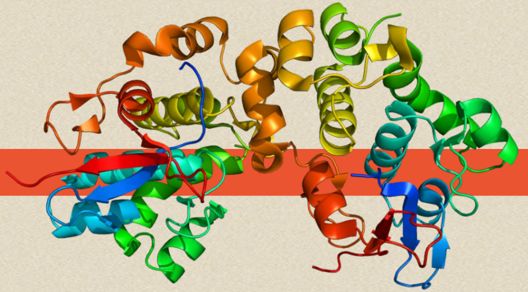
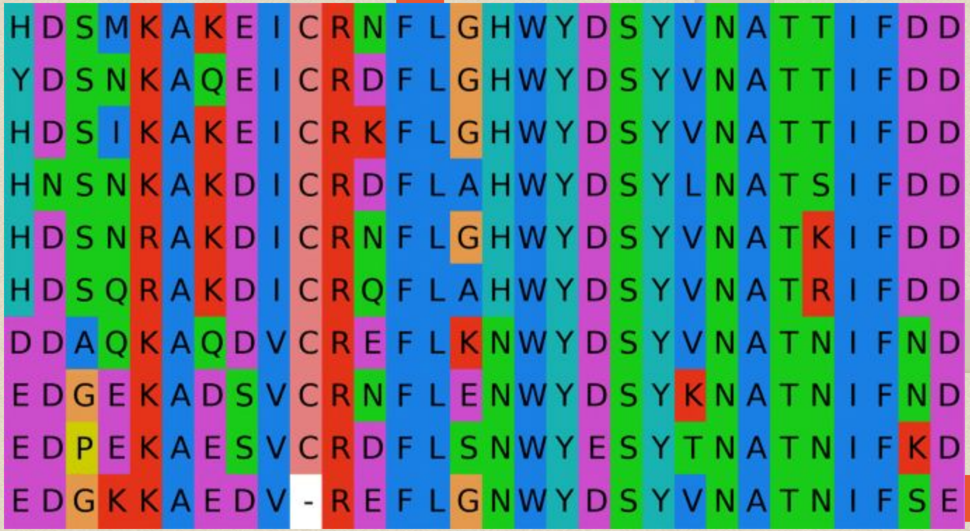


Evolutionary Multiple Sequence Alignment

Sanjana Bhagavatula, John Drohan, Rachel Utama, Sreevatsa Nukala

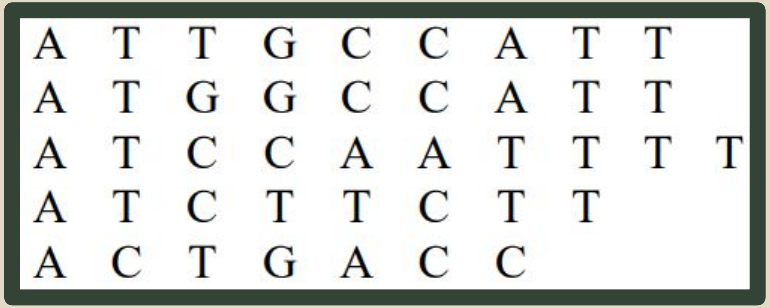
DS3500 Final Project



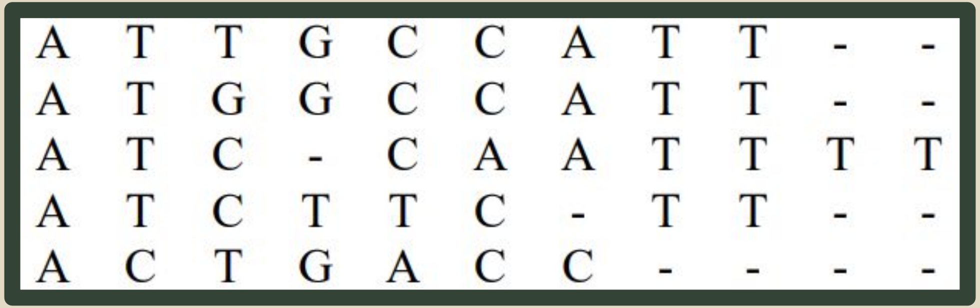
```
H D S M K A K E I C R N F L G H W Y D S Y V N A T T I F D D
Y D S N K A Q E I C R D F L G H W Y D S Y V N A T T I F D D
H D S I K A K E I C R K F L G H W Y D S Y V N A T T I F D D
H N S N K A K D I C R D F L A H W Y D S Y L N A T S I F D D
H D S N R A K D I C R N F L G H W Y D S Y V N A T K I F D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

Multiple sequence alignment (MSA)

- Alignment of 3 or more biological sequences of similar length



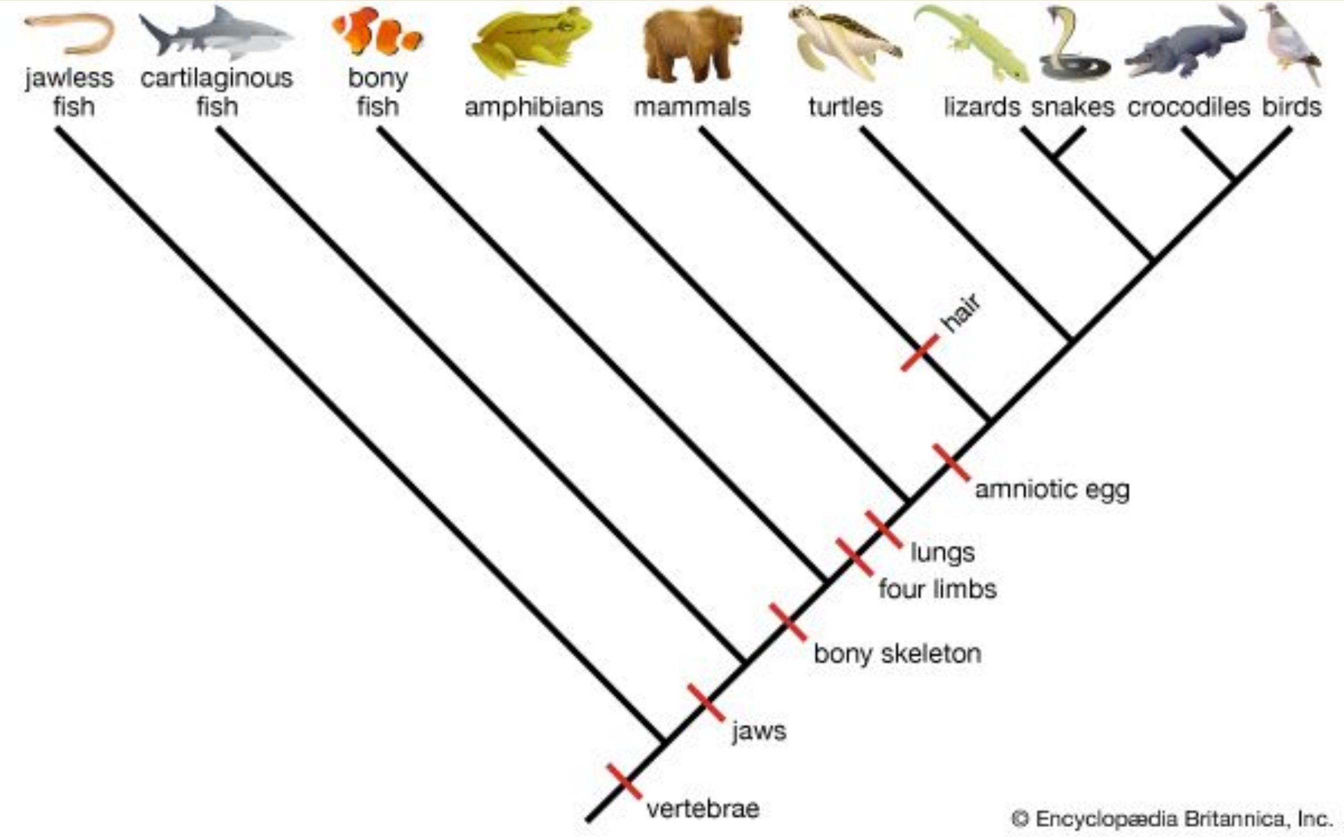
A	T	T	G	C	C	A	T	T	
A	T	G	G	C	C	A	T	T	
A	T	C	C	A	A	T	T	T	T
A	T	C	T	T	C	T	T		
A	C	T	G	A	C	C			



A	T	T	G	C	C	A	T	T	-	-
A	T	G	G	C	C	A	T	T	-	-
A	T	C	-	C	A	A	T	T	T	T
A	T	C	T	T	C	-	T	T	-	-
A	C	T	G	A	C	C	-	-	-	-

- Process of minimizing gaps and mismatches
- Detect regions of variability or conservation in a family of proteins
- May indicate functional, structural, or evolutionary relationships between biological sequences

Phylogenetic Tree

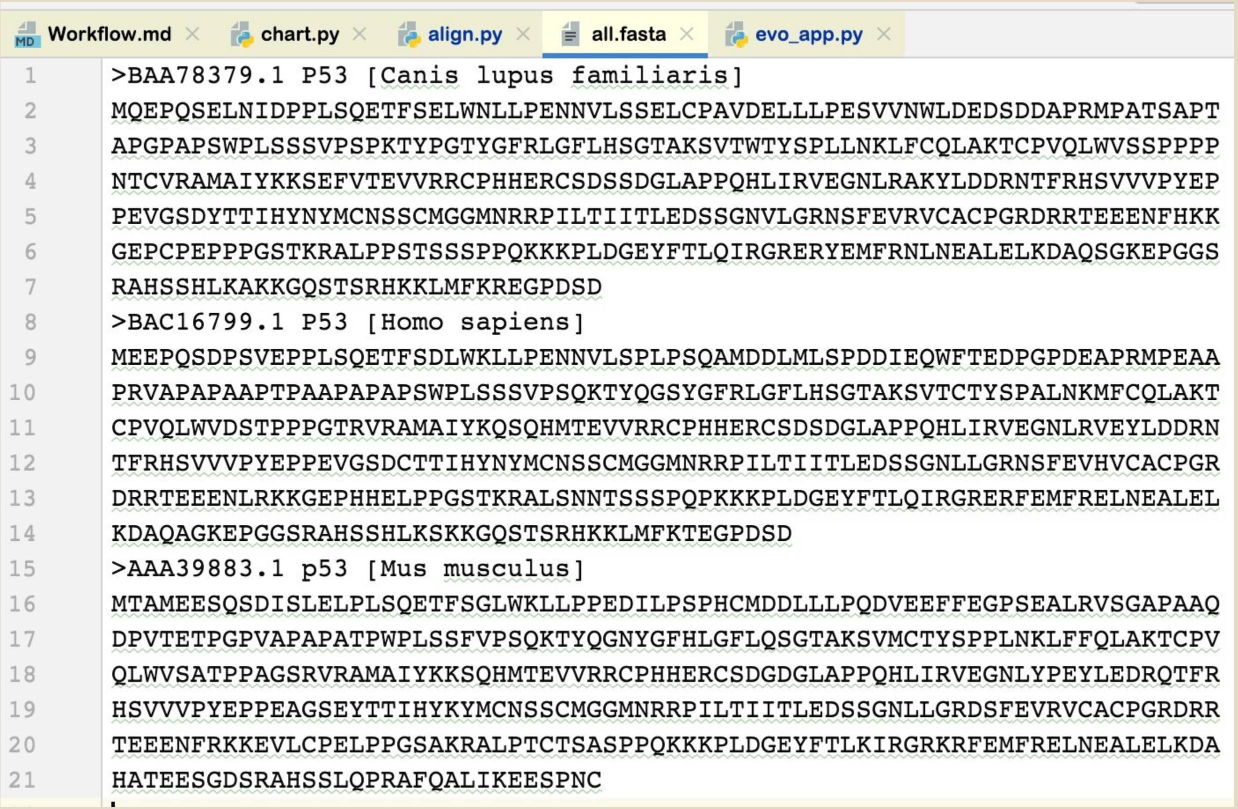


© Encyclopædia Britannica, Inc.

Implementation

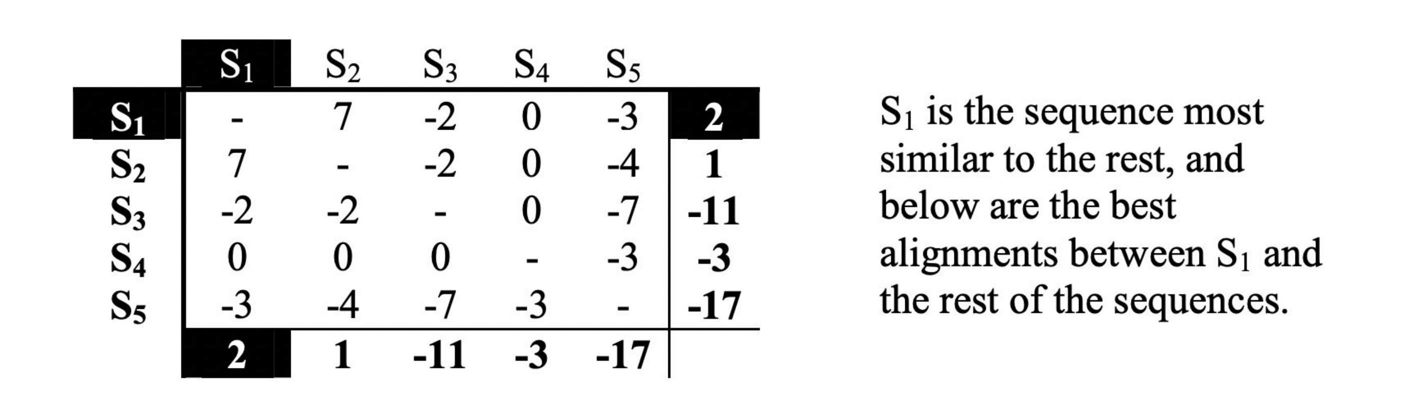
Evolutionary Multiple Sequence Alignment

Read Fasta File



```
>BAA78379.1 P53 [Canis lupus familiaris]
MQEPQSELNIDPPLSQETFSELWNLLPENNVLSSELCPAVDELLLPESVVNWLDEDSDDAPRMPATSAPT
APGPAPSWPLSSSVPSPKTYPGTYGFRLGFLHSGTAKSVTWTYSPLLNKLFCQLAKTCPVQLWVSSPPPP
NTCVRAMAIYKKSEFVTEVVRRCPHHERCSDSSDGLAPPQHLIRVEGNLRAKYLDDRNTFRHSVVVPYEP
PEVGSDYTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNVLGRNSFEVRVCACPGRDRRTEEENFHKK
GEPCPEPPPGSTKRALPPSTSSSPPQKKKPLDGEYFTLQIRGRERYEMFRNLNEALELKDAQSGKEPGGS
RAHSSHLKAKKGQSTSRHKKLMFKREGPDSD
>BAC16799.1 P53 [Homo sapiens]
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAA
PRVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKT
CPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRN
TFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVHVCACPGR
DRRTEEENLRKKGEPHHELPPGSTKRALSNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALEL
KDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD
>AAA39883.1 p53 [Mus musculus]
MTAMEESQSDISLELPLSQETFSGLWKLLPPEDILPSPHCMDDLLLPQDVEEFFEGPSEALRVSGAPAAQ
DPVTETPGPVAPAPATPWPLSSFVPSQKTYQGNYGFHLGFLQSGTAKSVMCTYSPPLNKLFFQLAKTCPV
QLWVSATPPAGSRVRAMAIYKKSQHMTEVVRRCPHHERCSDGDGLAPPQHLIRVEGNLYPEYLEDRQTFR
HSVVVPYEPPEAGSEYTTIHYKYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRDSFEVRVCACPGRDRR
TEEENFRKKEVLCPELPPGSAKRALPTCTSASPPQKKKPLDGEYFTLKIRGRKRFEMFRELNEALELKDA
HATEESGDSRAHSSLQPRAFQALIKEESPNC
```

Sum-Pairs Score Matrix



	S_1	S_2	S_3	S_4	S_5	
S_1	-	7	-2	0	-3	**2**
S_2	7	-	-2	0	-4	**1**
S_3	-2	-2	-	0	-7	**-11**
S_4	0	0	0	-	-3	**-3**
S_5	-3	-4	-7	-3	-	**-17**
	2	**1**	**-11**	**-3**	**-17**	

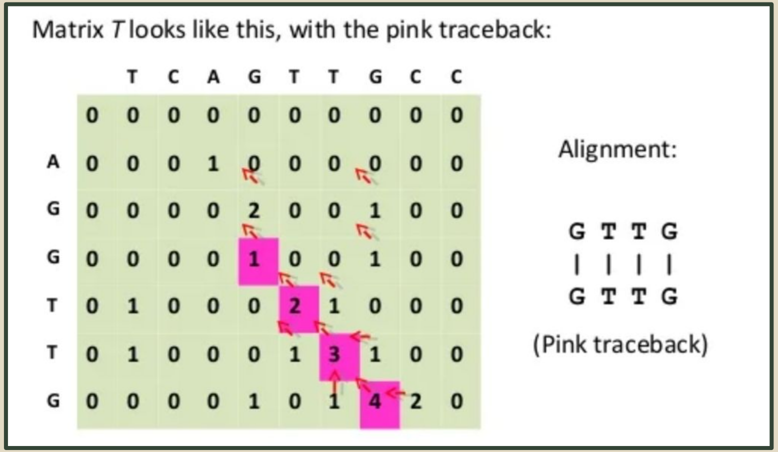
S_1 is the sequence most similar to the rest, and below are the best alignments between S_1 and the rest of the sequences.

Implementation

Evo framework

- Fitness criteria
  - Sum-pairs scores
  - BLOSUM Matrix
- Modification Agents
  - Smith waterman algorithm
- Evolve

- Visualizations
- Alignment Library



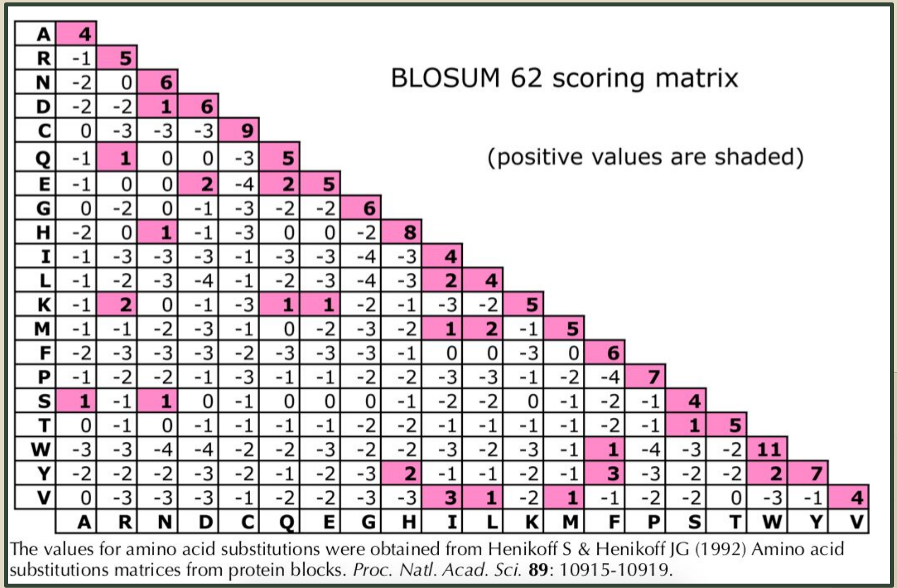
Matrix *T* looks like this, with the pink traceback:

		T	C	A	G	T	T	G	C	C
	0	0	0	0	0	0	0	0	0	0
A	0	0	0	1	0	0	0	0	0	0
G	0	0	0	0	2	0	0	1	0	0
G	0	0	0	0	1	0	0	1	0	0
T	0	1	0	0	0	2	1	0	0	0
T	0	1	0	0	0	1	3	1	0	0
G	0	0	0	0	1	0	1	4	2	0

Alignment:

```
G T T G
| | | |
G T T G
```

(Pink traceback)



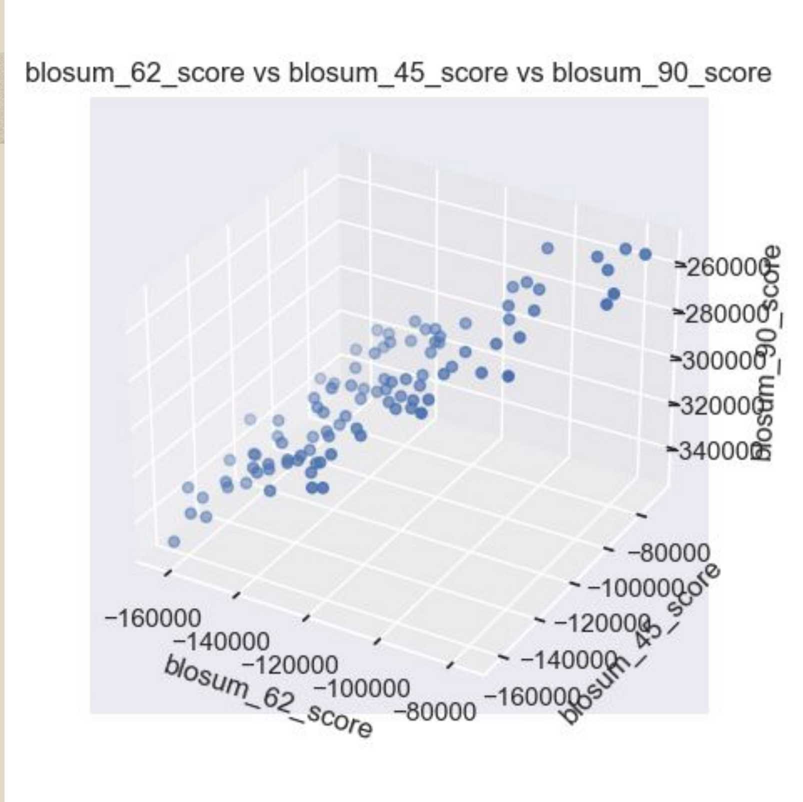
BLOSUM 62 scoring matrix

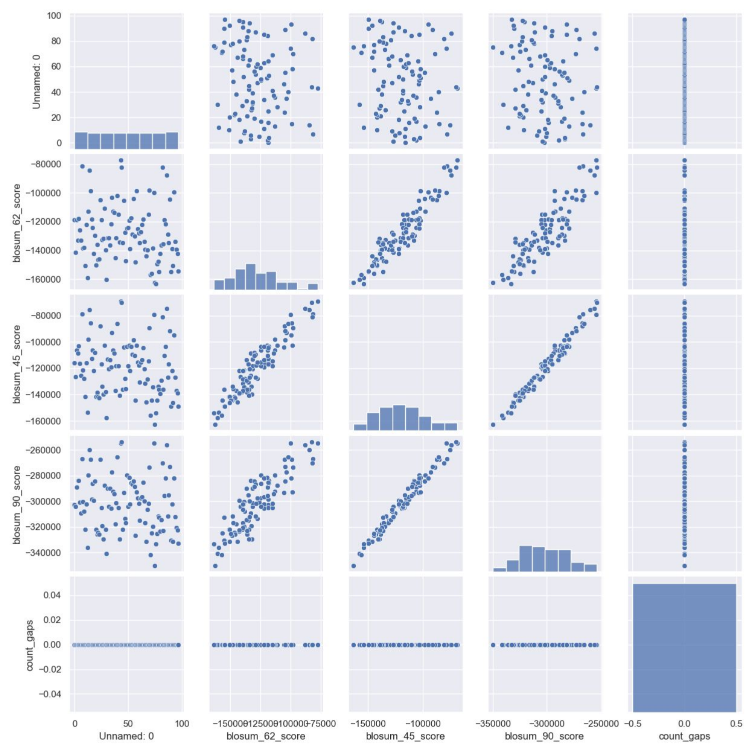
(positive values are shaded)

A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

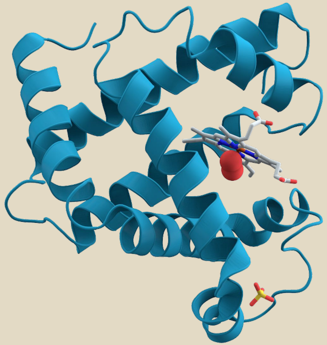
The values for amino acid substitutions were obtained from Henikoff S & Henikoff JG (1992) Amino acid substitutions matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89**: 10915-10919.

Preliminary Results





Potential Next Steps



1) **Add more Agents**
    a) Modification Agents -
        i) Needleman-Wunsch
2) **Add more Fitness Criteria**
3) **Convert Amino acid to numerical representation**
    a) Enable faster run time
    b) More capability to compare amino acids on a spectrum

References

1. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-3526-6
2. https://www.ebi.ac.uk/Tools/msa/
3. https://www.cs.princeton.edu/~mona/Lecture/msa1.pdf
4. https://d1w9csuen3k837.cloudfront.net/Pictures/2000x2000fit/7/9/6/138796_Hydrophobic-and-polar-amino-acids.jpg
5. https://d1w9csuen3k837.cloudfront.net/Pictures/2000x2000fit/7/9/6/138796_Hydrophobic-and-polar-amino-acids.jpg

Thank you!

Does anyone have any questions?