

BLAST

**Slides adapted & edited from a set by
Cheryl A. Kerfeld (UC Berkeley/JGI) &
Kathleen M. Scott (U South Florida)**

Kerfeld CA, Scott KM (2011) Using BLAST to Teach “E-value-tionary” Concepts.
PLoS Biology 9(2):e1001014

1

Starts with a Query Sequence in FASTA Format

Amino acid sequence:

```
>ribosomal protein L7/L12 [Thiomicrospira crunogena XCL-2]  
MAITKDDILEAVANMSVMEVVELVEAMEEKFGVSAAAVAVAGPAGDAGAA  
GEEQTEFDVVLTGAGDNKVAATKAVRGATGLGLKEAKSAVESAPFTLKEG  
VSKEEAETLANELKEAGIEVEVK
```

Nucleotide sequence:

```
>gi|118139508:333094-333465 Thiomicrospira crunogena XCL-2  
ATGGCAATTACAAAAGACGATATTTAGAACGAGTTGCTAACATGTCAGTAATGGAAG  
TTGTTGAACCTGTTGAAGCAATGGAAGAGAAGTTGGTGTTCGTCAGCAGCAGTTGC  
GGTTGCAGGTCCTGCAGGTGATGCTGGCGCTGCTGGTGAAGAACAAACAGAGTTTGAC  
GTTGTCTTGACTGGTGTGTTGACAACAAAGTTGCAGCAATCAAAGCCGTTTCGTGGCG  
CAACTGGCTCTGGGCTTAAAGAAGCGAAAAGTGCAGTTGAAAGTGCACCATTACGCT  
TAAAGAGGGTGTTCCTAAAGAAGAAGCAGAAAAGTCTTGCAAAATGAGCTTAAAGAAGCA  
GGTATTGAAGTCGAAGTTAAATAA
```

Note the description line
Starts with “>”, ends with carriage return
Not read as sequence data

Kerfeld and Scott, *PLoS Biology* 2011

2

2

NCBI BLAST Interface (blastp: for protein-protein alignments)

The screenshot shows the NCBI BLAST interface for protein-protein alignments. The main heading is "NCBI BLAST Interface (blastp: for protein-protein alignments)". The interface includes a navigation bar with "Home", "Recent Results", "Saved Strategies", and "Help". The main content area is titled "Enter Query Sequence" and contains a large yellow box with the text "(Paste FASTA format sequence here)". Below this, there are fields for "From" and "To" (query subrange), an "Or, upload file" section with a "Browse..." button, and a "Job Title" field. There are also checkboxes for "Align two or more sequences" and "Choose Search Set". The "Database" section is set to "Non-redundant protein sequences (nr)". The "Organism" field is empty, and the "Exclude" section has "Models (MXP)" and "Uncultured/environmental sample sequences" checked. The "Entrez Query" field is empty.

Kerfeld and Scott, PLoS Biology 2011 3

3

NCBI BLAST Results Page: Potential homologs retrieved from database

The screenshot shows the NCBI BLAST Results Page. At the top, there is a "Color key for alignment scores" with a scale from 0 to 180. The colors are: black (<40), blue (40-50), green (50-80), red (80-200), and white (>=200). Below the color key is a "Query" section with a red bar representing the query sequence. The main content is a table titled "Sequences producing significant alignments:".

Accession	Description	Max score	Total score	Query coverage	E value
NP_440048.1	potential FMN-protein [Synecocystis sp. PCC 6803] >sp P727	322	379	100%	1e-103
YP_001864235.1	flavin reductase domain-containing protein [Nostoc punctiforme]	199	199	100%	2e-49
YP_321888.1	flavin reductase-like, FMN-binding [Anabaena variabilis ATCC	198	198	98%	3e-49
NP_488484.1	flavoprotein [Nostoc sp. PCC 7120] >sp Q8YNW7.1 DFAA_ANA	197	197	98%	6e-49
CAO89562.1	dfad [Microcystis aeruginosa PCC 7806]	194	194	100%	3e-48
ZP_01630850.1	flavoprotein [Modularia spumigena CCY9414] >gb EAW44518	193	193	100%	6e-48

Below the table, there is a detailed view of a sequence alignment. The sequence is "flavin reductase domain protein FMN-binding [Cyanotheca sp. PCC 7425]". The alignment shows a score of 176 bits (446), an expectation of 1e-42, and a method of Compositional matrix adjust. The identities are 95/196 (48%), positives are 125/196 (63%), and gaps are 16/196 (8%). The alignment is shown as a red bar with a black bar below it.

Kerfeld and Scott, PLoS Biology 2011 4

4

Overview of BLAST

1. Segment the query sequence into short “words”
2. Use the query sequence segments to scan the database for matching sequences
3. Extend the matched segments in either direction to find local alignments.
4. Create a list of hits & alignments, with best matches first

5

BLAST Phase 1: Segment the query sequence and identify words that could form potential alignments

Query Sequence:

```
>gi|16329320 (residues 412 to 594)
SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVVTQTG
RHQGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFVLNLLQEGRS
VRRHFDHQPLPKDGNPPFSRLEHYSTQNGCLILAEALAYLECLVQSWNSI
GDHVLVYATVQAGQVLQPNGITAIRHRKSGGQY
```

Fragmentation into words:

```
SWVSQASFTPPGIM → SWV WVS VSQ SQA QAS ASF SFT ...
```

Selection of words scoring above threshold (for word SWV):

Substitution Matrix*						
	R	G	I	K	F	S
R	5	0	-1	-1	-2	1
G	6	-4	-2	-3	0	-2
I		4	-3	0	-2	-1
K			5	-3	0	-1
F				6	-2	-2
S					4	1
T						5
W						
V						

SWV (4+11+4 = 19)	} Synonyms above threshold 11... (others not shown)	
SWI (4+11+3 = 18)		
TWV (1+11+4 = 16)		
GWV (0+11+4 = 15)		
KWV (0+11+4 = 15)		
SWS (4+11-2 = 13)		
SFV (4+1+4 = 9)		} Synonyms below threshold 11... (others not shown)
SRV (4-3+4 = 5)		

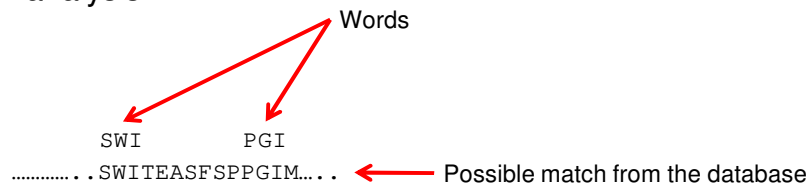
*A portion of the BLOSUM 62 matrix

- Segment the query sequence into pieces (“words”)
 - Default word length: 3 amino acids or 11 nucleic acids
- Create a list of synonyms and their scores for comparing query words to target words
 - Uses scoring matrix to calculate scores for synonyms that might be found in the database
- Save the scores (and synonyms) exceeding a given threshold T

6

BLAST Phase 2: Using the query sequence word list, scan the database for synonyms (hits)

- Scan the database for matches to the word list with acceptable T values
- Require two matches (“hits”) within the target sequence
- Set aside sequences with matches above T for further analysis



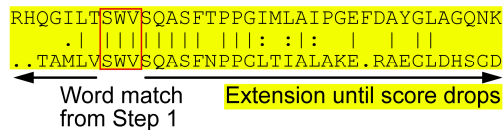
Kerfeld and Scott, PLoS Biology 2011

7

7

BLAST Phase 3: Extending the hits

- Search 5' and 3' of the word hit on both the query and target sequence
- Add up the score for sequence identity or similarity until value exceeds S
- Alignment is dropped from subsequent analyses if value never exceeds S



Kerfeld and Scott, PLoS Biology 2011

8

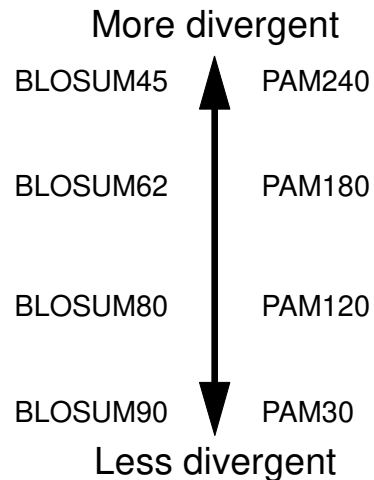
8

So, to summarize:

- BLAST segments query sequence into “words” and scores potential word matches
- Scans this list for alignments that meet a threshold score T
 - uses a scoring matrix to calculate this (e.g., **BLOSUM62**)
- Uses this list of ‘synonyms’ to scan the database
- Extends the alignments to see if they meet a cutoff score S
 - uses a scoring matrix to calculate this
- Reports the alignments that exceed S

PAM and BLOSUM Matrices

- Scoring matrices are calibrated to capture different degrees of sequence similarity
- In practice, this means choosing a matrix appropriate to the suspected degree of sequence identity between the query and its hits
- PAM: empirically derived for close relatives
- BLOSUM: empirically derived for distant relatives



Raw Scores (S values) from an Alignment

$$S = (\sum M_{ij}) - cO - dG,$$

where

M = score from a similarity matrix

for a particular pair of amino acids (ij)

c = number of gaps

O = penalty for the existence of a gap

d = total length of gaps

G = per-residue penalty for extending
the gap

Limitations of Raw Scores

- S values depend on the substitution matrix, gap penalties
- Impossible to compare S values from hits retrieved from BLAST searches when different matrices and gap penalties are used

Going from Raw Scores to Bit Scores

$$S' = [\lambda S - \ln(K)] / \ln(2)$$

where

S' = bit score

λ and K = normalizing parameters of the specific matrices and search spaces

(as in 0 vs 1)

- Larger raw scores result in larger bit scores
- **Allows user to compare scores obtained by using different matrices and search spaces**

13

Limitations of Bit Scores

- How high does a bit score have to be to suggest common ancestry?
 - Hard to evaluate hits as homologs or not, based solely on bit scores

14

E-value

- Number of distinct alignments with scores greater than or equal to a given value expected to occur in a search against a database of known size, based solely on chance, not homology.
 - Large E-values suggest that the query sequence and retrieved sequence similarities are due to chance
 - Small E-values suggest that the sequence similarities are due to shared ancestry (or potentially convergent evolution)

15

Calculating E-values

$$E = (n \times m) / 2^S$$

where

- m = effective length of the query sequence
= length of query sequence – average length of alignments
(Controls for fewer alignments occurring at the ends of the query sequence)
- n = effective length of the database sequence
(total number of bases)

The value of E decreases exponentially with increasing S

16

BLAST Parameters

- Expect →
- Word size →
- Matrix →
- Gap costs →
- Filter →
- Mask →

Algorithm parameters

General Parameters

Max target sequences: 100

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: Low complexity regions

Mask: Mask for lookup table only, Mask lower case letters

BLAST Search database nr using Blastp (protein-protein BLAST)

Show results in a new window

Kerfeld and Scott, PLoS Biology 2011

17

E value Threshold

- Alignments will be reported with E-values less than or equal to the expect values threshold
 - Setting a larger E threshold will result in more reported hits
 - Setting a smaller E threshold will result in fewer reported hits



Algorithm parameters

General Parameters

Max target sequences: 100

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: Low complexity regions

Mask: Mask for lookup table only, Mask lower case letters

BLAST Search database nr using Blastp (protein-protein BLAST)

Show results in a new window

Kerfeld and Scott, PLoS Biology 2011

18

Filter and Mask

- Filter: Low complexity
 - Replaces the following with N (nucleotides) or X (amino acids)
 - Dinucleotide repeats
 - Amino acid repeats
 - Leader sequences
 - Stretches of hydrophobic residues
- Mask: Lower case
 - Replaces lowercase letters in sequence with N or X
 - Lowercase letters typically indicate base or amino acid not known with certainty

The screenshot shows the 'Algorithm parameters' section of the BLAST interface. Under the 'Filters and Masking' sub-section, the 'Filter' checkbox is checked, and the 'Mask' checkbox is also checked. Red arrows point to the 'Filter' and 'Mask' labels. Other parameters shown include 'Max target sequences' (100), 'Short queries' (checked), 'Expect threshold' (10), 'Word size' (3), 'Matrix' (BLOSUM62), and 'Gap Costs' (Existence: 11 Extension: 1).

Kerfeld and Scott, PLoS Biology 2011

19

19

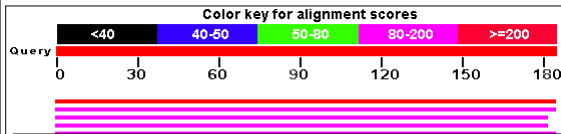
Parameter Summary is Found at the Bottom of the Output.....

Search Parameters		
Program	blastp	
Word size	3	
Expect value	10	
Hitlist size	100	
Gapcosts	11,1	
Matrix	BLOSUM62	
Filter string	F	
Genetic Code	1	
Window Size	40	
Threshold	11	
Composition-based stats	2	
Database		
Posted date	Sep 6, 2010 4:42 AM	
Number of letters	4,014,994,744	
Number of sequences	11,756,863	
Entrez query	none	
Karlin-Altschul statistics		
Lambda	0.319424	0.267
K	0.13352	0.041
H	0.397413	0.14
Results Statistics		
Length adjustment	129	
Effective length of query	54	
Effective length of database	2498359417	
Effective search space	134911408518	
Effective search space used	134911408518	

Kerfeld and Scott, PLoS Biology 2011

20

Evaluating BLAST Results



Accession	Description	Max score	Total score	Query coverage	E value
NP_440048.1	potential FMN-protein [Synecocystis sp. PCC 6803] >sp P727	379	379	100%	1e-103
YP_001304295.1	flavin reductase domain-containing protein [Nostoc punctiforme]	199	199	100%	2e-49
YP_321898.1	flavin reductase-like, FMN-binding [Anabaena variabilis ATCC	198	198	98%	3e-49
NP_418484.1	flavoprotein [Nostoc sp. PCC 7120] >sp Q81NW7.1 DFA4_ANA	197	197	98%	6e-49
CAO89562.1	flx4 [Microcystis aeruginosa PCC 7806]	194	194	100%	3e-48
ZP_01630858.1	flavoprotein [Modularia spumigena CCY9434] >cb EAW44518	193	193	100%	6e-48

```
>|ref|YP_002482587.1| G flavin reductase domain protein FMN-binding [Cyanothecae sp. PCC 7425]
|db|ACL44226.1| G flavin reductase domain protein FMN-binding [Cyanothecae sp. PCC 7425]
Length=585
GENE_ID: 7287783 Cyan7425_1859 | flavin reductase domain protein FMN-binding [Cyanothecae sp. PCC 7425]
Score = 176 bits (446), Expect = 1e-42, Method: Compositional matrix adjust.
Identities = 95/196 (48%), Positives = 125/196 (63%), Gaps = 16/196 (8%)
Query 1  SGANFARQLRTHKRQRIARQATTEQADRTQAVGRIGSIGVVTQTGRH----- 52
          +G++FA+ L+ K+OR RQ+ E Q+DR+QAVGRIGS+ V+T + H
Sbjct 393  AGSDFAQVLKAKKQKRSRQSIQVSDRTEQAVGRIGSLCVLTAKQQQTHPHEVEEP 452
Query 53  -----QGILTSVSVQASFTPPGIMLAIPECFDAYGLAGQNKAFVNLNLLQGRSVRRHFDH 107
          +L SVVSVQASF PPG+ +A+ E A GL AFVNL+L+EG ++RRHF
Sbjct 453  QLEVPTAMLVSVSVQASFPVPTLALAKE-RAEGLDHSQDAFVNLNLLKGMNLRHPSK 511
Query 108 QPLPKDCDNPFSRLEHYSTQNGCLLAEALAYLECLVQSWSNIGDHVLYATVQAGQVQLQ 167
          P G++ F+ L +NGC +L+ LAYLEC VQS GDH L+YATV G+VQLQ
Sbjct 512  SFAP--GEDRFAGLNQWAENGCPVLQDCLAYLECTVQSRMECGDHVLYATVYNNKGVQLQ 569
Query 168 PNGITAIRHRKSGGQY 183
          P G TA++HRKSG QY
Sbjct 570  PTGTTAVQHRKSGNQY 585
```

Kerfeld and Scott, PLoS Biology 2011

21

21

Examine the BLAST Alignment

```
>|ref|YP_002482587.1| G flavin reductase domain protein FMN-binding [Cyanothecae sp. PCC 7425]
|db|ACL44226.1| G flavin reductase domain protein FMN-binding [Cyanothecae sp. PCC 7425]
Length=585
GENE_ID: 7287783 Cyan7425_1859 | flavin reductase domain protein FMN-binding [Cyanothecae sp. PCC 7425]
Score = 176 bits (446), Expect = 1e-42, Method: Compositional matrix adjust.
Identities = 95/196 (48%), Positives = 125/196 (63%), Gaps = 16/196 (8%)
Query 1  SGANFARQLRTHKRQRIARQATTEQADRTQAVGRIGSIGVVTQTGRH----- 52
          +G++FA+ L+ K+OR RQ+ E Q+DR+QAVGRIGS+ V+T + H
Sbjct 393  AGSDFAQVLKAKKQKRSRQSIQVSDRTEQAVGRIGSLCVLTAKQQQTHPHEVEEP 452
Query 53  -----QGILTSVSVQASFTPPGIMLAIPECFDAYGLAGQNKAFVNLNLLQGRSVRRHFDH 107
          +L SVVSVQASF PPG+ +A+ E A GL AFVNL+L+EG ++RRHF
Sbjct 453  QLEVPTAMLVSVSVQASFPVPTLALAKE-RAEGLDHSQDAFVNLNLLKGMNLRHPSK 511
Query 108 QPLPKDCDNPFSRLEHYSTQNGCLLAEALAYLECLVQSWSNIGDHVLYATVQAGQVQLQ 167
          P G++ F+ L +NGC +L+ LAYLEC VQS GDH L+YATV G+VQLQ
Sbjct 512  SFAP--GEDRFAGLNQWAENGCPVLQDCLAYLECTVQSRMECGDHVLYATVYNNKGVQLQ 569
Query 168 PNGITAIRHRKSGGQY 183
          P G TA++HRKSG QY
Sbjct 570  PTGTTAVQHRKSGNQY 585
```

Does it cover the whole length of both the query and subject sequences?

Kerfeld and Scott, PLoS Biology 2011

22

22

High E-value: Discovery of a Distant Homolog or Garbage?

- Take another look at the target (subject) sequence(s) that have high E-values
 - Similar length?
 - Recurring motifs?
 - Similar biological functions?
- Use target sequences as query sequences for another BLAST search
 - Does the original query sequence come up in report?

Kerfeld and Scott, PLoS Biology 2011

23

23

Or to take a more topical BLAST search, a high-profile, now retracted, *bioRxiv* preprint:

Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag

This article has been withdrawn. Click here for details

Prashant Pradhan, Ashutosh Kumar Pandey, Akhilesh Mishra, Parul Gupta, Praveen Kumar Tripathi, Manoj Balakrishnan Menon, James Gomes, Perumal Vivekanandan, Bishwajit Kundu

“We ... compared the spike glycoprotein sequences of the 2019-nCoV to SARS ...we found that the 2019- nCoV spike glycoprotein contains 4 insertions”

“To further investigate if these inserts are present in any other corona virus, we performed a multiple sequence alignment of spike glycoprotein sequences of all available coronaviruses in NCBI refseq. We found that **these 4 insertions are unique to 2019-nCoV and are not present in other coronaviruses analyzed.”**

“To our surprise, **all the 4 inserts in the 2019-nCoV mapped to short segments of amino acids in the HIV-1 gp120 and Gag among all annotated virus proteins in the NCBI database. This uncanny similarity of novel inserts in the 2019- nCoV spike protein to HIV-1 gp120 and Gag is unlikely to be fortuitous.”**

24

Let's repeat their BLAST analysis: Wuhan coronavirus spike protein x nr database

U.S. National Library of Medicine
National Center for Biotechnology Information

BLAST⁺ » blastp suite » results for RID-3P99TT0P014

Job Title: ref|YP_009724390.1
RID: 3P99TT0P014
Program: BLASTP
Database: nr
Query ID: YP_009724390.1
Description: surface glycoprotein [Wuhan seafood market pneumonia virus]
Molecule type: amino acid
Query Length: 1273

Filter Results
Organism: only top 20 will appear
Type common name, binomial, taxid or group name
+ Add organism
Percent Identity: [] to []
E value: [] to []
Query Coverage: [] to []
Filter Reset

Sequences producing significant alignments

select all	Description	Max Score	Total Score	Query Cover	E value	Per Ident	Accession
<input checked="" type="checkbox"/>	spike glycoprotein [Wuhan seafood market pneumonia virus]	2640	2640	100%	0.0	100.00%	QH863250.1
<input checked="" type="checkbox"/>	surface glycoprotein [Wuhan seafood market pneumonia virus]	2637	2637	100%	0.0	100.00%	YP_009724390.1
<input checked="" type="checkbox"/>	spike glycoprotein [Bat coronavirus]	2634	2634	100%	0.0	99.92%	QH864449.1
<input checked="" type="checkbox"/>	spike protein [Bat SARS-like coronavirus]	2565	2565	100%	0.0	97.41%	QH863309.1
<input checked="" type="checkbox"/>	spike protein [Bat SARS-like coronavirus]	2105	2105	99%	0.0	80.32%	AVP78042.1
<input checked="" type="checkbox"/>	spike protein [Bat SARS-like coronavirus]	2092	2092	99%	0.0	81.00%	AVP78031.1
<input checked="" type="checkbox"/>	spike protein [Bat SARS-like coronavirus]	2066	2066	100%	0.0	77.07%	AT098205.1
<input checked="" type="checkbox"/>	spike protein [SARS-like coronavirus]	2066	2066	100%	0.0	76.92%	AT098157.1
<input checked="" type="checkbox"/>	spike protein [SARS-like coronavirus]	2065	2065	100%	0.0	77.07%	AB024652.1
<input checked="" type="checkbox"/>	spike protein [SARS-like coronavirus]	2054	2054	100%	0.0	77.38%	AC165703.1
<input checked="" type="checkbox"/>	spike protein [Bat SARS-like coronavirus]	2050	2050	99%	0.0	77.31%	AG248806.1
<input checked="" type="checkbox"/>	spike protein [Bat SARS-like coronavirus]	2049	2049	99%	0.0	77.23%	AT098132.1
<input checked="" type="checkbox"/>	spike glycoprotein [SARS coronavirus]	2048	2048	100%	0.0	76.27%	AS500023.1

A better top hit
Their top hit

Score %ID
2105 80%
vs
2048 76%

25

spike protein [Bat SARS-like coronavirus]
Sequence ID: AVP78042.1 Length: 1245 Number of Matches: 1

Range 1: 12 to 1245 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
2105 bits(5454)	0.0	Compositional matrix adjust.	1016/1265(80%)	1123/1265(88%)	33/12

Query 11 VSSQCWLLTRTLPAYNHSFRGVYDQKFRVSSLHSDQLFLPFSEHVTWFAIHV 70
V+SQC+LT RT L P YTNIS RGVYDQ+RS L +Q FLPF+SIVH+H+++
Sbjct 12 VHSQC-DLGRTPLNPHYTNSSQGVYDPTIYRSDTLVLSDQGLPFYSNVSWYYSL-T 69

Query 71 GTNGTKR DNPVLPFNDGVYFASTEKSNIRGNIIGFTLLDSKTSQSLIVNNATNVIVK 130
TKR DNP+L F Dg+YFA+TE SNI+RGNIGFTLLD+ +QSLIVNNATNV+IKV
Sbjct 70 NNAATKR DNPILDFKDGIFYAATEHSNIRVGNIGFTLLDNTSQSLIVNNATNVIVK 129

Query 131 CEFQFCNDPFLGVYHKNNKSESEFRVYSSANNCTFEYVSQPLMDLEGKQNFKNLR 190
C F FC DP+L YH NNK+ SE EF VVS NCTFEYVS+ F++++ G G F LR
Sbjct 130 CNFDFCYDPYLSG YH-NNKT SIREFAVYSFYANCTFEYVSKFNLNISGNGLFNTRLR 188

Query 191 EFVFNKIDGFKIVSKHTPILNRDLPQGSALPLVDLPIGINITRFQTLALHRSVLT 250
EFVFNKIDGFKIVSK TP+NL R LP G S L+PLV+LP+ INT+P+TLL +HR
Sbjct 189 EFVFNKIDGFKIVSKHTPILNRGLPTGLVSDLPVLELPSYINIKTFTLLTHR --- 244

Query 251 PGD--SSSGWVGAAYVYGVLPRTFLFKYNEIGITDADVCDLPLSEKCTLKSFV 308
GD -S+GWT +AA+VYGL+PRT+LKYNEIGITDADVCDLPLSEKCTLKSLV
Sbjct 245 -GDPMSNNGWVSAAYVYGLKPRFMLKYNEIGITDADVCDLPLSEKCTLKSLV 303

Query 309 EKIYQTSNFRVQPTSIIVRFPHITNLCPFGEVFNATRFASVYAIRKTSNCAVDYSL 368
+KGIYQTSNFRVQPTSIIVRFPHITN+CPE +VFNATRF SVYAM R +IS+C+ADY+V
Sbjct 304 QKGIYQTSNFRVQPTSIIVRFPHITNVCPEFHVFNATRFVSWAMERTKLSDCIADYTVF 363

Query 369 YIASASFTKCYGVSPKLNLDLCTFMVYDVSFVIRGDEVRQAPQGTGKIADYNYKLPDD 428
YIS SFSTFKCYGVSP+KL DLCT+VYAD+F+IR EVRQ+APQGT IADYNYKLPDD
Sbjct 364 YNISFTFKCYGVSPKLDLCTFSYADFTLIRFSEVRQAPQGTGVIADYNYKLPDD 423

Query 429 FTGCVIAMSNNLDSKVGNNYLYLFRKSNLKPFRDISEITVQAGSTPCNGVEGFINC 488
FTGCVIAMI+ D+ +Y YR R + LKPFERD+S++ NGV
Sbjct 424 FTGCVIAMIATKQDTG----HYFYSRHSRSTKLPFERDSSDE-----NGVR--- 466

Query 489 YPFLQSYGQPTNGVYQYRNVVLSFELLHAPATVCGPKSTNLVKNKCVNFHNLGTG 548
L +Y P P + +Q RVVLSFELLHAPATVCGPK ST LVKIH+CVNFHNLG G
Sbjct 467 --TLSTYFHNVFLYQATRVVLSFELLHAPATVCGPKLSTLVKIQCVNFHNLGK 524

Query 549 TGVLTSENKFLPFGQGRDIADTDAVRDQTLLEILDITPCSFSGVSV 548
TGVLT+S+H+F PFGQGD+D D+VRDQTLLEILDITPCSFSGVSV
Sbjct 525 TGVLTSSKRFQSFQFGKQDASDFIDSVRDPQTLLEILDITPCSFSGVSV 548

Query 609 AVLYQDVNCTVPAIHADQLTPMWRVYSGSHVFTQAGCLIGAEHVM 548
AVLYQDVNCT+VP IHADQLTP WR+Y+ G+VFTQ+AGCLIGAEHVM
Sbjct 585 AVLYQDVNCTVPTTIHADQLTPAWRIYAGTISVFTQAGCLIGAEHVM 548

Query 669 GICASYQTSNFRVQPTSIIVRFPHITNLCPFGEVFNATRFASVYAIRKTSNCAVDYSL 368
GICASY + S ++IAYTMSLGAENS+AY+HNSIATP
Sbjct 645 GICASYFAS---IILSTGQKATVAYTMSLGAENS IAYVNSIATP 368

Query 729 VSMKTSVDCVTHYICGDSFTECSNLLLOYGFSCTOLNRLATGIAVEQD 548

The actual top BLAST hit (bat coronavirus) has the insertions

It dates to 2018:

Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats

They tested 334 bats for coronaviruses from Zhoushan city, China

"we found that the virus can cause disease in suckling rats...to study the possibility of cross-species transmission"

26