Introduction to basic bioinformatics

1) Online databases
2) Making biological sense of DNA sequences: finding and predicting function of protein coding genes
3) Using NCBI
4) What is BLAST?
5) Using BLAST for sequence analysis

www.ncbi.nlm.nih.gov

Guide to readings on bioinformatics:

1) 19 MC4 Bioinformatics

- Intro
- The UCSC Genome Browser
- Algorithms, portals and methods
- BLAST and ClustalW
- Motif finding

2) Margaret Dayhoff, pioneer of bioinformatics (2019)

3) The beginners guide to genome annotation (2012)

4) BLOSUM 62: How this homology search algorithm works.

5) 2019-nCoV: sequencing, and two perspectives articles

6) Web sites (referred to in the notes)

Bioinformatics:
storage and analysis of biological information

- Nucleotide and protein sequence
- Macromolecular structures
- Gene expression patterns
- Biochemical pathways
- Evolutionary relationships
- Sorting/visualizing large data sets

Bioinformatics databases: repositories for biological information

Primary sequence databases:

- NCBI/Genbank
- DNA Databank of Japan (DDBJ)
- European Molecular Biology Laboratory (EMBL)

Annotated protein sequence databases:

- SWISS-PROT (most accurate annotation: structures, functions, protein families, with references)
- TrEMBL (most current, but not fully annotated)

Protein structure: The Protein Databank (pdb.org)

Many other databases exist, e.g ENCODE, UCSC genomes, etc.
https://academic.oup.com/nar/issue/49/D1

Genome sequencing projects

JGI: the Joint Genome Initiative at the US Dept. of Energy (DOE)

https://genome.jgi.doe.gov/portal/

The genome annotation pipeline

1. Make sure the genome assembly is ready. N50 scaffold length should be at least the median size of a gene.
2. Find repeat sequences and mask them (mark as repeats)
3. Gene prediction:
   - Align known proteins and ESTs (Expressed Sequence Tags) to the sequence to identify exons. Then identify splice sites
   - Also us *ab initio* gene identification software
4. Gene identification:
   - Automated annotation: multiple gene finders are run, and the consensus is used
   - Alignment data can be used to improve results
   - Manual curation (evidence-based decisions)
5. Assess the quality of annotation
   - Number of protein domains
   - Agreement of annotation with RNA info (EST, seq)
6. Visualize, share, and *update* the annotation

Yandell and Ence (2012) “A beginner’s guide to eukaryotic genome annotation”

A genome annotation is continuously updated as new experiments are done

Many predicted genes have unknown functions

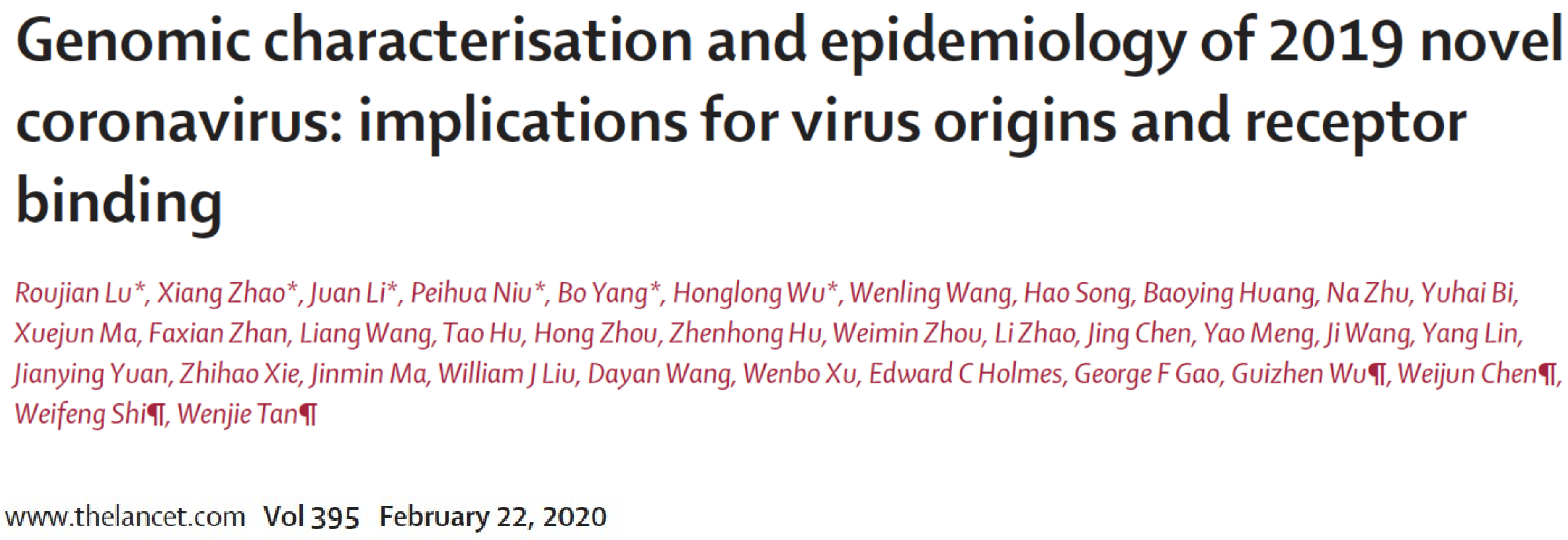
Many genome features are difficult to annotate

- Regulatory regions: promoters, DNA binding sites
- Non-coding RNAs
- Transposons
- Pseudogenes

“Like parenthood, annotation responsibilities do not end with birth. Incorrect and incomplete annotations poison every experiment that makes use of them.”

Yandell and Ence (2012) “A beginner’s guide to eukaryotic genome annotation”

Genome annotation in the news: nCoV-2, the virus causing COVID-19



Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding

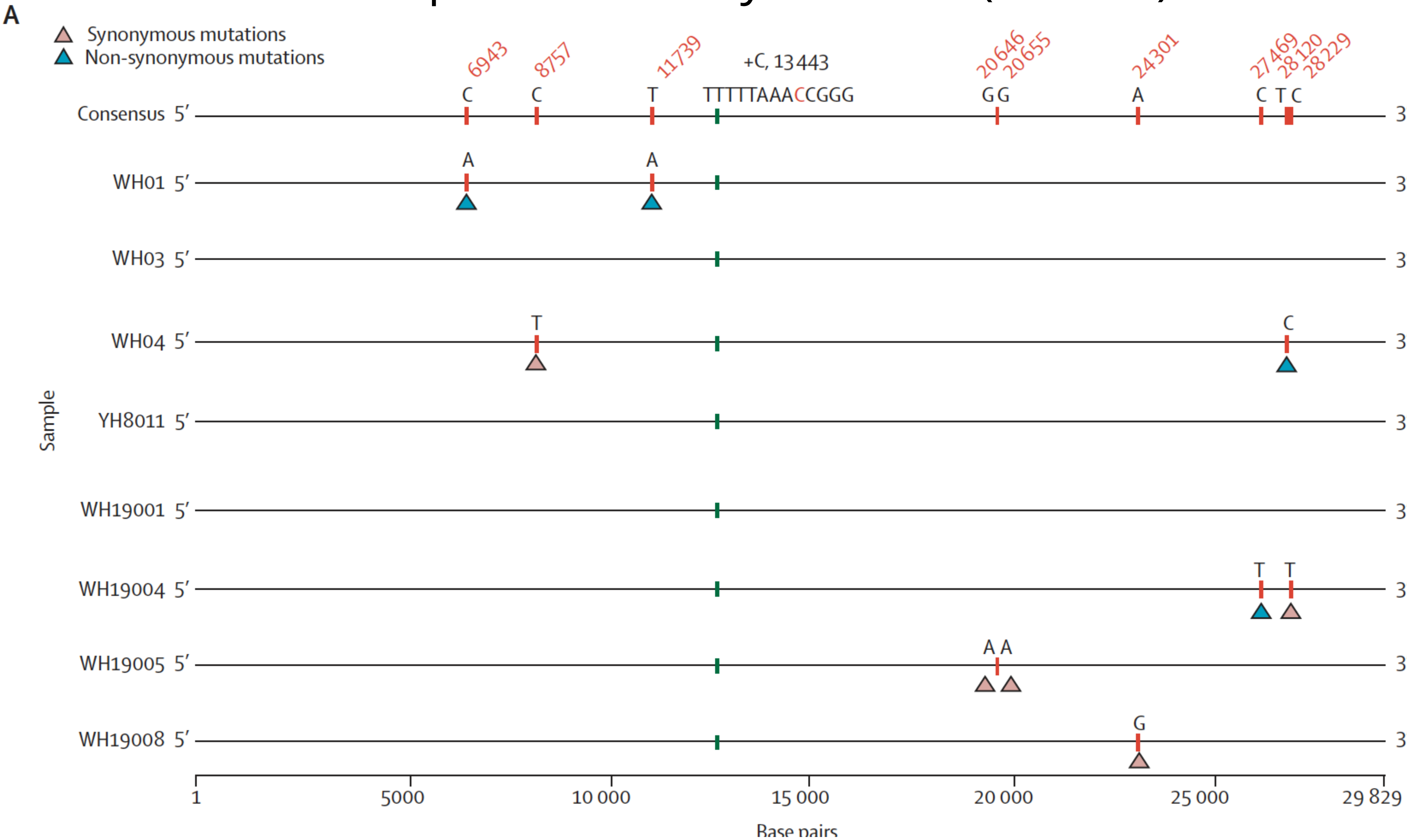
Roujian Lu, Xiang Zhao*, Juan Li*, Peihua Niu*, Bo Yang*, Honglong Wu*, Wenling Wang, Hao Song, Baoying Huang, Na Zhu, Yuhai Bi, Xuejun Ma, Faxian Zhan, Liang Wang, Tao Hu, Hong Zhou, Zhenhong Hu, Weimin Zhou, Li Zhao, Jing Chen, Yao Meng, Ji Wang, Yang Lin, Jianying Yuan, Zhihao Xie, Jinmin Ma, William J Liu, Dayan Wang, Wenbo Xu, Edward C Holmes, George F Gao, Guizhen Wu¶, Weijun Chen¶, Weifeng Shi¶, Wenjie Tan¶*

www.thelancet.com Vol 395 February 22, 2020

- Sequences of samples from 9 patients, several of whom had onset of symptoms in late December 2019

- Sequences were made freely available in January 2020

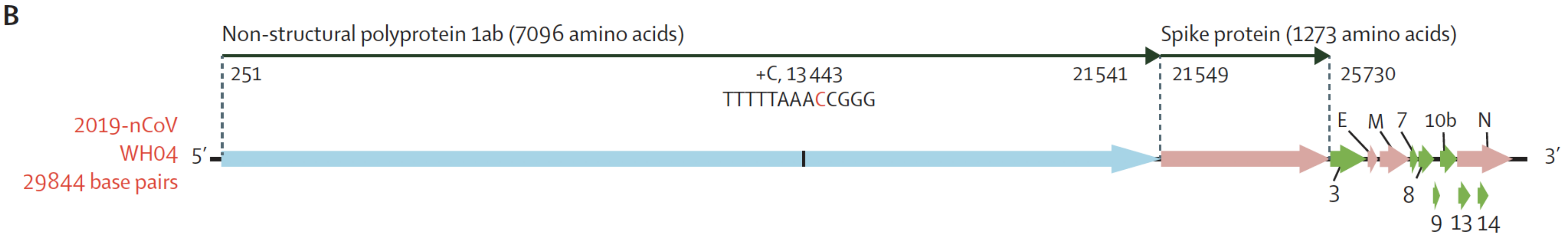
2019 nCoV-2 sequences: nearly identical (99.98%)



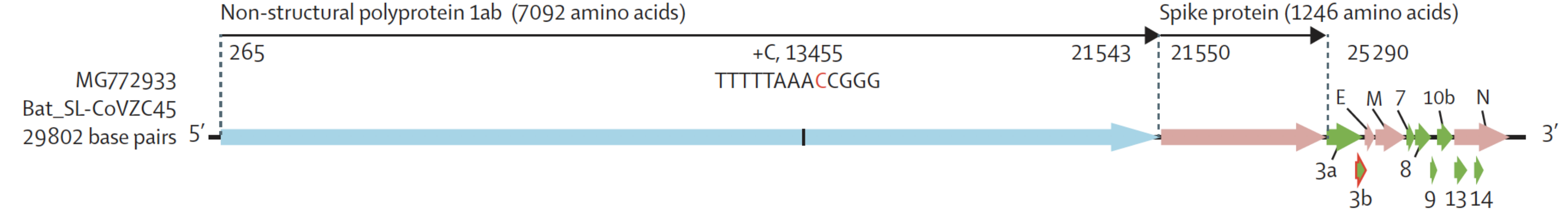
- therefore the virus has only made a recent entry into human populations

The viral genes were identified and the sequences compared with other coronaviral genomes

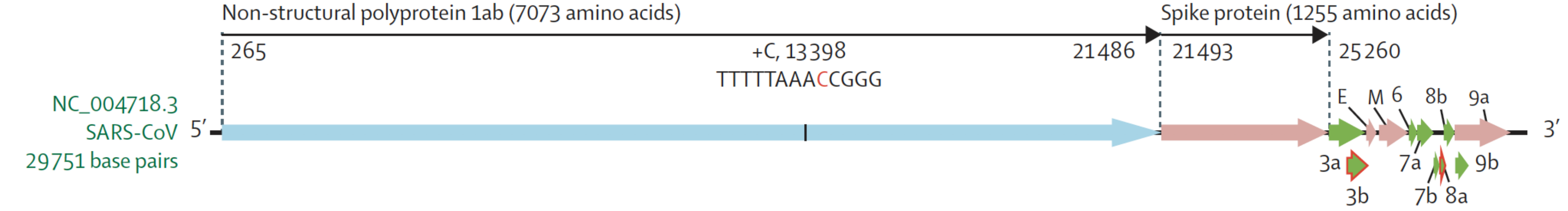
2019-nCoV genetic map



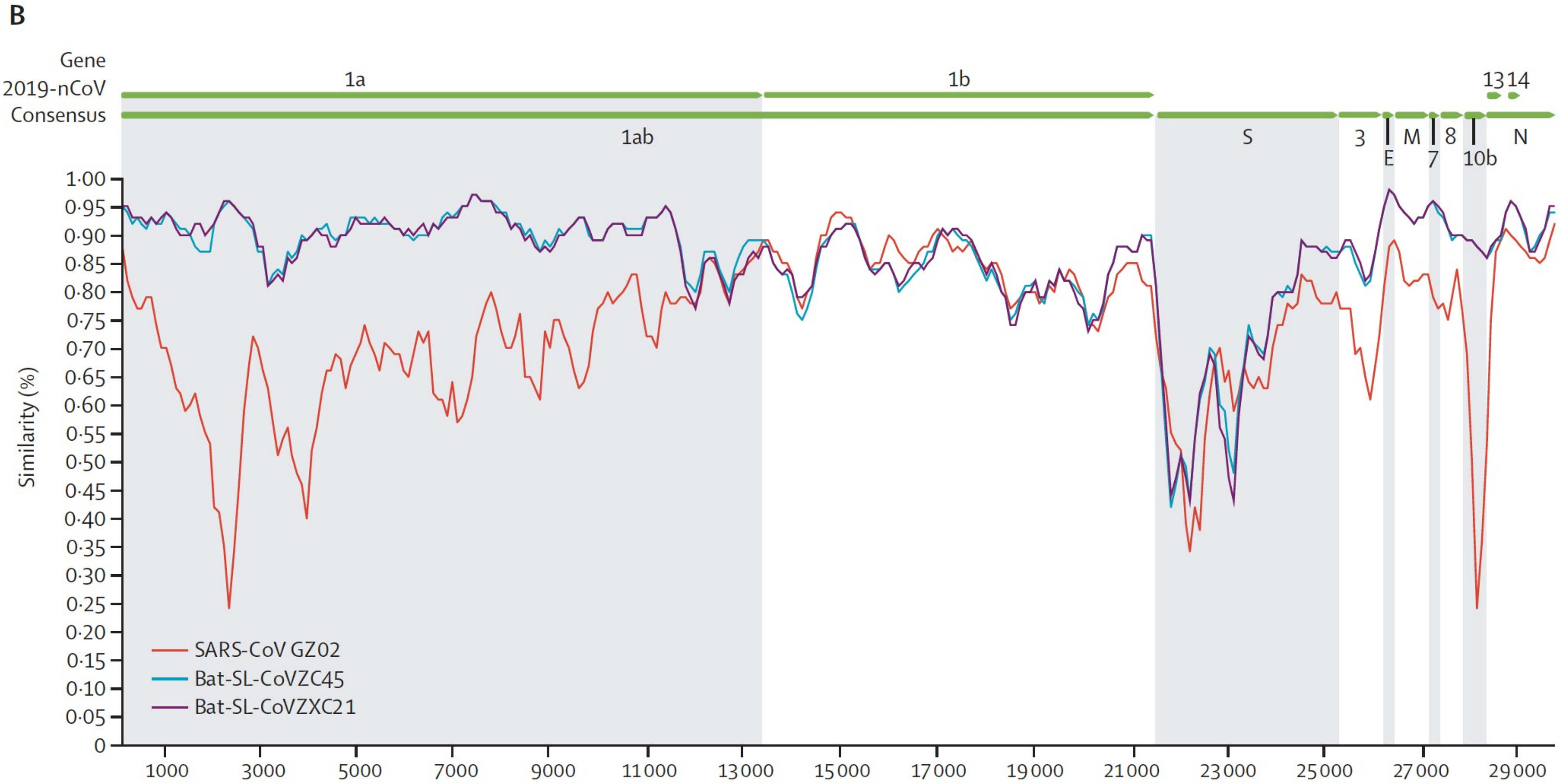
Bat SARS-like Beta coronavirus (88% sequence similarity)



SARS-CoV (79% sequence similarity)

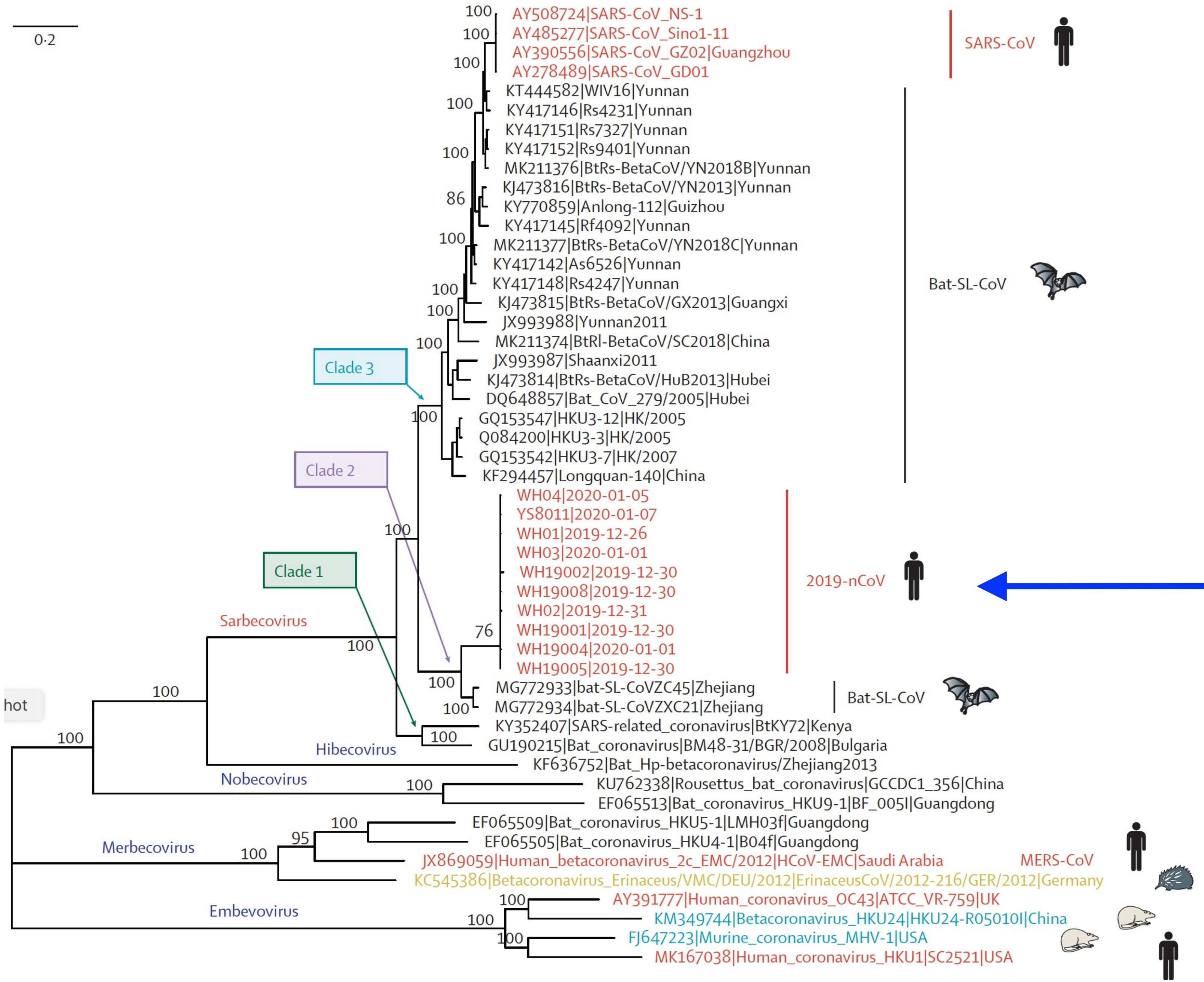


Sequence similarity of related viruses to 2019-nCov

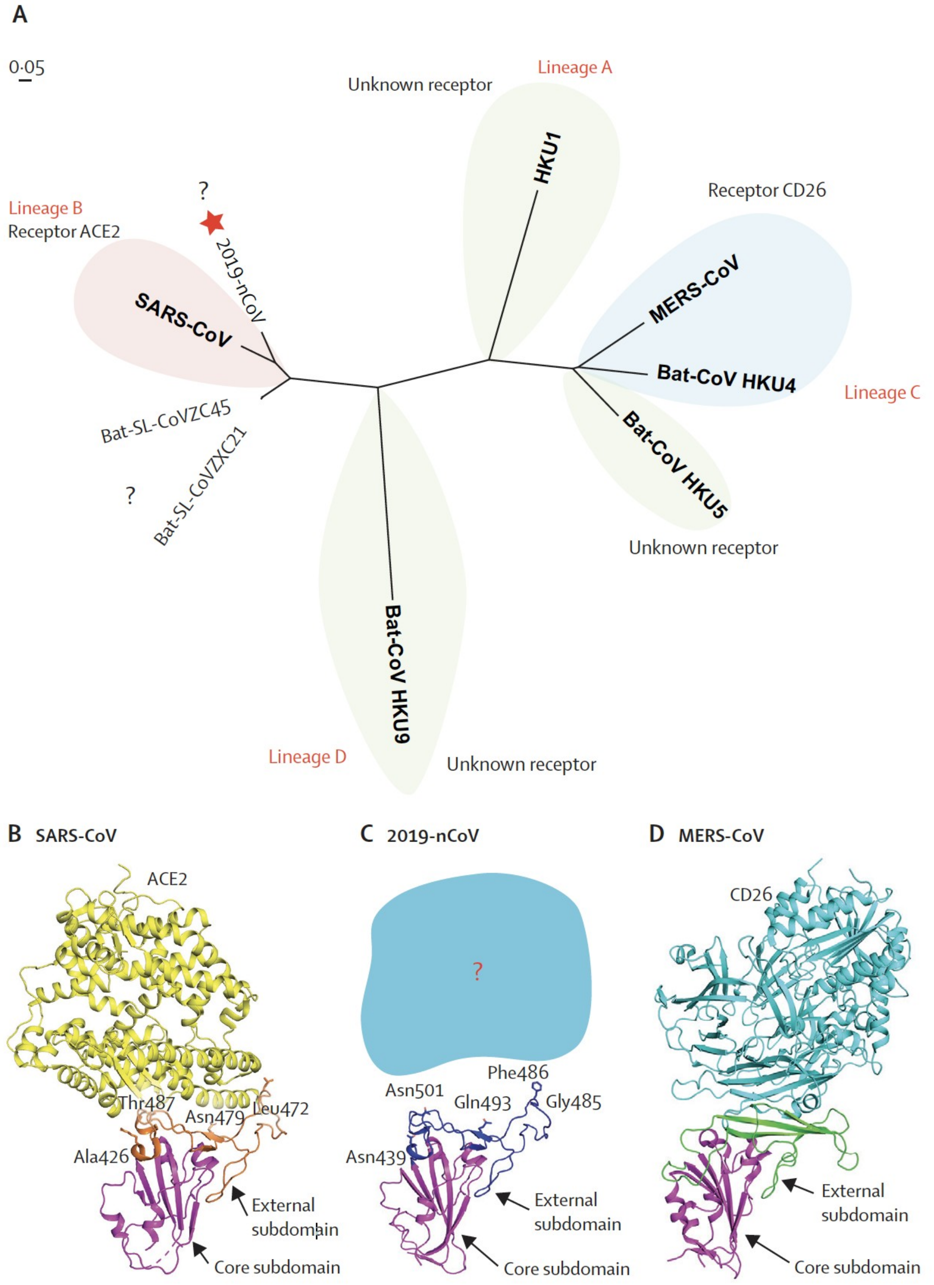


The lowest similarity with Bat CoV is in ‘spike’ protein

2019 nCoV-2 sequences are closest to bat CoV


0·2
100
AY508724|SARS-CoV_NS-1
100
AY485277|SARS-CoV_Sino1-11
AY390556|SARS-CoV_GZ02|Guangzhou
100
AY278489|SARS-CoV_GD01
SARS-CoV
KT444582|WIV16|Yunnan
100
KY417146|Rs4231|Yunnan
KY417151|Rs7327|Yunnan
100
KY417152|Rs9401|Yunnan
MK211376|BtRs-BetaCoV/YN2018B|Yunnan
KJ473816|BtRs-BetaCoV/YN2013|Yunnan
86
KY770859|Anlong-112|Guizhou
KY417145|Rf4092|Yunnan
100
MK211377|BtRs-BetaCoV/YN2018C|Yunnan
KY417142|As6526|Yunnan
KY417148|Rs4247|Yunnan
100
KJ473815|BtRs-BetaCoV/GX2013|Guangxi
100
JX993988|Yunnan2011
100
MK211374|BtRl-BetaCoV/SC2018|China
JX993987|Shaanxi2011
Clade 3
KJ473814|BtRs-BetaCoV/HuB2013|Hubei
DQ648857|Bat_CoV_279/2005|Hubei
100
GQ153547|HKU3-12|HK/2005
Q084200|HKU3-3|HK/2005
GQ153542|HKU3-7|HK/2007
KF294457|Longquan-140|China
Bat-SL-CoV
Clade 2
WH04|2020-01-05
YS8011|2020-01-07
100
WH01|2019-12-26
WH03|2020-01-01
WH19002|2019-12-30
Clade 1
WH19008|2019-12-30
2019-nCoV
WH02|2019-12-31
76
WH19001|2019-12-30
Sarbecovirus
WH19004|2020-01-01
100
WH19005|2019-12-30
100
MG772933|bat-SL-CoVZC45|Zhejiang
Bat-SL-CoV
100
hot
100
MG772934|bat-SL-CoVZXC21|Zhejiang
KY352407|SARS-related_coronavirus|BtKY72|Kenya
100
100
Hibecovirus
GU190215|Bat_coronavirus|BM48-31/BGR/2008|Bulgaria
KF636752|Bat_Hp-betacoronavirus/Zhejiang2013
Nobecovirus
100
KU762338|Rousettus_bat_coronavirus|GCCDC1_356|China
EF065513|Bat_coronavirus_HKU9-1|BF_005I|Guangdong
100
EF065509|Bat_coronavirus_HKU5-1|LMH03f|Guangdong
95
EF065505|Bat_coronavirus_HKU4-1|B04f|Guangdong
Merbecovirus
100
JX869059|Human_betacoronavirus_2c_EMC/2012|HCoV-EMC|Saudi Arabia
MERS-CoV
KC545386|Betacoronavirus_Erinaceus/VMC/DEU/2012|ErinaceusCoV/2012-216/GER/2012|Germany
100
AY391777|Human_coronavirus_OC43|ATCC_VR-759|UK
Embevovirus
100
KM349744|Betacoronavirus_HKU24|HKU24-R05010I|China
100
FJ647223|Murine_coronavirus_MHV-1|USA
MK167038|Human_coronavirus_HKU1|SC2521|USA

How does 2019-nCoV attach to human cells?



Receptor binding domain of spike protein closely related to SARS and bat CoV

SARS CoV interacts with human ACE2 receptor, likely same for 2019-nCov

Takeaways from the 2019-nCoV sequence and annotation

- Single strand RNA genome -- 29,829 bases in length

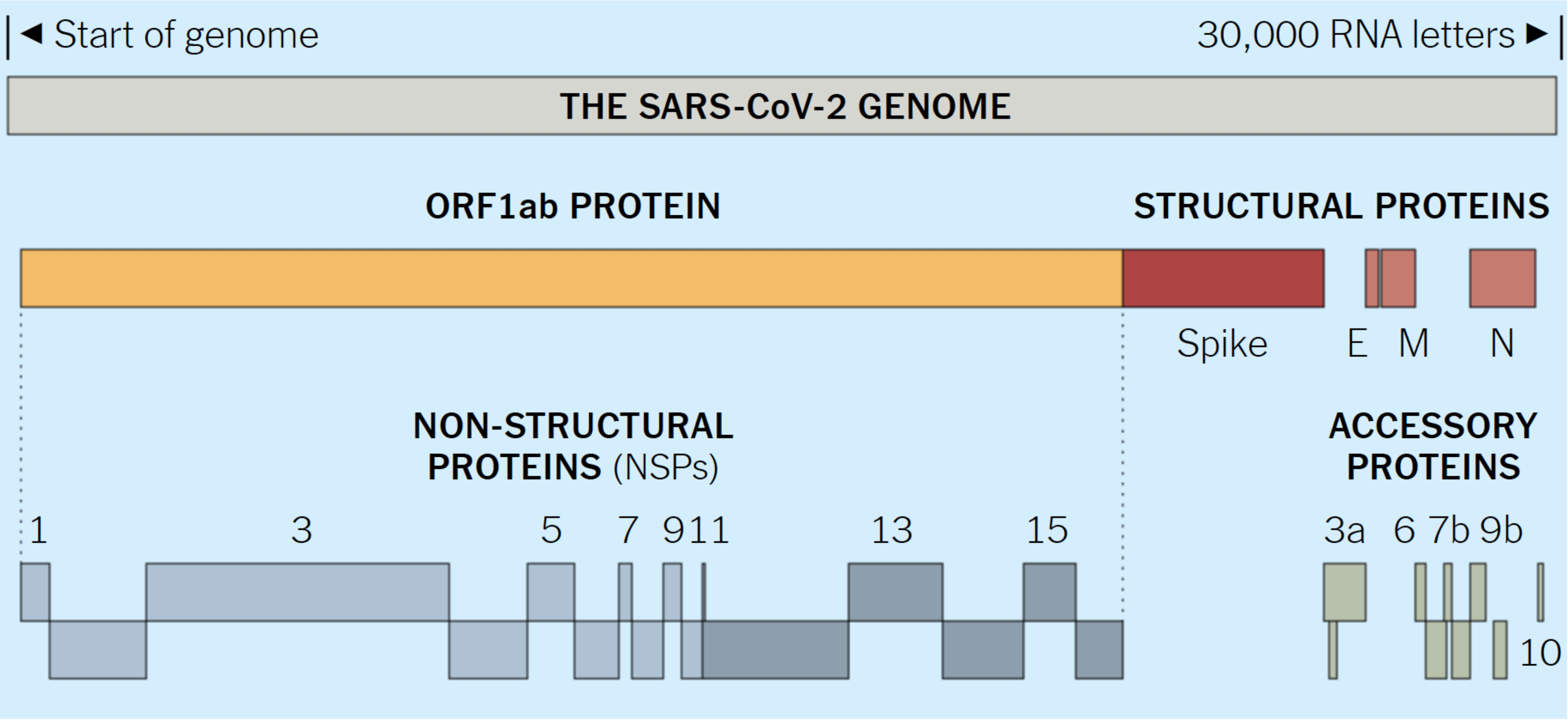
- Single introduction of the virus into humans, and then human to human spread

- Most similar to bat beta coronaviruses, but may not have come directly from bats, but instead from intermediary species

- The spike protein may interact with human ACE2 receptor protein (this is supported by recent structural work)

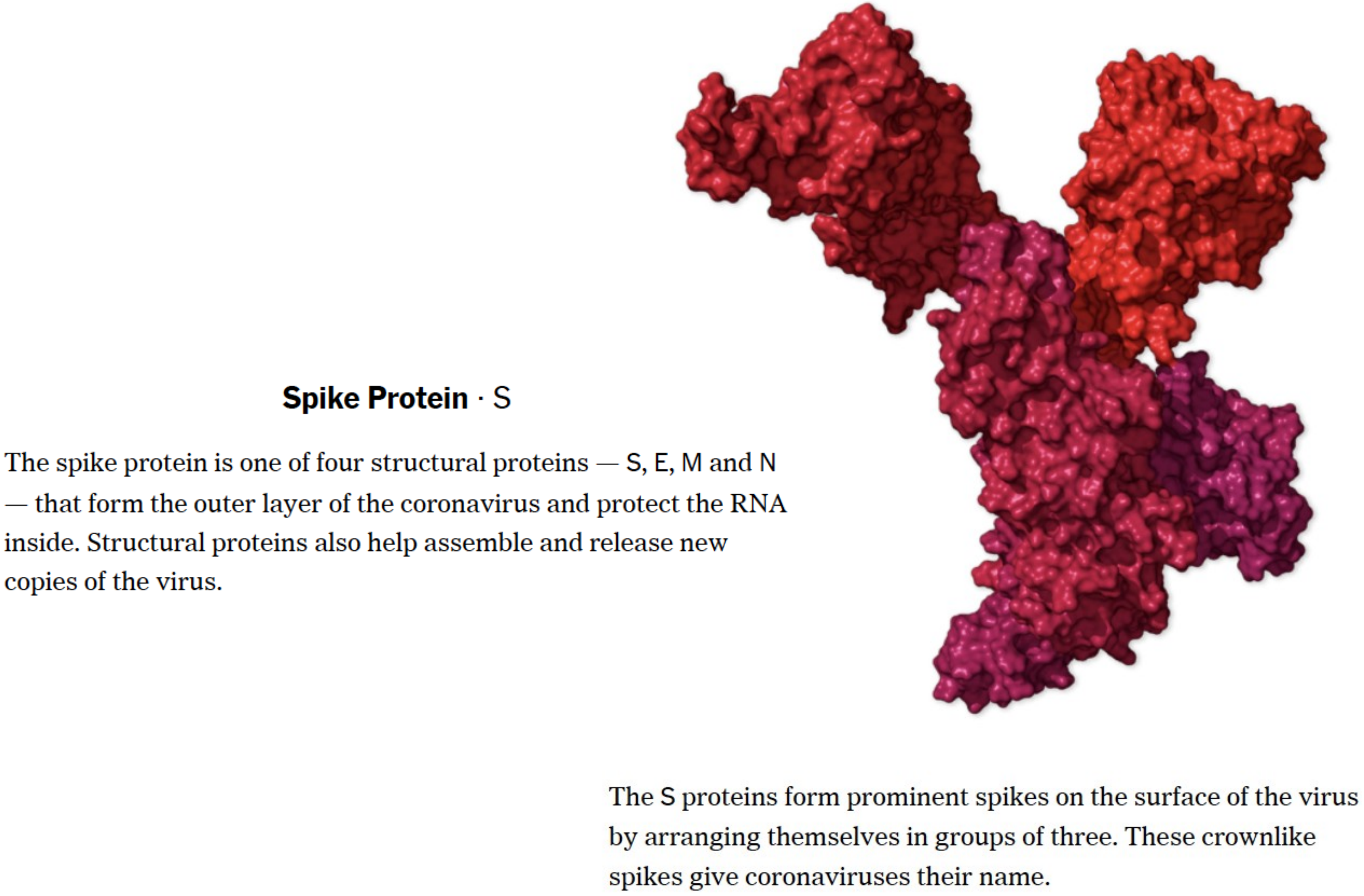
An annotation of 2019-nCoV

‘Bad news wrapped in protein’ Corum & Zimmer 4/3/20



Each protein-coding sequence is discussed in order

Protein structures can be predicted by comparison to closely related examples, as can functions

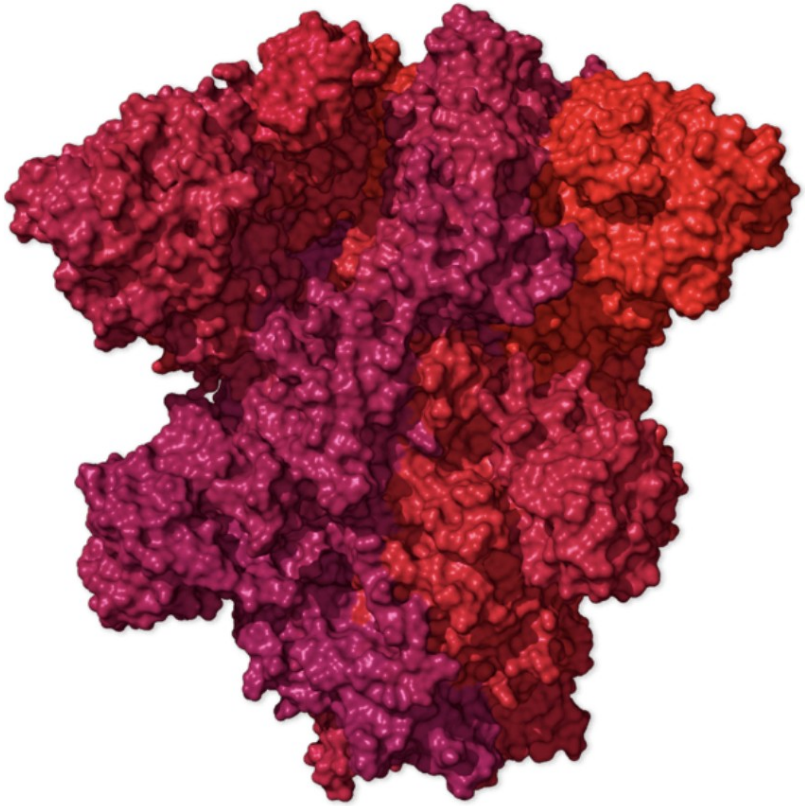


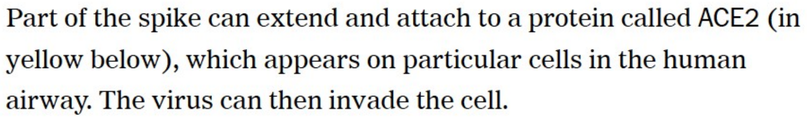
Spike Protein · S

The spike protein is one of four structural proteins — S, E, M and N — that form the outer layer of the coronavirus and protect the RNA inside. Structural proteins also help assemble and release new copies of the virus.

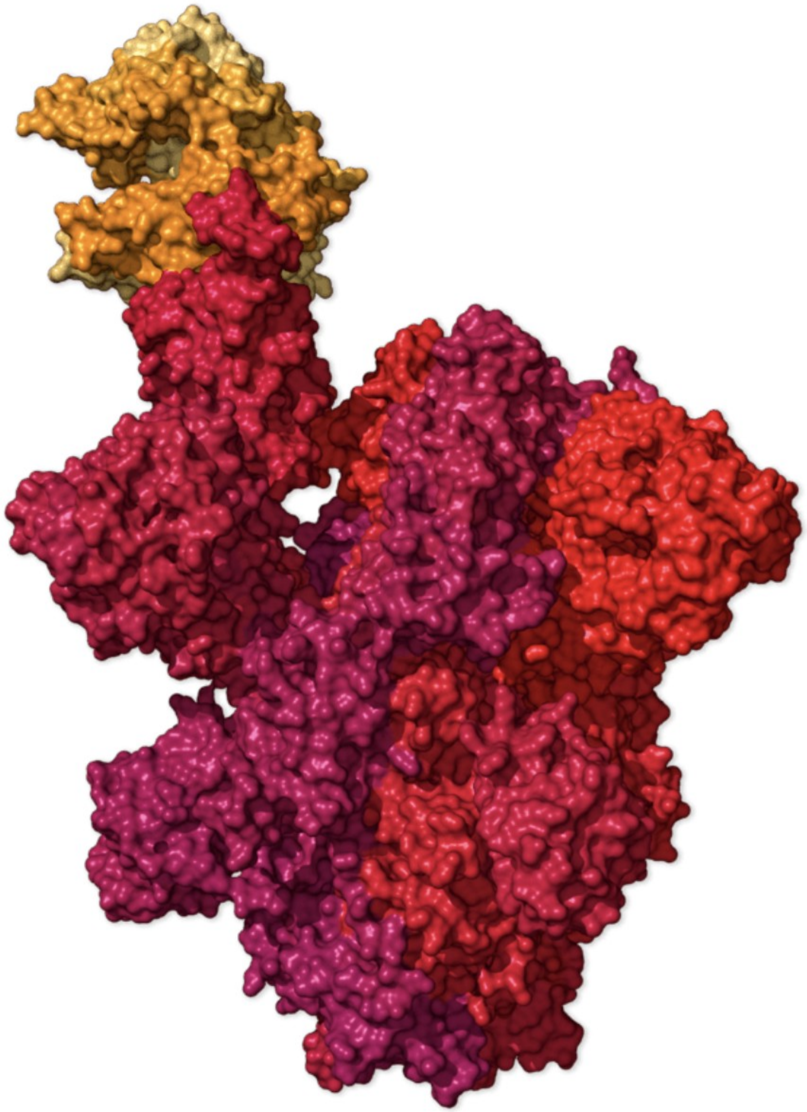
The S proteins form prominent spikes on the surface of the virus by arranging themselves in groups of three. These crownlike spikes give coronaviruses their name.

The spike protein is a target for the immune system, as well as for design of one class of antiviral therapies





Part of the spike can extend and attach to a protein called ACE2 (in yellow below), which appears on particular cells in the human airway. The virus can then invade the cell.



The spike protein changes conformation to interact with ACE2

What’s in a genome?

1) Genes

   a) Protein-coding
      - Where are the open reading frames?
      - What are the ORFs most similar to? (What is the function/structure/evolution history?)

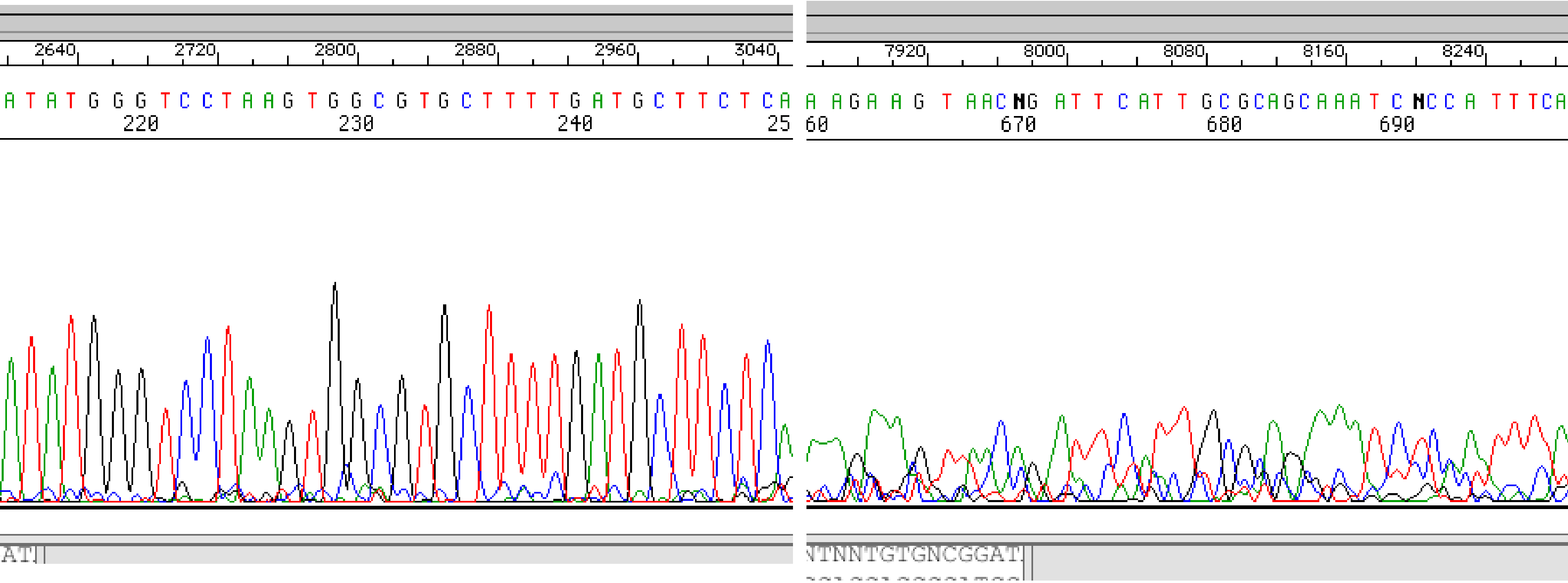
   b) RNA

2) Non-genes

   a) Regulation: promoters and factor-binding sites
   b) Transactions: replication, repair, and segregation, DNA packaging (chromatin)

Sequence output

Raw data



Computer calls

GNNTNNTGTGNCGGATACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATGCACCACCAC
CACCACCACCCCATGGGTATGAATAAGCAAAAGGTTTGTCCTGCTTGTGAATCTGCGGAACTTATTTATGATCCAGAAAG
GGGGGAAATAGTCTGTGCCAAGTGCGGTTATGTAATAGAAGAGAACATAATTG**ATATGGGTCCTAAGTGGCGTGCTTTTG**
ATGCTTCTCAAAGGGAACGCAGGTCTAGAACTGGTGCACCAGAAAGTATTCTTCTTCATGACAAGGGGCTTTCAACTGCA
ATTGGAATTGACAGATCGCTTTCCGGATTAATGAGAGAGAAGATGTACCGTTTGAGGAAGTGGCANTCCANATTANGAGT
TAGTGATGCAGCANANAGGAACCTAGCTTTTGCCCTAAGTGAGTTGGATAGAATTNCTGCTCAGTTAAAACTTCCNNGAC
ATGTAGAGGAAGAAGCTGCAANGCTGNACANAGANGCAGNGNGANAGGGACTTATTNGANGCAGATCTATTGAGAGCGTT
ATGGCGGCANGTGTTTACCCTGCTTGTAGGTTATTAAAAGNTCCCGGGACTCTGGATGAGATTGCTGATATTGCTAGAGC

What does this sequence do?

```
atgttgtatttgtctgaagaaaataaatccgtatccactccttg
ccctcctgataagattatctttgatgcagagaggggggagtaca
tttgctctgaaactggagaagttttagaagataaaattatagat
caagggccagagtggagggccttcacgccagaggagaaagaaaa
gagaagcagagttggagggcctttaaacaatactattcacgata
ggggtttatccactcttatagactggaaagataaggatgctatg
ggaagaactttagaccctaagagaagacttgaggcattgagatg
gagaaagtggcaaattaga
```

Perhaps it encodes a protein...

Does the DNA encode a protein?
Find an open reading frame (ORF) using “ ORF Finder”

https://www.ncbi.nlm.nih.gov/orffinder/

- a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database

- Identifies all open reading frames using the standard or alternative genetic codes

- Deduced amino acid sequence can be searched against other sequence databases, e.g. using the WWW BLAST server

Look up the orfs using NCBI ORF finder:

ORF identification: things to consider

- The identification of ORFs catches most, but not all protein coding genes

- Not all genes initiate with ATG, e.g. in certain microbes (e.g. archaea)

- What is the shortest possible length of a real ORF? 100 amino acids is the typical boundary, but:
  - There are many ORFs of 100-150 codons that don't encode proteins
  - There are some ORFs of less than 100 codons that do encode proteins

- In eukaryotes, identifying the full protein coding region is complicated by the presence of introns

What is the function of the ORF?

Classical methods (slow, but reliable)

- mutate gene, observe phenotype for clues to function (genetics)
- purify protein product, test activity *in vitro* (biochemistry)

Similarity of ORF to other genes

- if a gene has been previously studied, you want to know right away!
- gene sequences that have high sequence identity often have the same or similar functions

Homology of proteins

Homology: similarity of biological structure, physiology, development, and evolution, based on common ancestry

Homologous proteins: statistically similar sequence *may* indicate similar function

Alignment of sequences

The principle: two homologous sequences derived from the same ancestral sequence will have at least some identical (similar) amino acid residues that allow them to be aligned

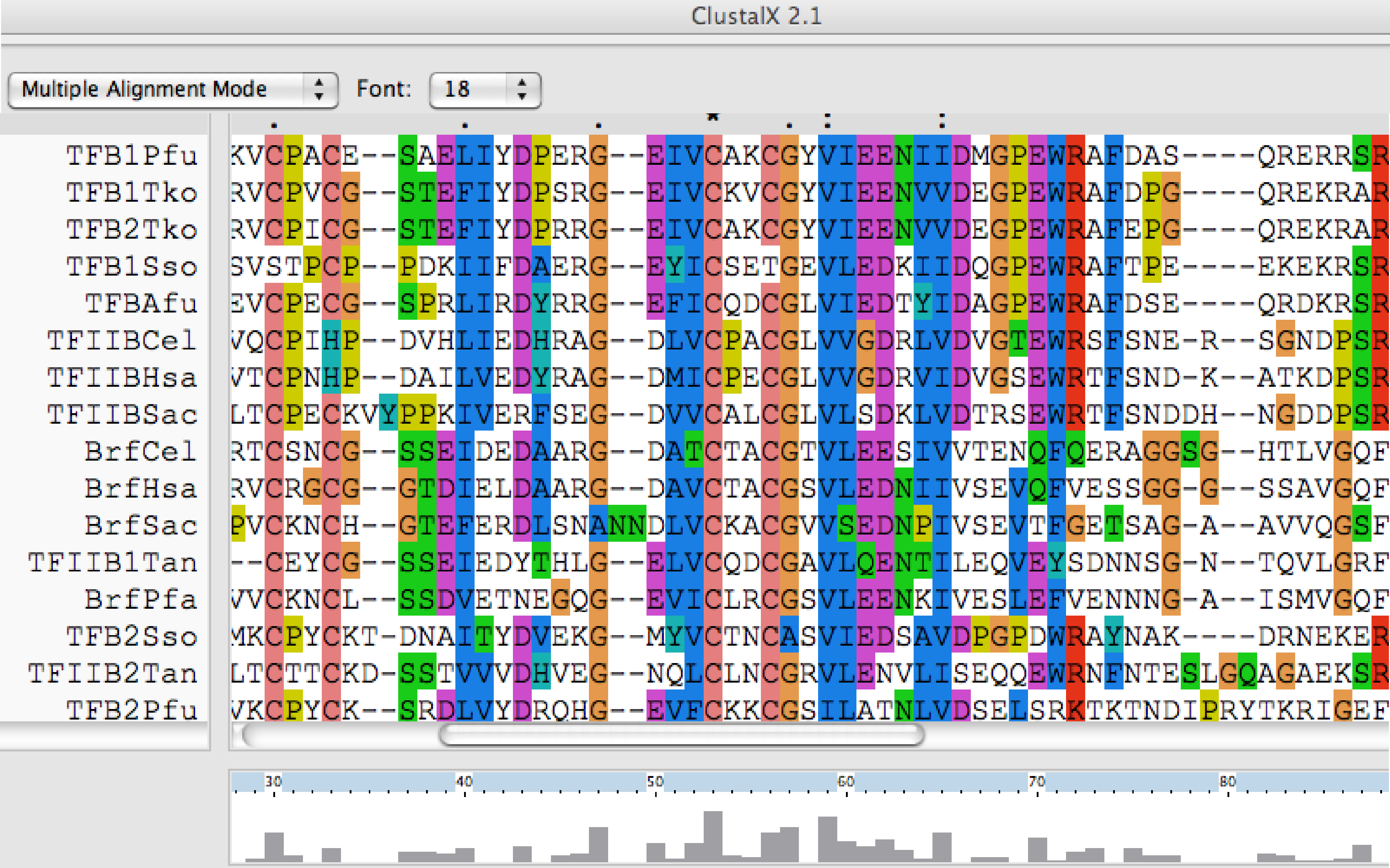
Alignment quality is judged by **three** things:

1) Percent identity: fraction of identical amino acids as a measure of structural/functional similarity

2) Similarity score: amino acids that have similar physical/chemical properties are more likely to substitute for each other in important functional regions, therefore contribute to alignment quality

3) Gaps in similar/homologous sequences are infrequent, and are cause penalty scores in alignment

Alignment of specified sequences: archaeal TFB and eukaryotic TFIIB and Brf. Similar sequence, and similar structure/function



```
               .         .      .     *   . :     :
TFB1Pfu    KVCPACE--SAELIYDPERG--EIVCAKCGYVIEENIIDMGPEWRAFDAS----QRERRSR
TFB1Tko    RVCPVCG--STEFIYDPSRG--EIVCKVCGYVIEENVVDEGPEWRAFDPG----QREKRAR
TFB2Tko    RVCPICG--STEFIYDPRRG--EIVCAKCGYVIEENVVDEGPEWRAFEPG----QREKRAR
TFB1Sso    SVSTPCP--PDKIIFDAERG--EYICSETGEVLEDKIIDQGPEWRAFTPE----EKEKRSR
TFBAfu     EVCPECG--SPRLIRDYRRG--EFICQDCGLVIEDTYIDAGPEWRAFDSE----QRDKRSR
TFIIBCel   VQCPIHP--DVHLIEDHRAG--DLVCPACGLVVGDRLVDVGTEWRSFSNE-R--SGNDPSR
TFIIBHsa   VTCPNHP--DAILVEDYRAG--DMICPECGLVVGDRVIDVGSEWRTFSND-K--ATKDPSR
TFIIBSac   LTCPECKVYPPKIVERFSEG--DVVCALCGLVLSDKLVDTRSEWRTFSNDDH--NGDDPSR
BrfCel     RTCSNCG--SSEIDEDAARG--DATCTACGTVLEESIVVTENQFQERAGGSG--HTLVGQF
BrfHsa     RVCRGCG--GTDIELDAARG--DAVCTACGSVLEDNIIVSEVQFVESSGG-G--SSAVGQF
BrfSac     PVCKNCH--GTEFERDLSNANNDLVCKACGVVSEDNPIVSEVTFGETSAG-A--AVVQGSF
TFIIB1Tan  --CEYCG--SSEIEDYTHLG--ELVCQDCGAVLQENTILEQVEYSDNNSG-N--TQVLGRF
BrfPfa     VVCKNCL--SSDVETNEGQG--EVICLRCGSVLEENKIVESLEFVENNNG-A--ISMVGQF
TFB2Sso    MKCPYCKT-DNAITYDVEKG--MYVCTNCASVIEDSAVDPGPDWRAYNAK----DRNEKER
TFIIB2Tan  LTCTTCKD-SSTVVVDHVEG--NQLCLNCGRVLENVLISEQQEWRNFNTESLGQAGAEKSR
TFB2Pfu    VKCPYCK--SRDLVYDRQHG--EVFCKKCGSILATNLVDSELSRKTKTNDIPRYTKRIGEF
```

NCBI: National Center for Biotechnology Information

NCBI home page --Go to www.ncbi.nlm.nih.gov for the following (and much, much more)

PubMed, PubMed Central: search tool for scientific literature--search by author, subject, title words, etc.

BLAST: Basic Local Alignment Search Tool

Nucleotide, Genome, Gene, Protein: databases for each

OMIM: Online Mendelian Inheritance in Man

Bookshelf: many online textbooks available

GEO: Gene Expression Omnibus, for analysis of microarray and related data

PubChem: Information of biological activities of small molecules

Guide to NCBI: see first entry at this link
https://academic.oup.com/nar/issue/49/D1

Using NCBI: Education and Tutorials pages

“ The Handbook” :
https://www.ncbi.nlm.nih.gov/books/NBK143764/

Video guides and tutorials:
https://www.ncbi.nlm.nih.gov/home/learn/

Other training and tutorials:
https://www.ncbi.nlm.nih.gov/guide/training-tutorials/

Help manual:
https://www.ncbi.nlm.nih.gov/books/NBK3831/

BLAST guide:
ftp://ftp.ncbi.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf

Note: Lots of other non-NCBI servers exist for bioinformatics tools, see index at:
https://academic.oup.com/nar/issue/49/D1

What does BLAST do?

1) Searches the chosen sequence database and identifies sequences with similarity to test sequence
2) Ranks similar sequences by degree of homology (E value)
3) Illustrates alignment between test sequence and similar sequences

Programs available for BLAST searches

Protein sequence

blastp--compares an amino acid query sequence against a protein sequence database

tblastn--compares a protein query sequence against a nucleotide sequence database translated in all reading frames

DNA sequence

blastn--compares a nucleotide query sequence against a nucleotide sequence database

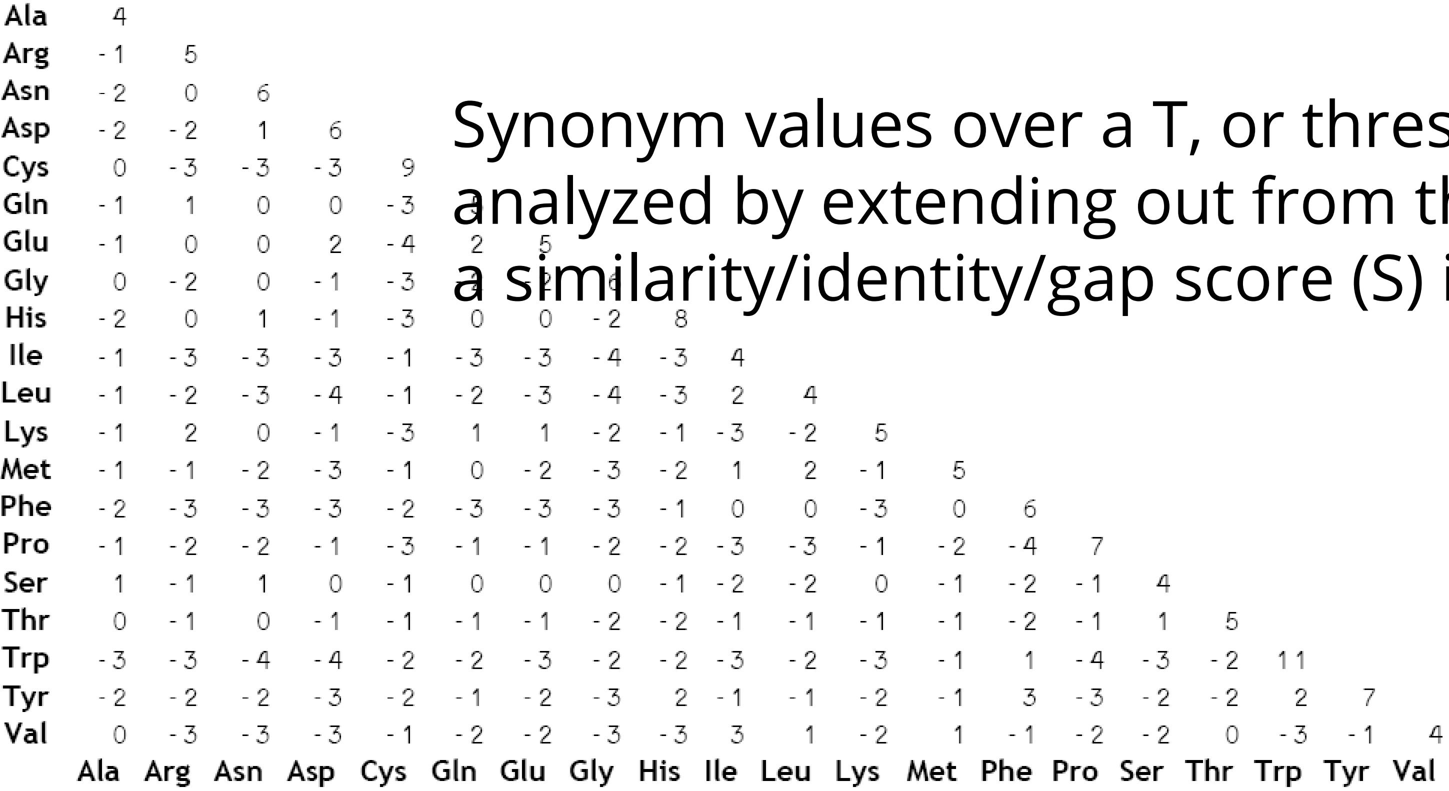
blastx--compares a nucleotide query sequence translated in all reading frames against a protein sequence database

tblastx--compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

How a protein BLAST search works

A query protein sequence is first converted into overlapping segments (words)

Synonyms for each query word are located in a sequence database, with “ scores” for each, based on a Dayhoff matrix (e.g. BLOSUM 62: BLOcks of amino acid SUbstitution Matrix)



	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Synonym values over a T, or threshold value, are analyzed by extending out from the words, and a similarity/identity/gap score (S) is generated

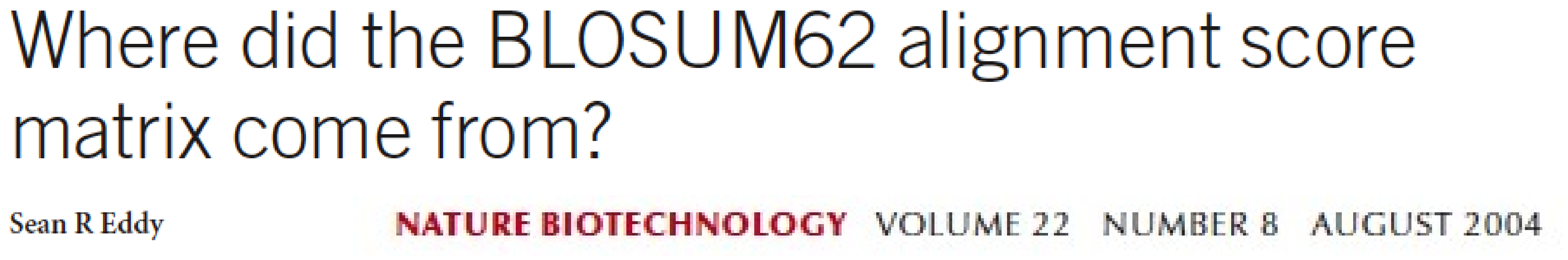
Margaret Dayhoff, founder of bioinformatics

https://www.whatisbiotechnology.org/index.php/people/summary/Dayhoff

https://www.smithsonianmag.com/science-nature/how-margaret-dayhoff-helped-bring-computing-scientific-research-180971904/

- **1948**: PhD in quantum chemistry
- **1961** and onward: worked as evolutionary biologist, and developed the first substitution matrices used to judge the significance of alignments between protein sequences

  These are the PAM matrices (Position Accepted Mutation), available today for your BLAST searches
- Invented the single letter codes for amino acids (very important for computation)
- **1965**: Initiated first collection of protein sequences, “Atlas of Protein Sequences and Structure”
- **1966**: with Richard Eck, created first computationally derived phylogenetic tree reconstruction
- **1971**: Created first computer database for protein sequences (the Protein Information Resource)
- **1980**: made largest nucleotide sequence database freely available by telephone network



Where did the BLOSUM62 alignment score matrix come from?

Sean R Eddy

NATURE BIOTECHNOLOGY VOLUME 22 NUMBER 8 AUGUST 2004

“ *...details in BLOSUM62 that may seem counterintuitive at first glance.* For instance, tryptophan (W/W) pairs score +11, while leucine (L/L) pairs only score +4; why shouldn’ t all identities get the same score? The rarer the amino acid is, the more surprising it would be to see two of them align together by chance. In the homologous alignment data that BLOSUM62 was trained on, leucine/leucine (L/L) pairs were in fact more common than tryptophan/tryptophan (W/W) pairs (*pLL* = *0.0371, pWW = 0.0065), but tryptophan* is a much rarer amino acid (*fL = 0.099, fW = 0.013). Run those numbers (with BLOSUM62* ’ s original $\lambda = 0.347$) and you get +3.8 for L/L and +10.5 for W/W, which were rounded to +4 and +11.”

BLAST scores

In any given alignment:

- matches (identical or similar) RAISE score
- mismatches LOWER score
- gaps LOWER score

Three criteria are used in the display of the highest scoring ('best') sequences:

1) percent identity
2) similarity score
3) E-value--probability that two sequences will have the similarity they have by chance (lower numbers mean a higher probability of evolutionary homology, and so a higher probability of similar function)

What is the E-value? (https://youtu.be/nO0wJgZRZJs)

The E value (also called Expect value) represents the chance that the similarity is random and therefore insignificant.

…the E value describes the random background noise that exists for matches between sequences. For example, an E value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance.

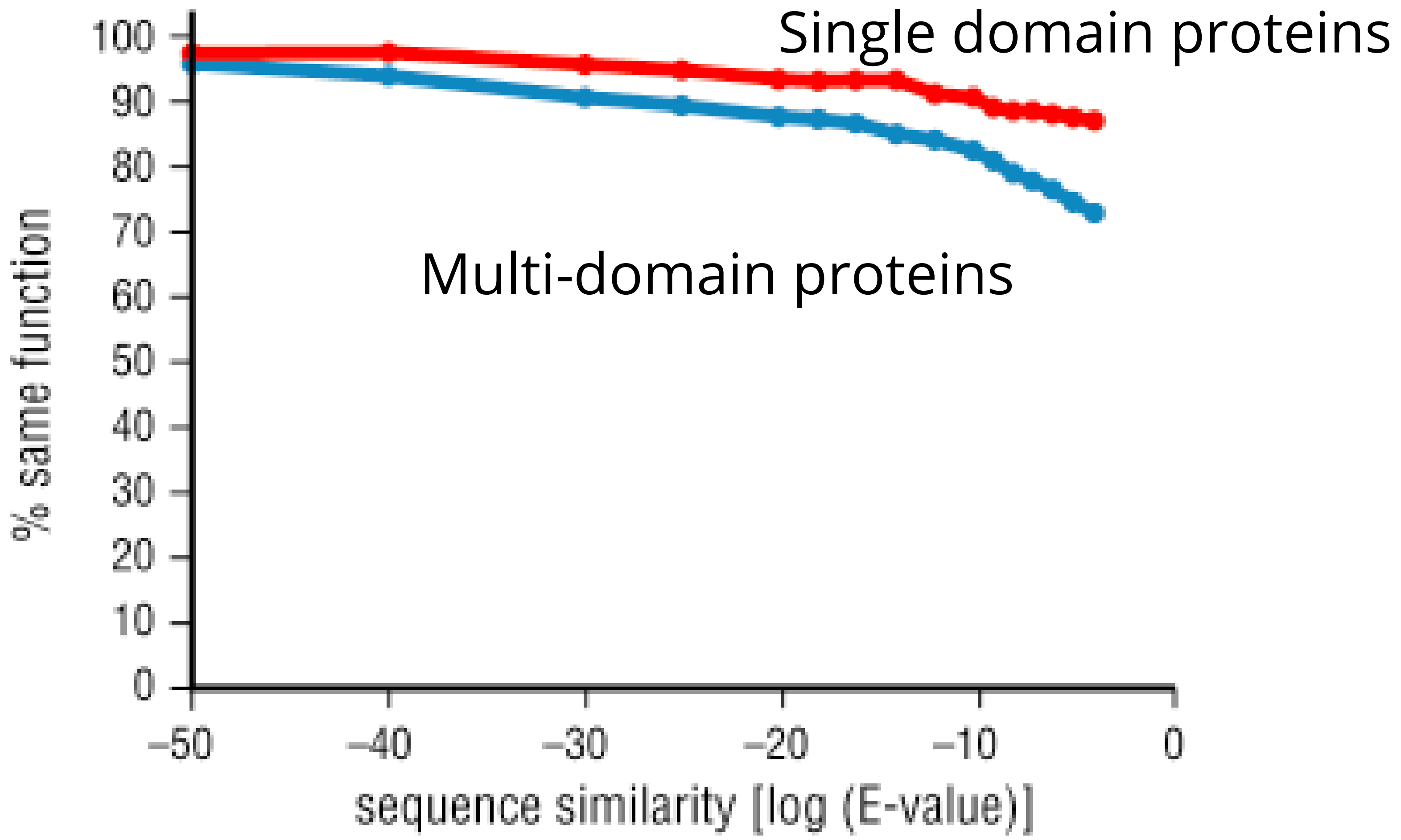
You can change the Expect value threshold on most main BLAST search pages. When the Expect value is increased from the default value of 10, a larger list with more low-scoring hits will be listed.

E values (continued)

From the BLAST tutorial:

Although hits with E values much higher than 0.1 are unlikely to reflect true sequence relatives, it can be useful to examine hits with lower significance (E values between 0.1 and 10) for short regions of similarity. In the absence of longer similarities, these short regions may allow the tentative assignment of biochemical activities to the ORF in question. The significance of any such regions must be assessed on a case by case basis.

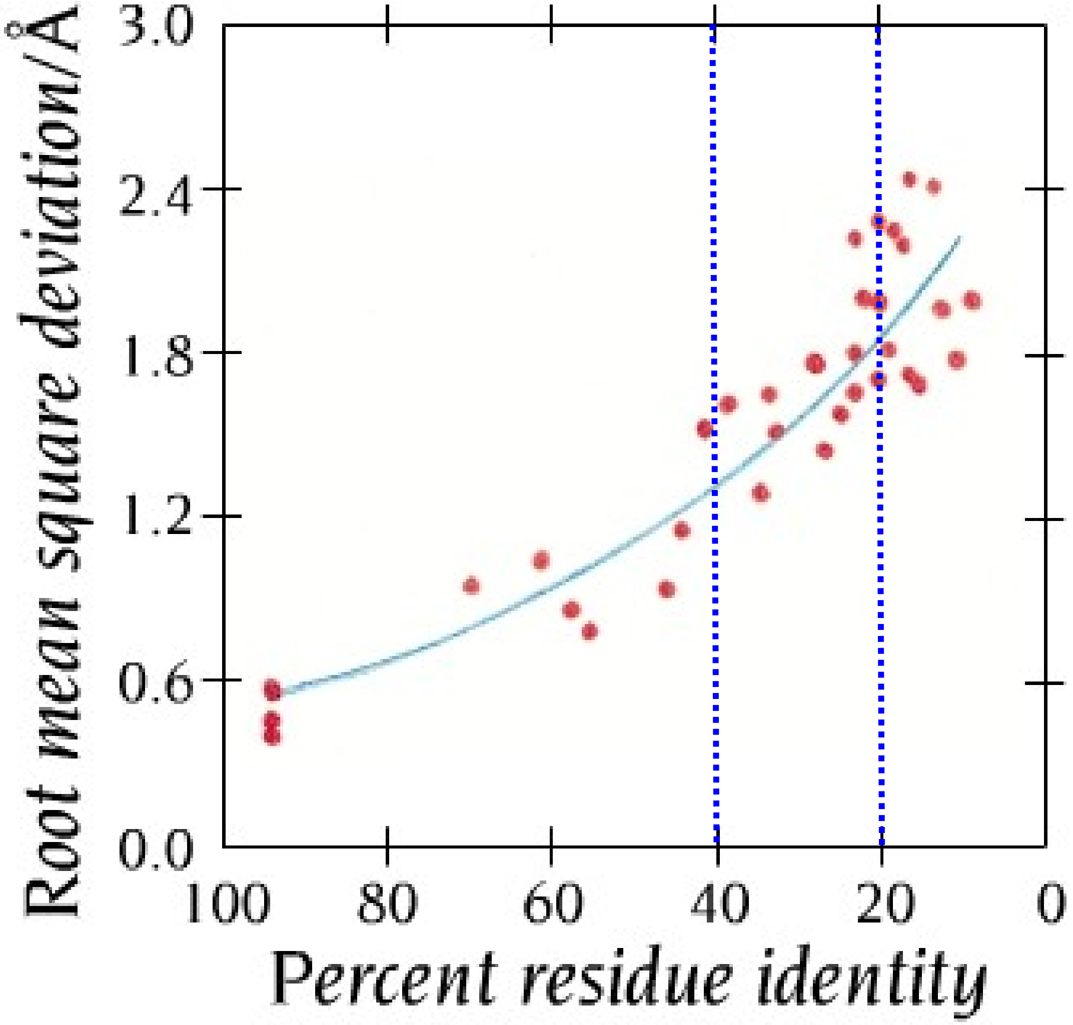
Relationship between E-value and function



E value greater than 10^{-10}, could be a similar structure but may have a different function

Similar sequence = similar structure

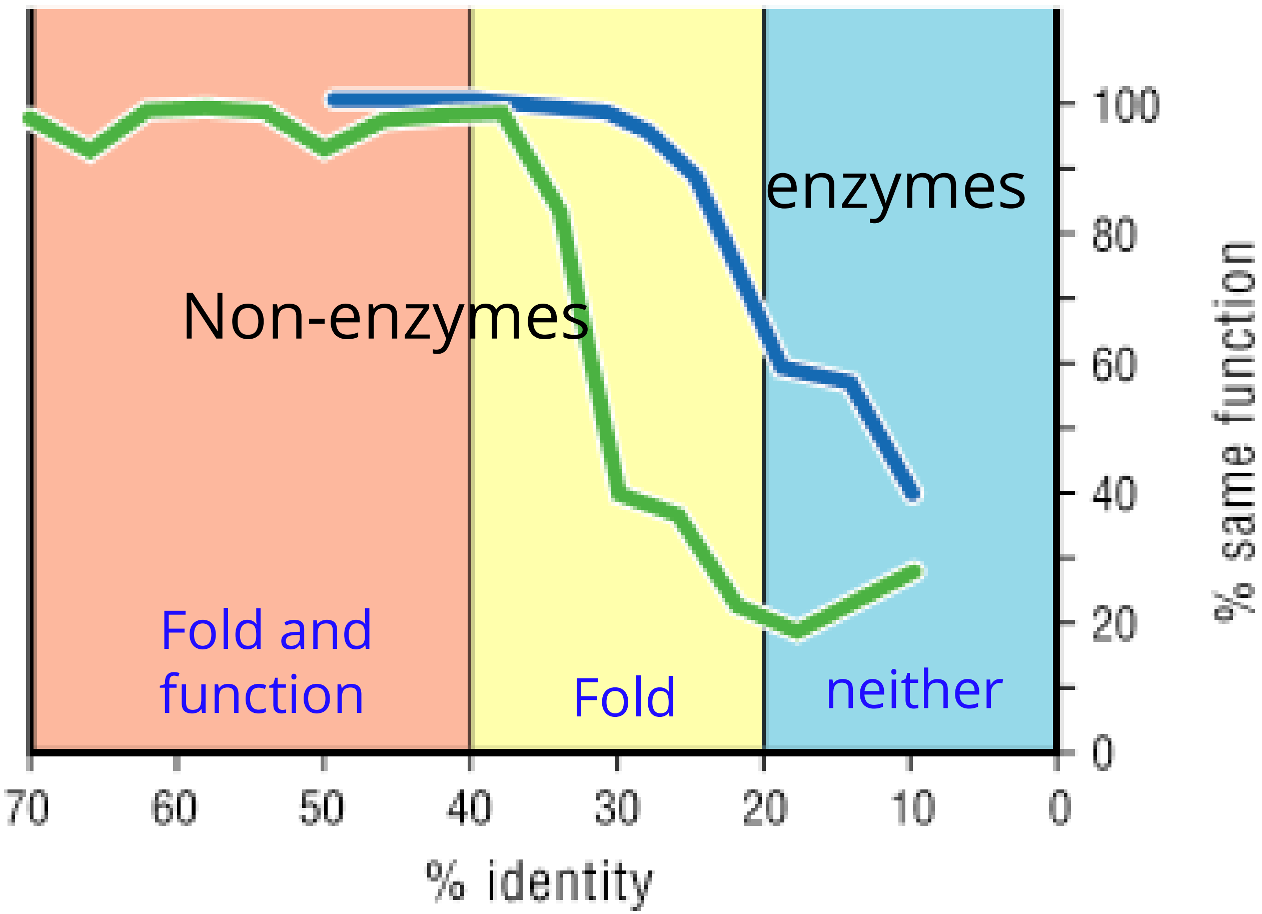
Deviation in main chain atoms in protein cores increases as percent identity of protein sequence decreases



(Plot calculated from the known structures of 32 pairs of homologous proteins)

High sequence similarity correlates with functional similarity

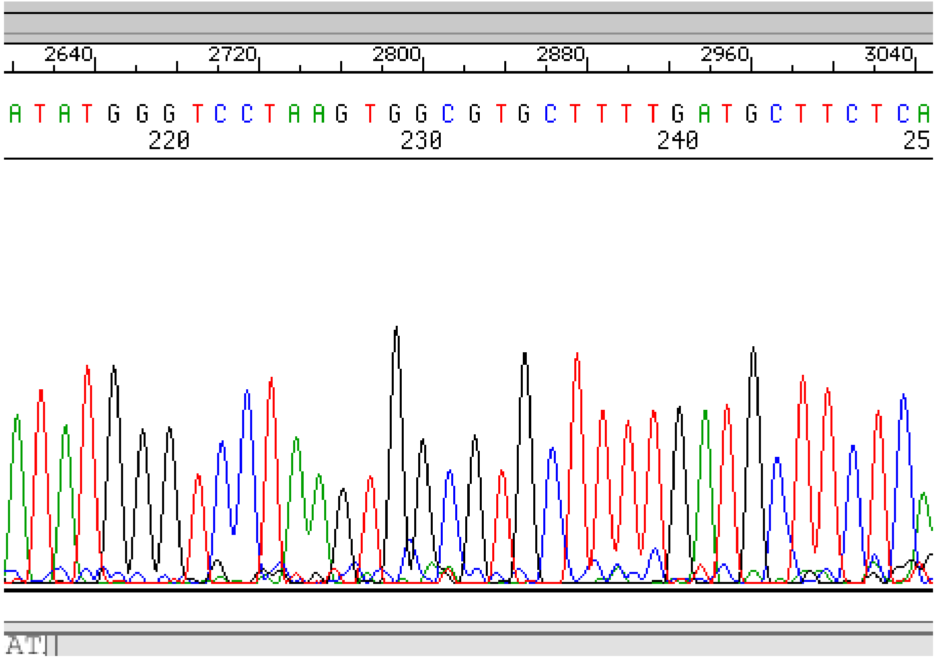
Reliable prediction?

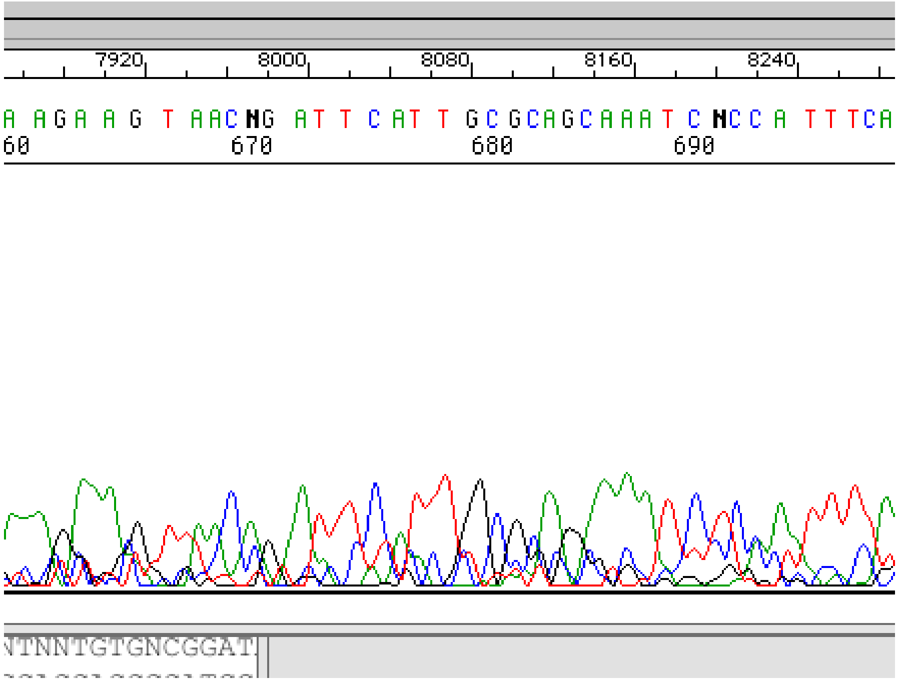


40-20% identity: fold can be predicted by similarity but precise function cannot be predicted (the 40% rule)

Biological function of a new sequence? BLAST…..

Raw data





Computer calls GNNTNNTGTGNCGGATACAATTCCCCTCTAGAAATAATT
TTGTTTAACTTTAAGAAGGAGATATACATATGCACCACCAC
CACCACCACCCCATGGGTATGAATAAGCAAAAGGTTTGTCCTGCTTGTGAATCTGCGGAACTTATTTATGATCCAGAAAG
GGGGGAAATAGTCTGTGCCAAGTGCGGTTATGTAATAGAAGAGAACATAATTGATATGGGTCCTAAGTGGCGTGCTTTTG
ATGCTTCTCAAAGGGAACGCAGGTCTAGAACTGGTGCACCAGAAAGTATTCTTC
TTCATGACAAGGGGCTTTCAACTGCA
ATTGGAATTGACAGATCGCTTTCCGGATTAATGAGAGAGAAGATGTACCGTTTGAGGAAGTGGCANTCCANATTANGAGT
TAGTGATGCAGCANANAGGAACCTAGCTTTTGCCCTAAGTGAGTTGGATAGAATTNCTGCTCAGTTAAAACTTCCNNGAC
ATGTAGAGGAAGAAGCTGCAANGCTGNACANAGANGCAGNGNGANAGGGACTTATTNGANGCAGATCTATTGAGAGCGTT
ATGGCGGCANGTGTTTACCCTGCTTGTAGGTTATTAAAAGNTCCCGGGACTCTGGATGAGATTGCTGATATTGCTAGAGC

Find the open reading frame(s) and translate

```
MKCPYCKSRDLVYDRQHGEVFCKKCGSILATNLVDSELSRKT
KTNDIPRYTKRIGEFTREKIYRLRKWQKKISSERNLVLAMSE
LRRLSGMLKLPKYVEEEAAYLYREAAKRGLTRRIPIETTVAA
CIYATCRLFKVPRTLNEIASYSKTEKKEIMKAFRVIVRNLNL
TPKMLLARPTDYVDKFADELELSERVRRRTVDILRRANEEGI
TSGKNPLSLVAAALYIASLLEGERRSQKEIARVTGVSEMTVR
NRYKELA
```

Query Sequence in FASTA Format

Amino acid sequence of a protein, in FASTA format:

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

Nucleotide sequence of a gene, in FASTA format:

```
>gi|118139508:333094-333465 Thiomicrospira crunogena XCL-2
ATGGCAATTACAAAAGACGATATTTTAGAAGCAGTTGCTAACATGTCAGTAATGGAAGTTGT
TGAACTTGTTGAAGCAATGGAAGAGAAGTTTGGTGTTTCTGCAGCAGCAGTTGCGGTTGCAG
GTCCTGCAGGTGATGCTGGCGCTGCTGGTGAAGAACAAACAGAGTTTGACGTTGTCTTGACT
GGTGCTGGTGACAACAAAGTTGCAGCAATCAAAGCCGTTCGTGGCGCAACTGGTCTTGGGCT
TAAAGAAGCGAAAAGTGCAGTTGAAAGTGCACCATTTACGCTTAAAGAGGGTGTTTCTAAAG
AAGAAGCAGAAACTCTTGCAAATGAGCTTAAAGAAGCAGGTATTGAAGTCGAAGTTAAATAA
```

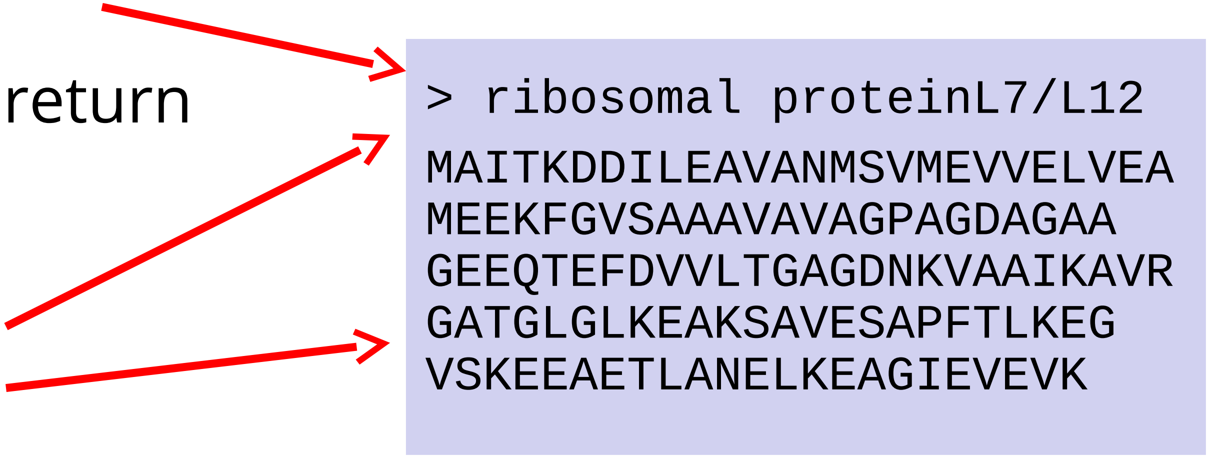
Kerfeld and Scott, PLoS
Biology 2011

The FASTA format

” description line” (not read as sequence data)

- Begins with >
- Ends with a hard return

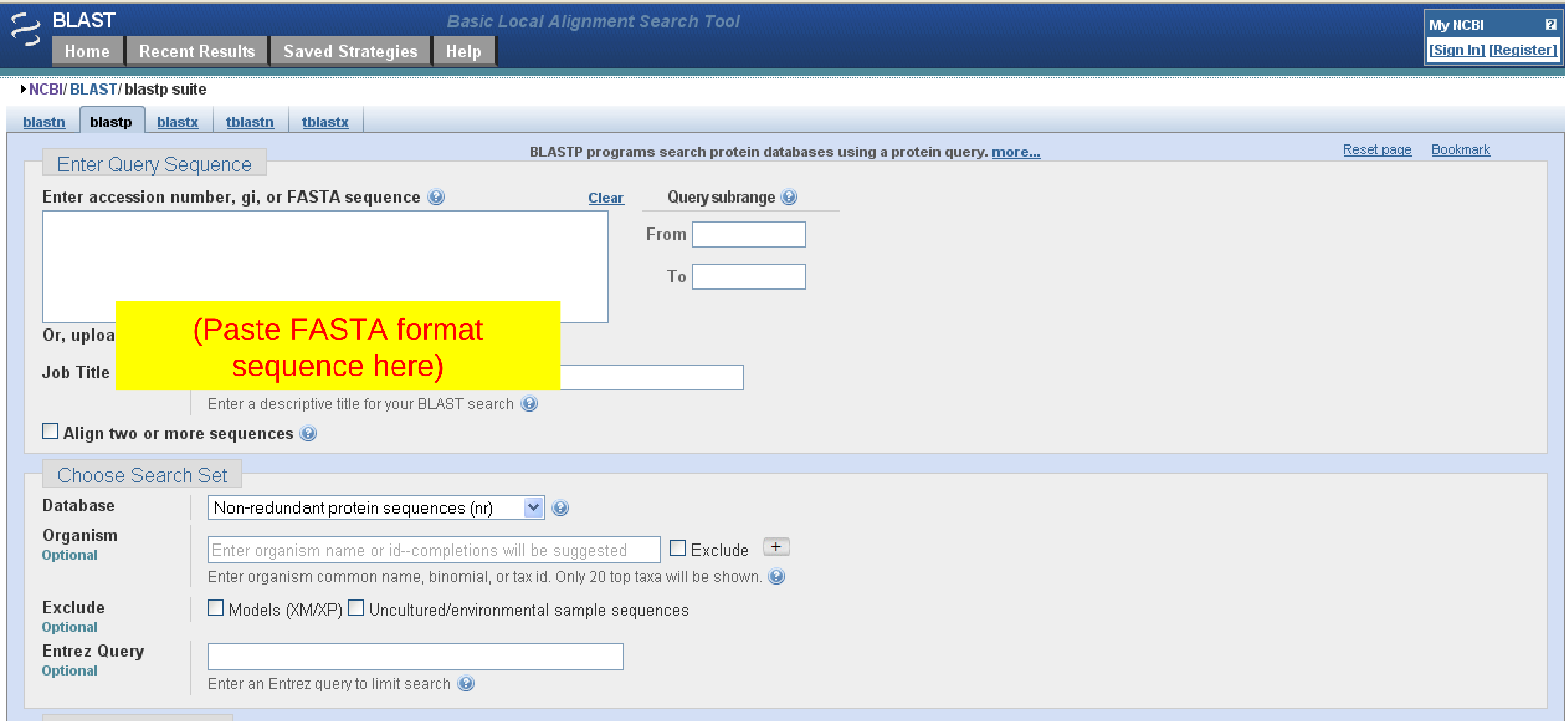
Sequence data (amino acid in this case)



```
> ribosomal proteinL7/L12
MAITKDDILEAVANMSVMEVVELVEA
MEEKFGVSAAAVAVAGPAGDAGAA
GEEQTEFDVVLTGAGDNKVAAIKAVR
GATGLGLKEAKSAVESAPFTLKEG
VSKEEAETLANELKEAGIEVEVK
```

Kerfeld and Scott, PLoS Biology 2011

NCBI BLAST Interface (blastp: Proteins)



BLASTP programs search protein databases using a protein query. more...

Enter accession number, gi, or FASTA sequence

Query subrange

From

To

Or, uploa

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be suggested □ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query (Optional)

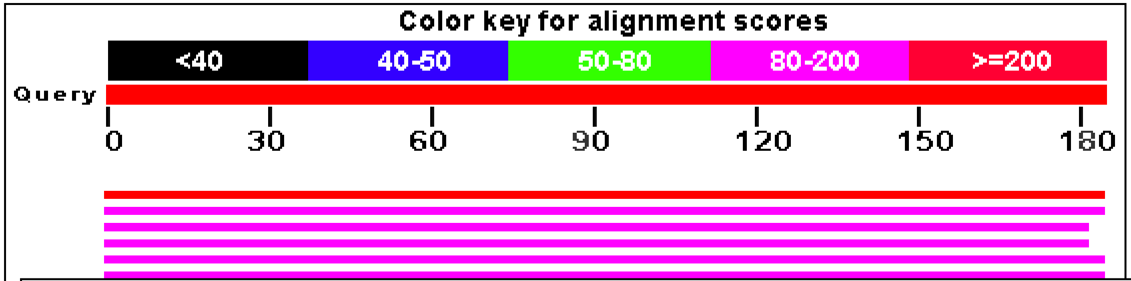
Enter an Entrez query to limit search

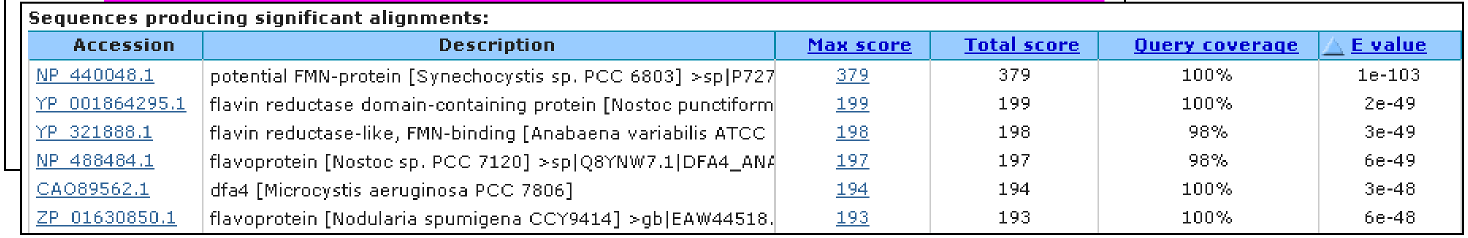
(Paste FASTA format sequence here)

Kerfeld and Scott, PLoS Biology 2011

NCBI BLAST Results Page:

Potential homologs retrieved from database

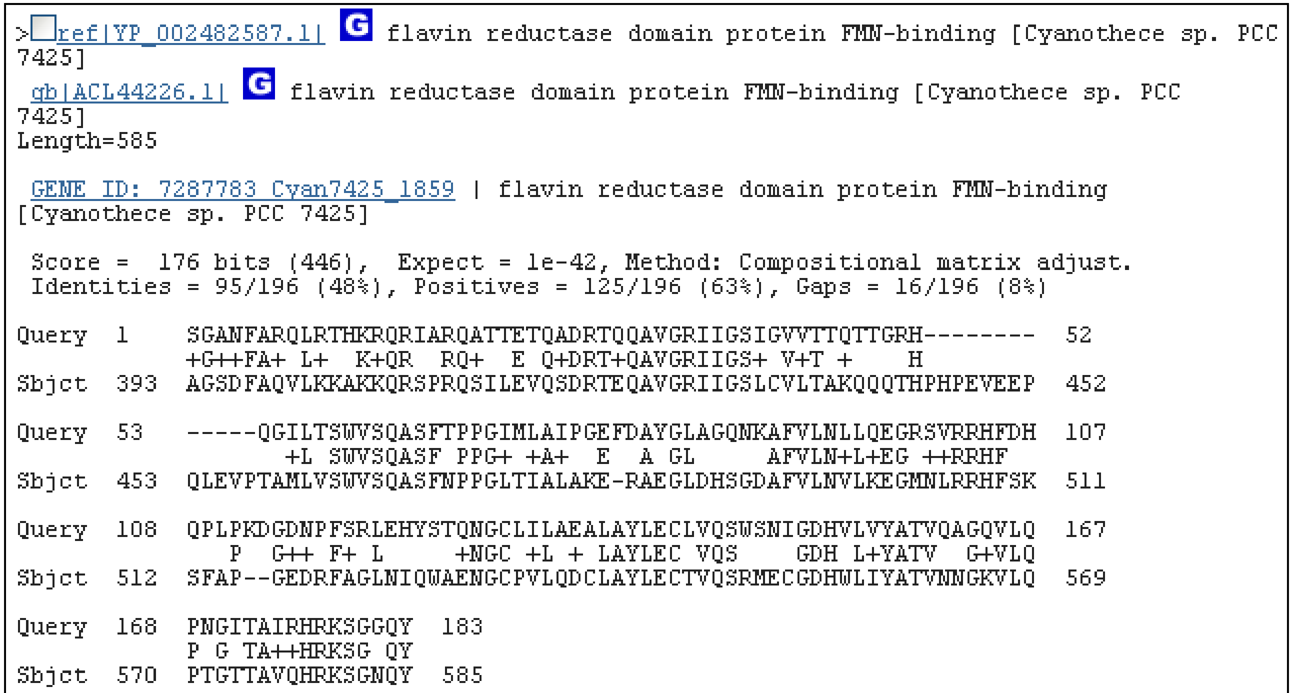




Sequences producing significant alignments:

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_001864295.1	flavin reductase domain-containing protein [Nostoc punctiform	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|DFA4_ANA	197	197	98%	6e-49
CAO89562.1	dfa4 [Microcystis aeruginosa PCC 7806]	194	194	100%	3e-48
ZP_01630850.1	flavoprotein [Nodularia spumigena CCY9414] >gb|EAW44518.	193	193	100%	6e-48

Kerfeld and Scott, PLoS Biology 2011



```
>□ref|YP_002482587.1| G flavin reductase domain protein FMN-binding [Cyanothece sp. PCC
7425]
 gb|ACL44226.1| G flavin reductase domain protein FMN-binding [Cyanothece sp. PCC
7425]
Length=585

 GENE ID: 7287783 Cyan7425_1859 | flavin reductase domain protein FMN-binding
[Cyanothece 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    SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVVTTQTTGRH--------  52
            +G++FA+ L+  K+QR  RQ+  E Q+DRT+QAVGRIIGS+ V+T +     H
Sbjct  393  AGSDFAQVLKKAKKQRSPRQSILEVQSDRTEQAVGRIIGSLCVLTAKQQQTHPHPEVEEP  452

Query  53   -----QGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFVLNLLQEGRSVRRHFDH  107
                  +L SWVSQASF PPG+ +A+   E   A GL     AFVLN+L+EG ++RRHF
Sbjct  453  QLEVPTAMLVSWVSQASFNPPGLTIALAKE-RAEGLDHSGDAFVLNVLKEGMNLRRHFSK  511

Query  108  QPLPKDGDNPFSRLEHYSTQNGCLILAEALAYLECLVQSWSNIGDHVLVYATVQAGQVLQ  167
               P  G++ F+ L      +NGC +L + LAYLEC VQS     GDH L+YATV  G+VLQ
Sbjct  512  SFAP--GEDRFAGLNIQWAENGCPVLQDCLAYLECTVQSRMECGDHWLIYATVNNGKVLQ  569

Query  168  PNGITAIRHRKSGGQY  183
            P G TA++HRKSG QY
Sbjct  570  PTGTTAVQHRKSGNQY  585
```

BLAST page: https://blast.ncbi.nlm.nih.gov/Blast.cgi

-- go to "protein blast"

-- Program: blastp

-- Search set: leave blank, or choose taxonomic group (example: archaea) https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi

-- choose algorithm: blastp (default)

--PSI-, PHI-, and DELTA-BLAST are variants that can find homologs with relatively low sequence identity (distant homologs), DELTA- is most modern and seems to work best (esp. cross-domain homology)

-- click " BLAST" button

View Report:

1) Distribution of hits: query sequence followed color-coded positions 'hit' sequences that gave alignments
2) Sequences producing significant alignments (sorted by Max score)
   - Accession number (this takes you to the sequence that yielded the hit: gene or contig)
   - Description: gene name, organism
   - Max score, total score, query coverage
   - E-value

DELTA-BLAST:

- Domain Enhanced Lookup Time Accelerated BLAST

- The algorithm constructs a specific PSSM (position-specific scoring matrix) using the results of a Conserved Domain Database (CDD) search

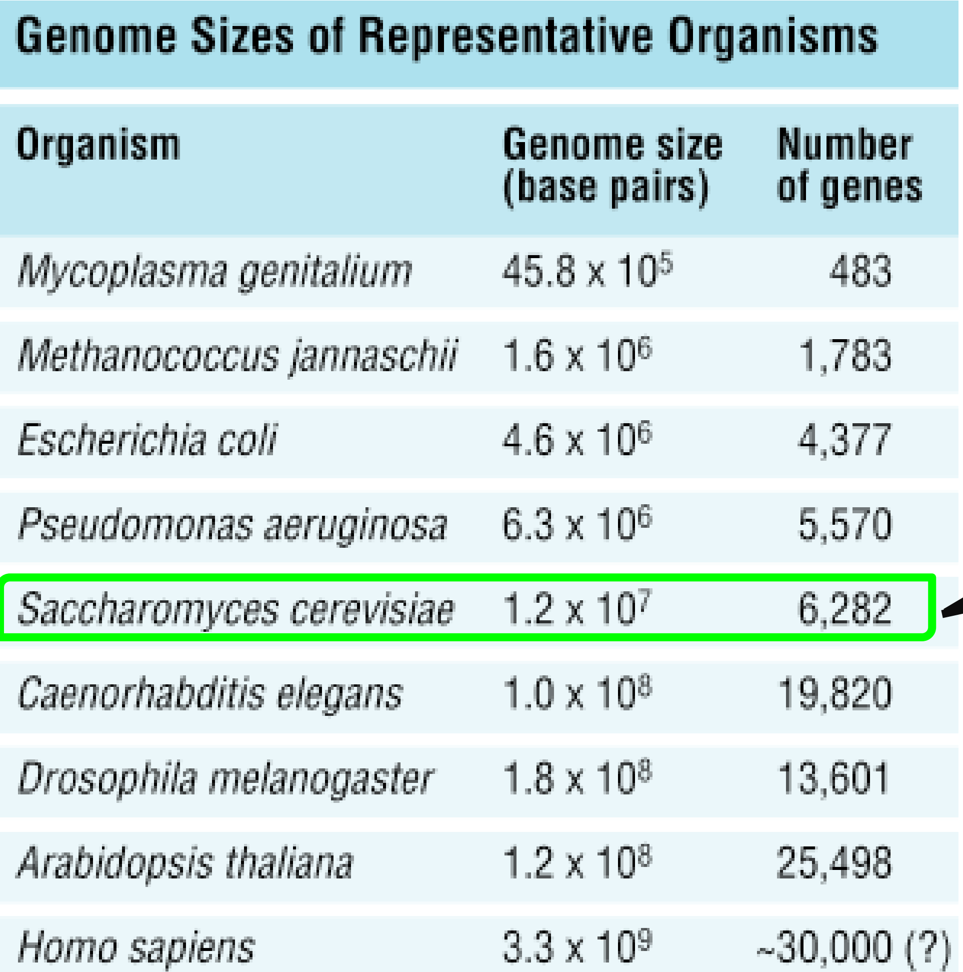
- A PSSM differs from typical PAM and BLOSUM matrices in which scores are position independent

- The constructed PSSM is used to search sequence database

BLAST is good, but not perfect

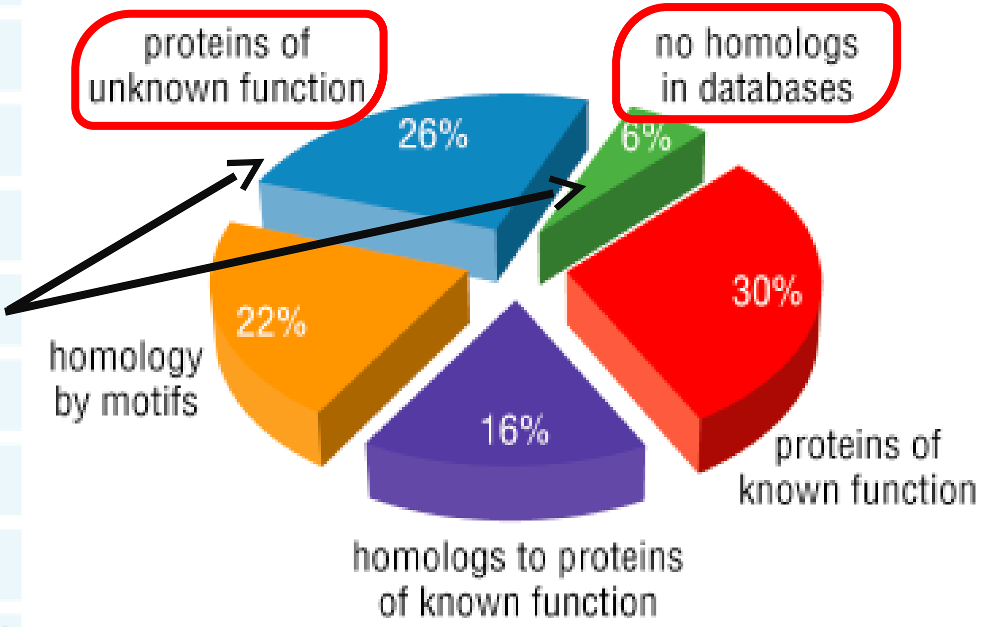
1) High homology? *function is inferred* but not known

2) Many coding proteins *cannot be assigned function* based on homology



Genome Sizes of Representative Organisms

Organism	Genome size (base pairs)	Number of genes
Mycoplasma genitalium	45.8×10^{5}	483
Methanococcus jannaschii	1.6×10^{6}	1,783
Escherichia coli	4.6×10^{6}	4,377
Pseudomonas aeruginosa	6.3×10^{6}	5,570
Saccharomyces cerevisiae	1.2×10^{7}	6,282
Caenorhabditis elegans	1.0×10^{8}	19,820
Drosophila melanogaster	1.8×10^{8}	13,601
Arabidopsis thaliana	1.2×10^{8}	25,498
Homo sapiens	3.3×10^{9}	~30,000 (?)



Bioinformatics:
making sense of biological sequence

- New DNA sequences are analyzed for ORFs (Open Reading Frames: protein)

- Any DNA or protein sequence can then be compared to all other sequences in databases, and similar sequences identified

- There is much more -- a huge number of bioinformatics tools are available