strain KMP104 transposon Tn917 rRNA methylase (ermB) gene, complete cds

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](#)

<https://www.ncbi.nlm.nih.gov/nuccore/87042723/>



Find the 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 \(Weighted\)](#)

Recent activity

 [Streptococcus
transposon Tn](#)

 [DQ355148.1](#)




Ah, The Sequence in FASTA Format

>DQ355148.1 Streptococcus agalactiae strain KMP104 transposon Tn917 rRNA methylase (ermB) gene, complete cds
ATGAACAAAAATATAAAATATTCTCAAACTTTTTAACGAGTGAAAAAGTACTCAACCAAATAATAAAAC
AATTGAATTTAAAAGAAACCGATACCGTTTACGAAATTGGAACAGGTAAAGGGCATTTAACGACGAAACT
GGCTAAAATAAGTAAACAGGTAACGTCTATTGAATTAGACAGTCATCTATTCAACTTATCGTCAGAAAAA
TTAAAACTGAACATTCGTGTCACCTTTAATTCACCAAGATATTCTACAGTTTCAATTCCCTAACAAACAGA
GGTATAAAATTGTTGGGAATATTCCTTACCATTTAAGCACACAAATTATTAAAAAAGTGGTTTTTTGAAAG
CCATGCGTCTGACATCTATCTGATTGTTGAAGAAGGATTCTACAAGCGTACCTTGGATATTCACCGAACA
CTAGGGTTGCTCTTGACACTCAAGTCTCGATTCAGCAATTGCTTAAGCTGCCAGCGGAATGCTTTTCATC
CTAAACCAAAAGTAAACAGTGTCTTAATAAACTTACCCGCCATACCACAGATGTTCCAGATAAATATTG
GAAGCTATATACGTACTTTGTTTCAAATGGGTCAATCGAGAATATCGTCAACTGTTTACTAAAAATCAG
TTTCATCAAGCAATGAAACACGCCAAAGTAAACAATTTAAGTACCGTTACTTATGAGCAAGTATTGTCTA
TTTTTAATAGTTATCTATTATTTAACGGGAGGAAATAA



Blast Website

 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](#)


NEWS

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...](#)

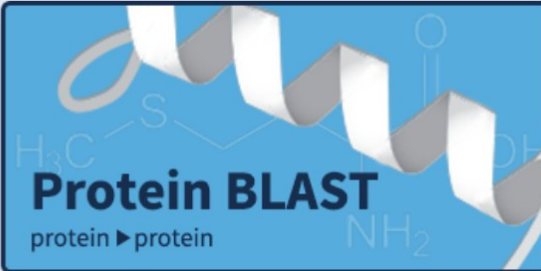
Web BLAST



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
ATGAACAAAAATATAAAATATTCTCAAAACTTTTTAACGAGTGAAAAAGTACTCAACCAAATAATAAAAC
AATTGAATTTAAAGAAACCGATACCGTTTACGAAATTGGAACAGGTAAAGGGCATTTAACGACGAAACT
GGCTAAAATAAGTAAACAGGTAACGTCTATTGAATTAGACAGTCATCTATTCAACTTATCGTCAGAAAAA
TTAAAACTGAACATTCGTGTCACTTTAATTCACCAAGATATTCTACAGTTTCAATTCCCTAACAAACAGA
GGTATAAAATTGTTGGGAATATTCCTTACCATTTAAGCACACAAATTATTAATAAAAGTGGTTTTTGAAAG
CCATGCGTCTGACATCTATCTGATTGTTGAAGAAGGATTCTACAAGCGTACCTTGATATTACCGAACA
```

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

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

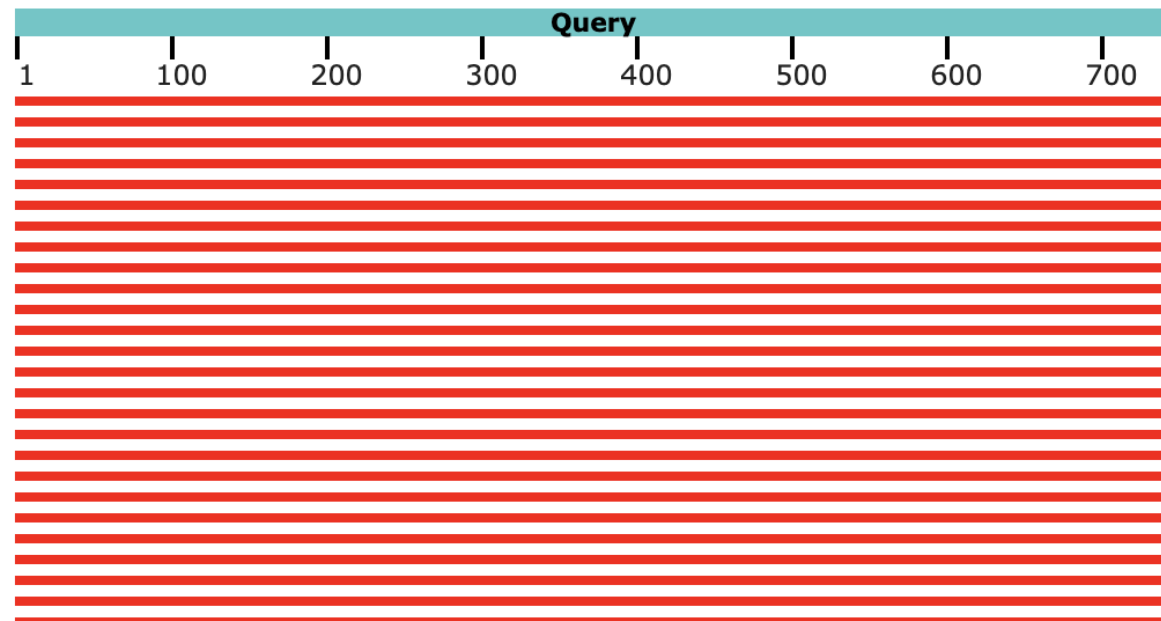
 50 - 80

 80 - 200

 >= 200

100 sequences selected 

Distribution of the top 111 Blast Hits on 100 subject sequences



Results

Descriptions

Graphic Summary

Alignments

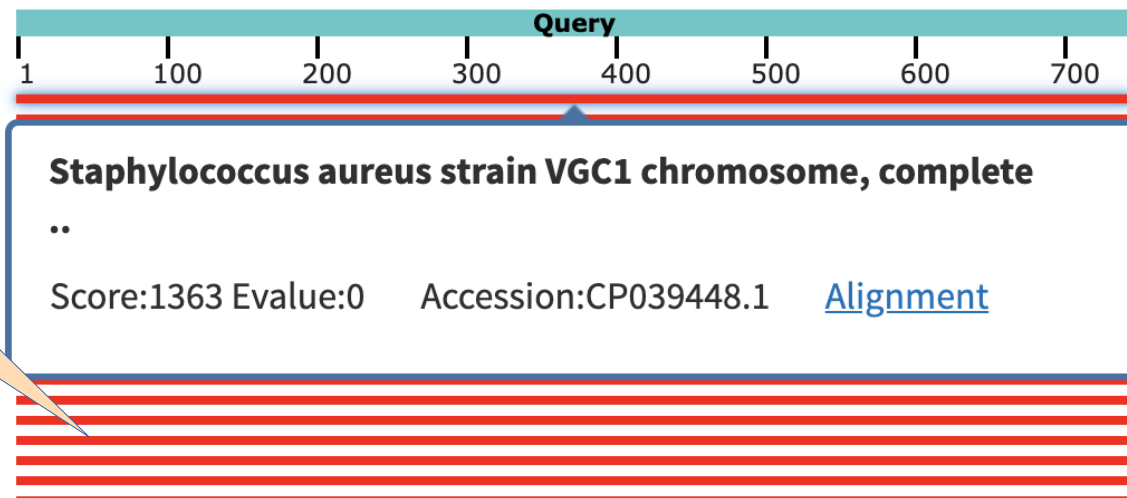
Taxonomy

🖱 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





Back to 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.

Blast is Cool!





Your Turn to Investigate!!!

- Investigate a gene of resistance: *ermA*
- Questions:
 - What does this gene do? (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?



THINK

GitHub Activity Repository:

<https://classroom.github.com/a/DuBJW7yi>

Due at 12:15 on 17 Oct. 2019

Make a directory: **act1**

Workfile: **act1/blastWork.md**



Predicted Protein Results?

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		MSVKHAISRGLLEGRSFLQIGLLKSGGRVAAKLRA DRFRVGPSVRTV		
Sbjct 1		MSVKHAISRGLLEGRSFLQIGLLKSGGRVAAKLRA DRFRVGPSVRTV		
Query 61		RTSLRGLAAQLQSAGFRRRFTGASPRNRAVFLAFGLGVGLIEQQLEE		
Sbjct 61		RTSLKGLAAQLQSAGFRRRFTGASPRNRAVFLAFGLGVGLIEQQLED		
Query 121		VFKKKKIQSTLRPFTSGFKLEDYVIGNQIGKGSNAAVYEAAAQFAHF		
Sbjct 121		VFKKKKIQSTLRPFTSGFKLEDYVIGNQIGKGSNAAVYEAAAQFSHE		
Query 181		DNEVEVQNVRSAACCSLRNFPLAIKMLWNFGAGSSSEAILKSMSQEI		
Sbjct 181		DNEEEVQNVRSPSCCSLRNFPLAIKMLWNFGAGSSSEAILKSMSQEI		
Query 241		HITLDGHFGVLPKRVS AHPNVIRVYRAFTADVPLLPGAEEEEYPDVLE		
Sbjct 241		QITLDGRFGVLP RRVSAHPNVIRVYRAFTADVPLLPGAQEEYPDVLE		
Query 301		LFLVMKNYPYTLRQYLQVSTPNRRQGSLMVLQLL	334	
Sbjct 301		LFLVMKNYPCTLRQYLQVSTPNRRQGSLMVLQLL	334	

The central dogma of molecular biology



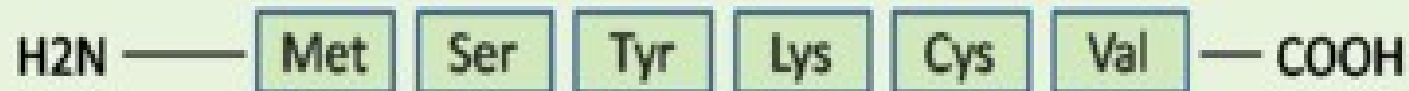
Transcription



Translation

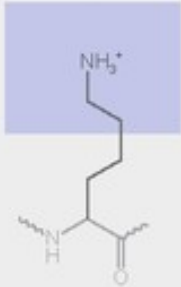
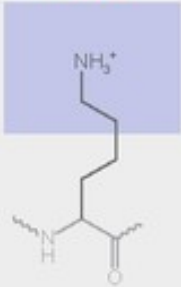
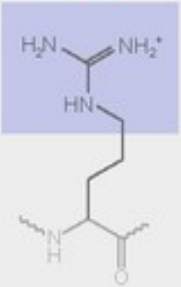
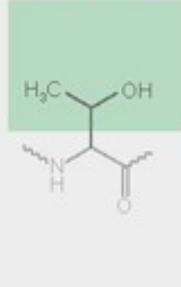
Protein
N-terminus

C-terminus



More About Silent Mutations

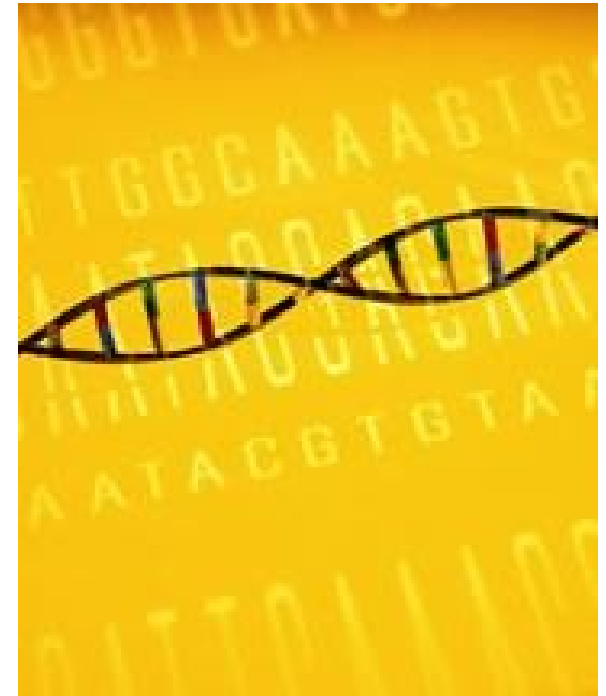
- Redundant codons mean ~1/3 of DNA mutations often do not alter protein sequence

	Point mutations				
	No mutation	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
					
	basic	basic		basic	polar

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 has still has 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

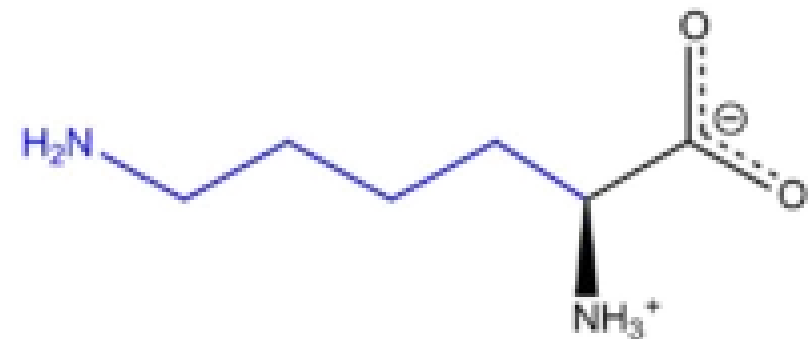
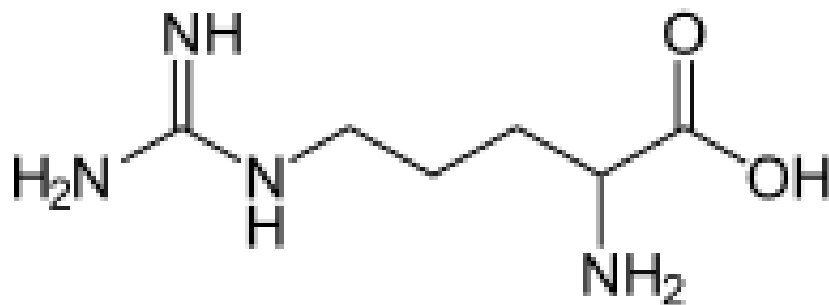


		Second letter				
		U	C	A	G	
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp	U C A G
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C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

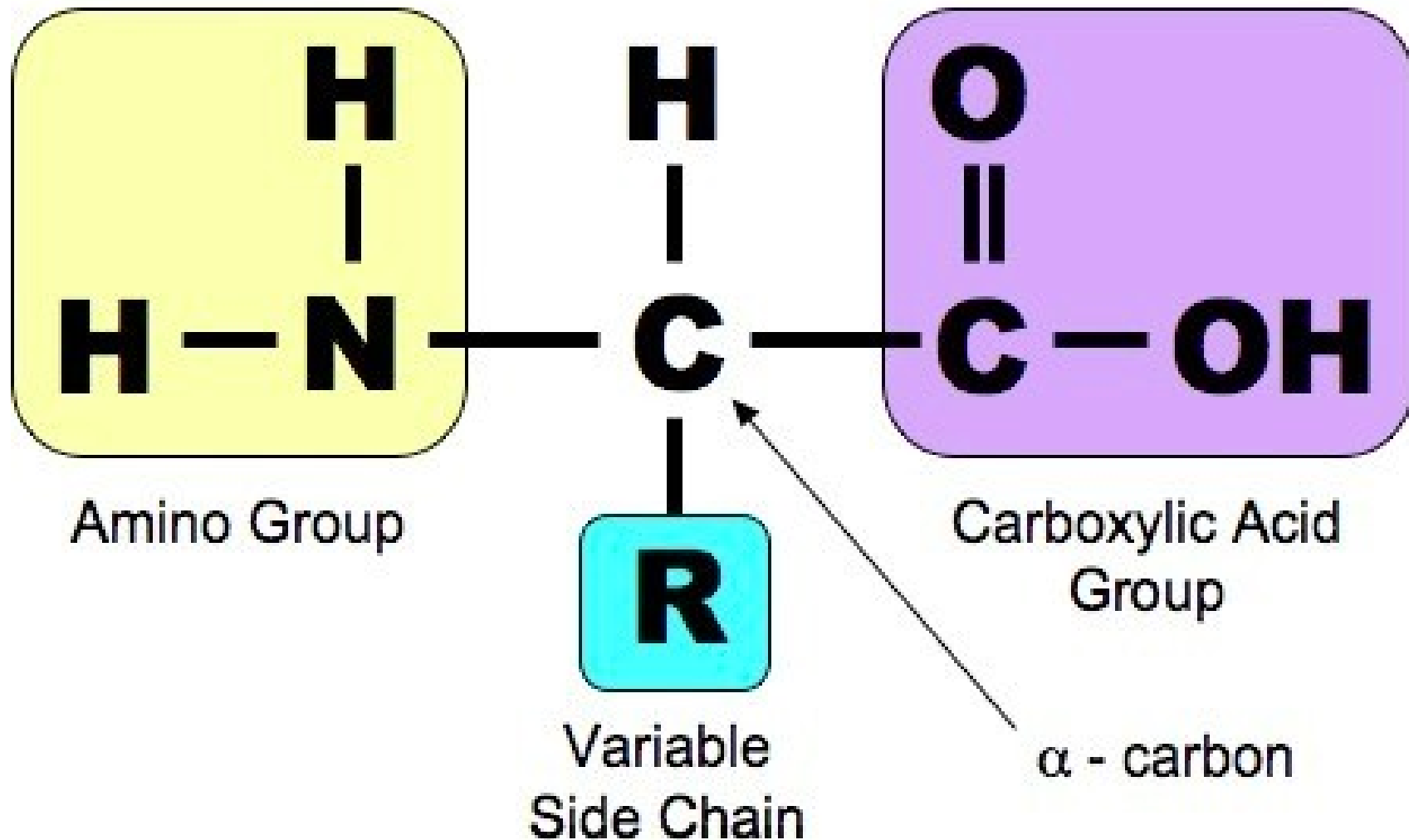
Third letter

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 significance
- Amino acid changes are not equally harmful to protein structure
 - Chemical complexes being replaced by similar chemical complex.
 - Ex: Arginine (Arg) and Lysine (Lys)

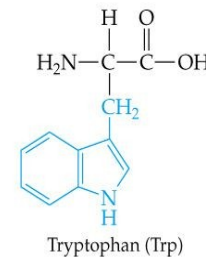
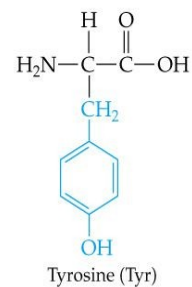
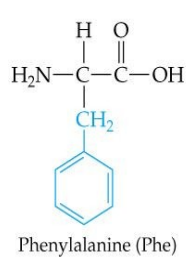
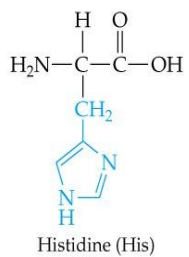
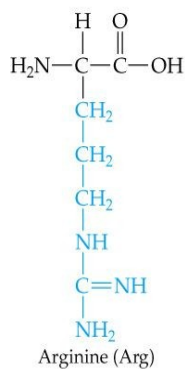
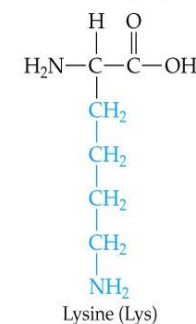
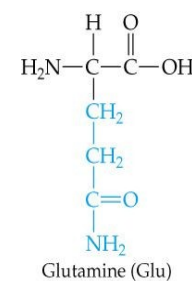
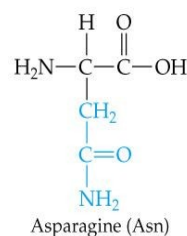
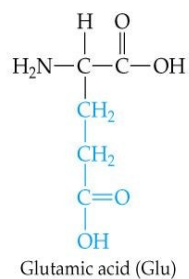
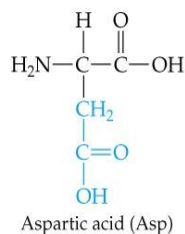
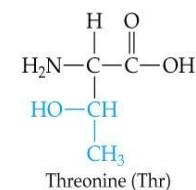
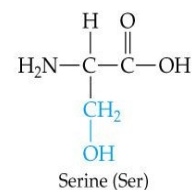
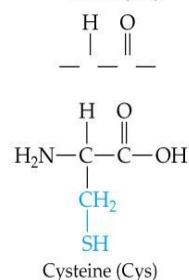
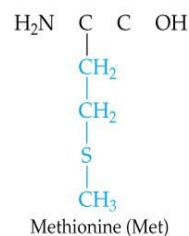
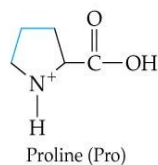
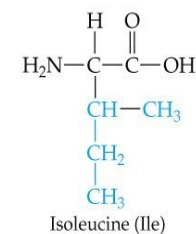
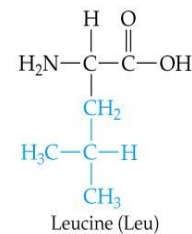
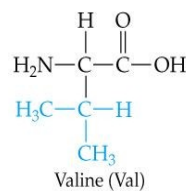
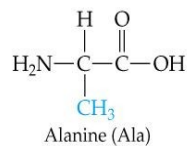
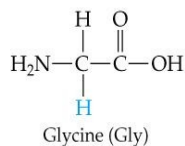


Amino Acid Substitutions





Amino Acid Substitutions

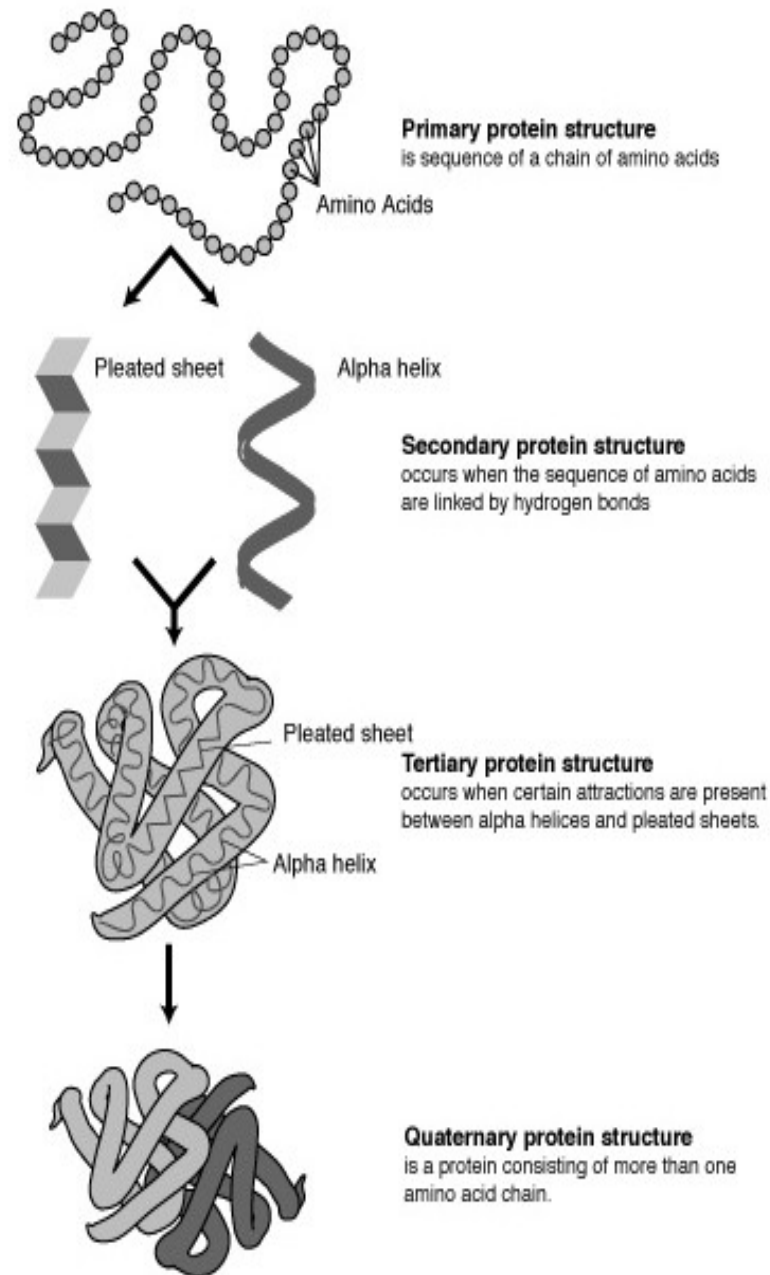




Amino Acid Components

- **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 Acid 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 mutations

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

		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



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

BLOSUM 80

PAM 1

Less divergent

BLOSUM 62

PAM 120

BLOSUM 45

PAM 250

More divergent





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



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
- 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 $N \times M$ 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:

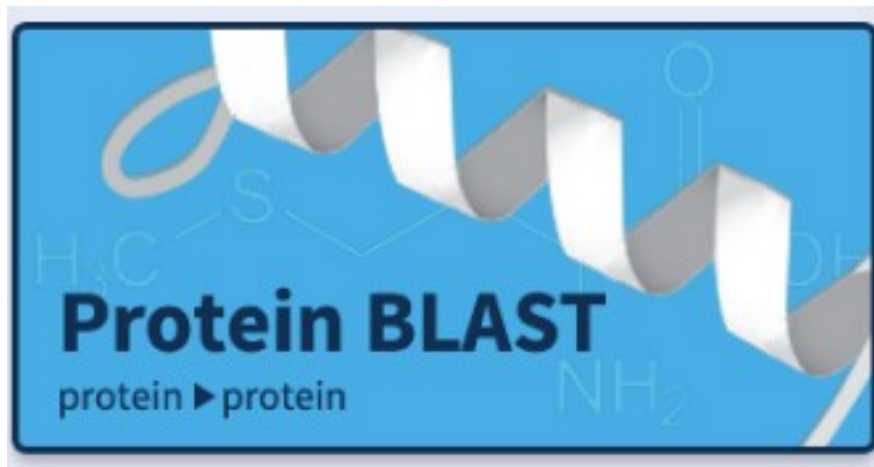
Protein Alignment – Chap 5

- Create $N \times M$ 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



THINK