

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
GEEQTEFDVVLGTGAGDNKVAATKAVRGATGLGLKEAKSAVESAPFTLKEG
VSKEEAETLANELKEAGIEVEVK
```

Nucleotide sequence:

```
>gi|118139508:333094-333465 Thiomicrospira crunogena XCL-2
ATGGCAATTACAAAAGACGATATTTTAGAAGCAGTTGCTAACATGTCAGTAATGGAAG
TTGTTGAACTTGTGAAGCAATGGAAGAGAAGTTTGGTGTCTTCTGCAGCAGCAGTTGC
GGTTGCAGGTCTGCAGGTGATGCTGGCGCTGCTGGTGAAGAACAAACAGAGTTTGAC
GTTGTCTTGACTGGTGCTGGTGACAACAAAGTTGCAGCAATCAAAGCCGTTTCGTGGCG
CAACTGGTCTTGGGCTTAAAGAAGCGAAAAGTGCAGTTGAAAGTGCACCATTTACGCT
TAAAGAGGGTGTCTTAAAGAAGAAGCAGAAACTCTTGCAAATGAGCTTAAAGAAGCA
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 search bar with a yellow box prompting the user to "Paste FASTA format sequence here". Below the search bar are fields for "Enter accession number, gi, or FASTA sequence", "Query subrange" (From and To), and "Or, upload file". There are also fields for "Job Title" and "Align two or more sequences". The "Choose Search Set" section includes options for "Database" (Non-redundant protein sequences), "Organism" (with a dropdown and "Exclude" checkbox), and "Exclude" (Models (MXP) and Uncultured/environmental sample sequences). There is also an "Entrez Query" field.

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 scale is color-coded: red for scores <40, blue for 40-50, green for 50-80, yellow for 80-200, and red for scores >=200. Below the color key is a table titled "Sequences producing significant alignments:".

Accession	Description	Max score	Total score	Query coverage	E value
NP_440048.1	potential FMN-protein [Synechocystis sp. PCC 6803] >sp P727	379	379	100%	1e-103
YP_001384295.1	flavin reductase domain-containing protein [Nostoc punctiforme]	199	199	100%	2e-49
YP_321889.1	flavin reductase-like, FMN-binding [Anabaena variabilis ATCC	198	198	98%	3e-49
NP_408464.1	flavoprotein [Nostoc sp. PCC 7120] >sp O51N77.1 DFA4_ANF	197	197	98%	6e-49
CAO89562.1	flaK [Microcystis aeruginosa PCC 7806]	194	194	100%	3e-48
ZP_01630850.1	flavoprotein [Nodularia spumigena CCY9414] >gb EAW44518	193	193	100%	6e-48

Below the table, there is a section for "Sequences producing significant alignments:" with a detailed view of the top hit. The top hit is "flavin reductase domain protein FMN-binding [Cyanobacteria sp. PCC 7425]". The alignment shows the query sequence (SGANFARQLKTHORRIARQATTETQADRTQAVRIGSIGVYTTTGRH) and the subject sequence (AOSDFAQVLRKAGKGRSPKSTLIVQSDRTEQAVRIGSIGVLTAFQGGTTPHPEVEEP). The alignment score is 176 bits (446), Expect = 1e-42, Method: Compositional matrix adjust. Identical residues = 95/196 (48%), Positives = 125/196 (63%), Gaps = 16/196 (8%).

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:

```
>gil16329320 (residues 412 to 594)
SGANFARQLRTHKQRRIARQATTETQADRTQQAVGRIGSIGVVTTQTG
RHQGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFVLMNLLQEGRS
VRRHFDHQPLPKDGDNPFSLRLEHYSTQNGCLILAEALAYLECLVQSWNSI
GDHVLVYATVQAGQVLQPNGITAIRHRKSGGQY
```

Fragmentation into words:

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

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

	R	G	I	K	F	S	T	W	V
R	5	0	-1	-1	-2	1	0	-3	0
G	6	4	-2	-3	0	-2	-2	-3	
I		4	-3	0	-2	-1	-3	3	
K			5	-3	0	-1	-3	-2	
F				6	-2	-2	1	-1	
S					4	1	-3	-2	
T						5	-2	0	
W							11	-3	
V									4

*A portion of the BLOSUM 62 matrix

```
SWV (4+11+4 = 19)
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)
SRV (4-3+4 = 5)
```

Synonyms above threshold 11... (others not shown)

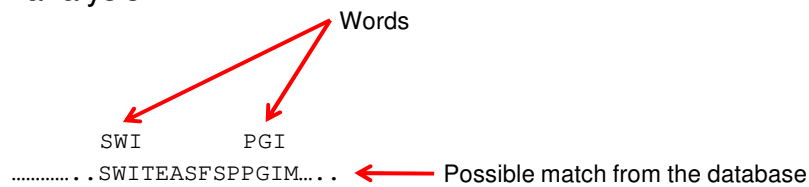
Synonyms below threshold 11... (others not shown)

- 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



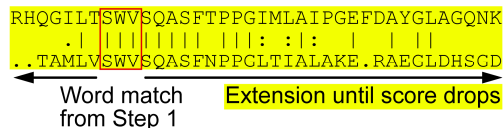
Kerfeld and Scott, PLoS Biology 2011

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

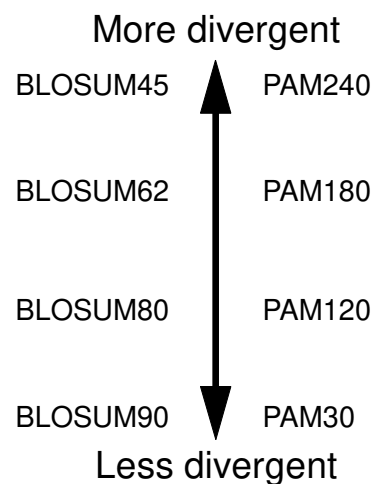
8

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

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

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)

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

BLAST Parameters

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

The screenshot shows the BLAST web interface with the following parameters:

- 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

Buttons: **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



The screenshot shows the BLAST web interface with the following parameters:

- 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

Buttons: **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 a BLAST search interface. Under the 'Filters and Masking' sub-section, there are three checkboxes: 'Filter' (checked), 'Mask for lookup table only' (unchecked), and 'Mask lower case letters' (unchecked). Red arrows point to the 'Filter' and 'Mask' labels. Below this section is a 'BLAST' button and a checkbox for 'show results in a new window'.

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

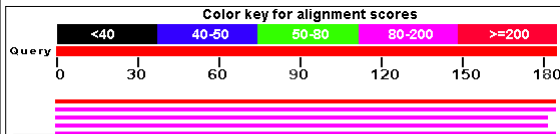
Karlin-Altschul statistics		
Lambda	0.319424	0.267
K	0.13352	0.041
H	0.397413	0.14

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



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value
WP_440348.1	potential FMN-protein [Synedocystis sp. PCC 6803] >sp P727	323	379	100%	1e-103
YP_001864235.1	flavin reductase domain-containing protein [Nostoc punctiforme]	192	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 DFA4_ANA	197	197	98%	6e-49
CAO89562.1	dfad [Microcystis aeruginosa PCC 7806]	194	194	100%	3e-48
ZP_01630850.1	flavoprotein [Modularia spumigena CCY9434] >gb 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  SGANFARQLRTHKRQRIARQATTETQADRTQAVGRIIGSIGVVTITQITGRH----- 52
          +G++FA+ L+ K+QR RQ+ E Q+DRT+QAVGRIIGS+ V+T + H
Sbjct 393  AGSDFAQVLRKAKKQRSRPSQSIILEVQSDRTEQAVGRIIGSLCVLTARQQQTHPHEVEEP 452
Query 53  -----QGILTSUVVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFVNLNLLQEGRSVRRHFDH 107
          +L SUVVSQASF PPG+ +A+ E A GL AFVLM+L+EG ++RRHF
Sbjct 453  QLEVPTAMLVSUVVSQASFNPPGLTIALAKE-RAEGLDHSGDFAFVNLNLLQEGMRLRRHFSK 511
Query 108 QPLPKDGDNPFRLEHYSTQNGCLILAEALAYLECLVQSWSNIGDHVLYATVQAGQVLQ 167
          P G++ F+ L +NGC +L + LAYLEC VQS GDH L+YATV G+VLQ
Sbjct 512  SFAP--GEDRFAGLNIQWAENGCPVLQDCLAYLECTVQSRMECGDHWLIYATVYNNGRVLQ 569
Query 168 PNGITAIRHRKSGGQY 183
          P G TA++HRKSG QY
Sbjct 570  FTGTTAVQHRKSGNQY 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  SGANFARQLRTHKRQRIARQATTETQADRTQAVGRIIGSIGVVTITQITGRH----- 52
          +G++FA+ L+ K+QR RQ+ E Q+DRT+QAVGRIIGS+ V+T + H
Sbjct 393  AGSDFAQVLRKAKKQRSRPSQSIILEVQSDRTEQAVGRIIGSLCVLTARQQQTHPHEVEEP 452
Query 53  -----QGILTSUVVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFVNLNLLQEGRSVRRHFDH 107
          +L SUVVSQASF PPG+ +A+ E A GL AFVLM+L+EG ++RRHF
Sbjct 453  QLEVPTAMLVSUVVSQASFNPPGLTIALAKE-RAEGLDHSGDFAFVNLNLLQEGMRLRRHFSK 511
Query 108 QPLPKDGDNPFRLEHYSTQNGCLILAEALAYLECLVQSWSNIGDHVLYATVQAGQVLQ 167
          P G++ F+ L +NGC +L + LAYLEC VQS GDH L+YATV G+VLQ
Sbjct 512  SFAP--GEDRFAGLNIQWAENGCPVLQDCLAYLECTVQSRMECGDHWLIYATVYNNGRVLQ 569
Query 168 PNGITAIRHRKSGGQY 183
          P G TA++HRKSG QY
Sbjct 570  FTGTTAVQHRKSGNQY 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?

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

Sequences producing significant alignments

Accession	Description	Max Score	Total Score	Query Cover	E value	Per Ident	Accession
QHR63256.1	spike glycoprotein [Wuhan seafood market pneumonia virus]	2640	2640	100%	0.0	100.00%	QHR63256.1
YP_009724390.1	surface glycoprotein [Wuhan seafood market pneumonia virus]	2637	2637	100%	0.0	100.00%	YP_009724390.1
QHR64449.1	spike glycoprotein [Bat coronavirus]	2634	2634	100%	0.0	99.92%	QHR64449.1
QHR63300.1	spike glycoprotein [Bat SARS-like coronavirus]	2565	2565	100%	0.0	97.41%	QHR63300.1
AVP78042.1	spike protein [Bat SARS-like coronavirus]	2105	2105	99%	0.0	80.32%	AVP78042.1
AVP78031.1	spike protein [Bat SARS-like coronavirus]	2092	2092	99%	0.0	81.00%	AVP78031.1
AT096205.1	spike protein [SARS-like coronavirus WIV15]	2066	2066	100%	0.0	77.07%	AT096205.1
AT096157.1	spike protein [recombinant coronavirus]	2066	2066	100%	0.0	76.92%	AT096157.1
AK024657.1	spike protein [SARS-like coronavirus RSHC014]	2065	2065	100%	0.0	77.07%	AK024657.1
AC_650703.1	spike glycoprotein [recombinant coronavirus]	2054	2054	100%	0.0	77.38%	AC_650703.1
AG248806.1	spike protein [Bat SARS-like coronavirus]	2050	2050	99%	0.0	77.31%	AG248806.1
AT098132.1	spike protein [Bat SARS-like coronavirus]	2049	2049	99%	0.0	77.23%	AT098132.1
ASS00003.1	spike glycoprotein [SARS coronavirus B229]	2048	2048	100%	0.0	76.27%	ASS00003.1

Score %ID
2105 80%
vs
2048 76%

A better top hit → (points to QHR63256.1)

Their top hit → (points to AVP78042.1)

25

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

Range 1: 12 to 1245

Query	Score	Expect	Method	Identities	Positives	Gaps
11	2105	0.0	Compositional matrix adjust.	1016/1265(80%)	1123/1265(88%)	33/12

Query 11 VSSQCWLTTRTQLPPAYTNSFRGVYDQKFRSSVLSHSTQDLFLPFSSNIVTWFHAIHV 70
 V+SQC +LT RT L P YTNIS RGVYYPD +RS L +Q FLPF+SIVW++++
 Sbjct 12 VISOQ-DLGRPLNPNYTNISQRGVYYPDTIYRSQDLVLSQGYFLPFSSNIVSHYSL- T 69

Query 71 STNIGTKR DNPVLPFNQVYFASTEKSHIIRGHIFFGTLTLDKSTQSLINNNATNVTKV 130
 TKR DNP+L F DG+YFA+TE SNI+RGNIGFGTLLD +QSLINNNATNV+IKV
 Sbjct 70 NMAATR DNPILDFKDGIFYAATEHSNIVRGNIGFGTLLDNTSQSLINNNATNVIKV 129

Query 131 CEFQFCNDPFLGVYHKINIKSMESEFRVYSSAMNCTFEVYSQFPLNDELGKQGNKILR 190
 C F C DP+L YH NIK+ EF VYS NCTFEVYS+ F+++ G G F LR
 Sbjct 130 CNIDFCYDPYLSG YH-NNK+ STREFAVYSYANCTFEVYSKFMNINISGNGLFNTLR 188

Query 191 EFVFNIDGFKIYKSHPTINLVRDLPGFSAL EPLVDLPIGINITRFQTL LALHRVLT 250
 EFVF+H+DG+FKIYSK TP+NL R LP G S L+PLV+LP+ INIT+F+TLL +HR
 Sbjct 189 EFVFNVDGDFKIYSKFPVNLNRGLPTGLSVPQLVLPVLSINTKFRTLTTHR- - - 244

Query 251 PGD-SSSGMTRGAAAYVYGLQRTFELLYKNIENGTITDAVDCALDPLSEKCTLKSFV 308
 GD S++GMTR +AAAY+VGYL+PRF+LKYNIENGTITDAVDCALDPLSEKCTLKLS +V
 Sbjct 245 GDMHSHGMTR +SAAAYVYGLKPRTFMLKYNIENGTITDAVDCALDPLSEKCTLKLSLV 303

Query 309 EKGIVQTSNFRVQPTISVIRFPNITNLCPEGVFNATRFASVAMNKRISNCAVDYSVL 368
 +KGIYQTSNFRVQPTISVIRFPNITNLCPE +VFNATRF SVAM R +IS+C+ADY+V
 Sbjct 304 QKGIYQTSNFRVQPTISVIRFPNITNLCPEGVFNATRFASVAMERTKISDCIADYTVF 363

Query 369 YNISASFSTFKCYGSPKLNLDLCTFNVAQDSVIRGDEVRQIAPQGTCKIADVNYKLPDD 428
 YNIS SFSTFKCYGSPKLNLDLCTF+VYAD+FR EVRQ+APQGT IADVNYKLPDD
 Sbjct 364 YNISFSTFKCYGSPKLNLDLCTFVYADTFLIRSEVRQAPQGTGIADVNYKLPDD 423

Query 429 FTGCVIAMNSMLDSKVGNYNYLRFKSNLKPFRDISEIYQAGSTPCNVEGFHC 488
 FTGCVIAMN+ D+ +Y R R+ LKPFERD+S+ NGV
 Sbjct 424 FTGCVIAMNTAKQDTG- - - -HYFYRSHRSTKLPFRERDLSDE- - - -NGV- - - 466

Query 489 YFPLQSYGQPTNGVGYQYRUVVLSFELLHAPATVCGPKSTNLVKNKCVNFNGLTG 548
 L +Y F P + YQ RVVLSFELL+APATVCGPK ST LVKN+CVNFNGL G
 Sbjct 467 -TLSTYDFHPHWLEYPATRVVLSFELLNAPATVCGPKSLTVLKNKCVNFNGLKLG 524

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

It dates to 2018:

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"

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

Dem Huo^{1,2}, Chengyang Zhu¹, Lili Ai¹, Ting He¹, Yi Wang¹, Fuxiang He¹, Lu Yang¹, Cheni Ding¹, Xuhui Zhu¹, Ruiheng Li¹, Jin Zhu¹, Baojun Hui^{1,2}, Youjun Peng¹, Weibing Tan¹ and Chengxin Wang¹

26