

Eric Nawrocki

Scientist - National Library of Medicine



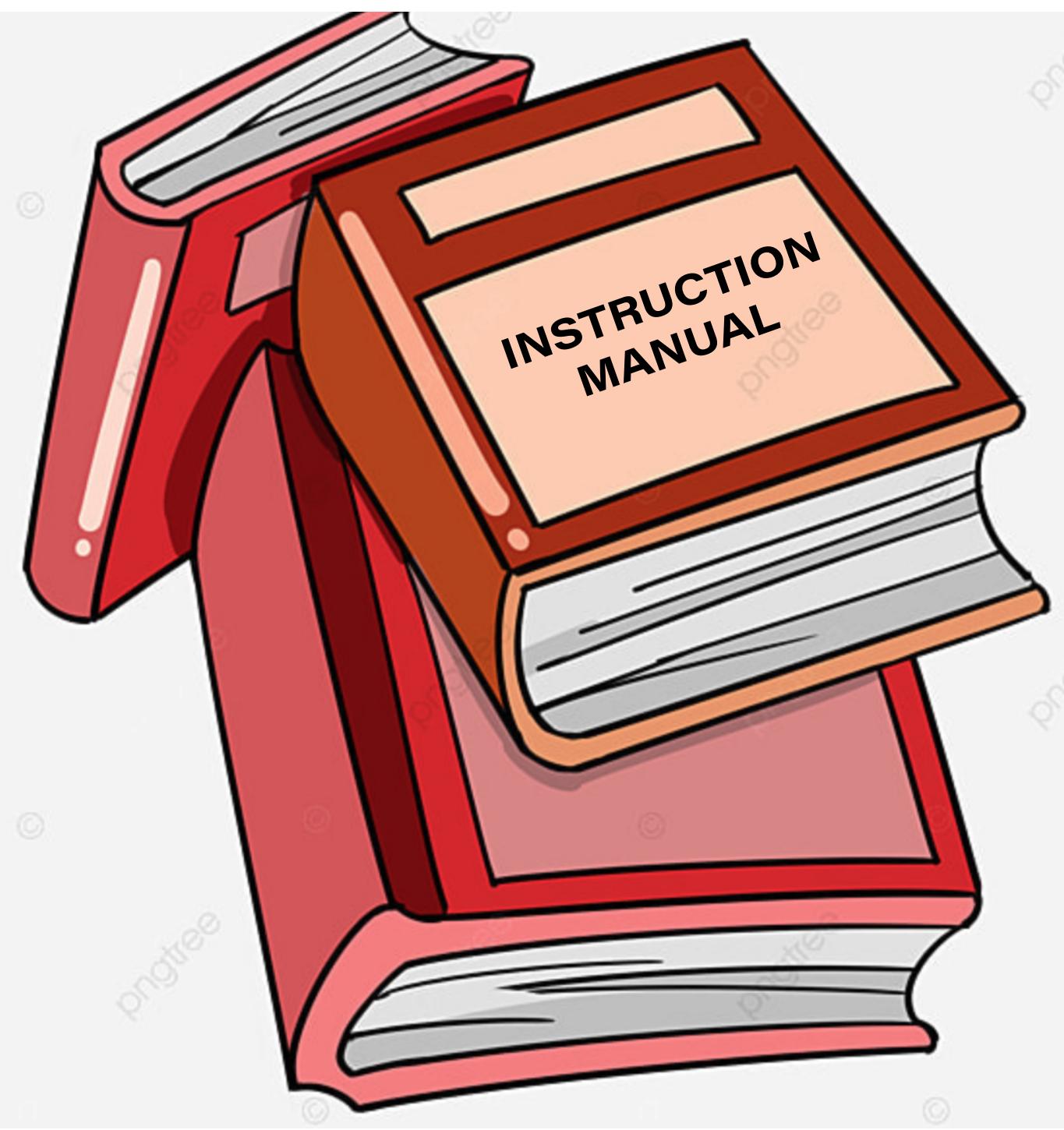
Career day advice from my daughter:

“Bring candy”



# Library: scientific books and journals







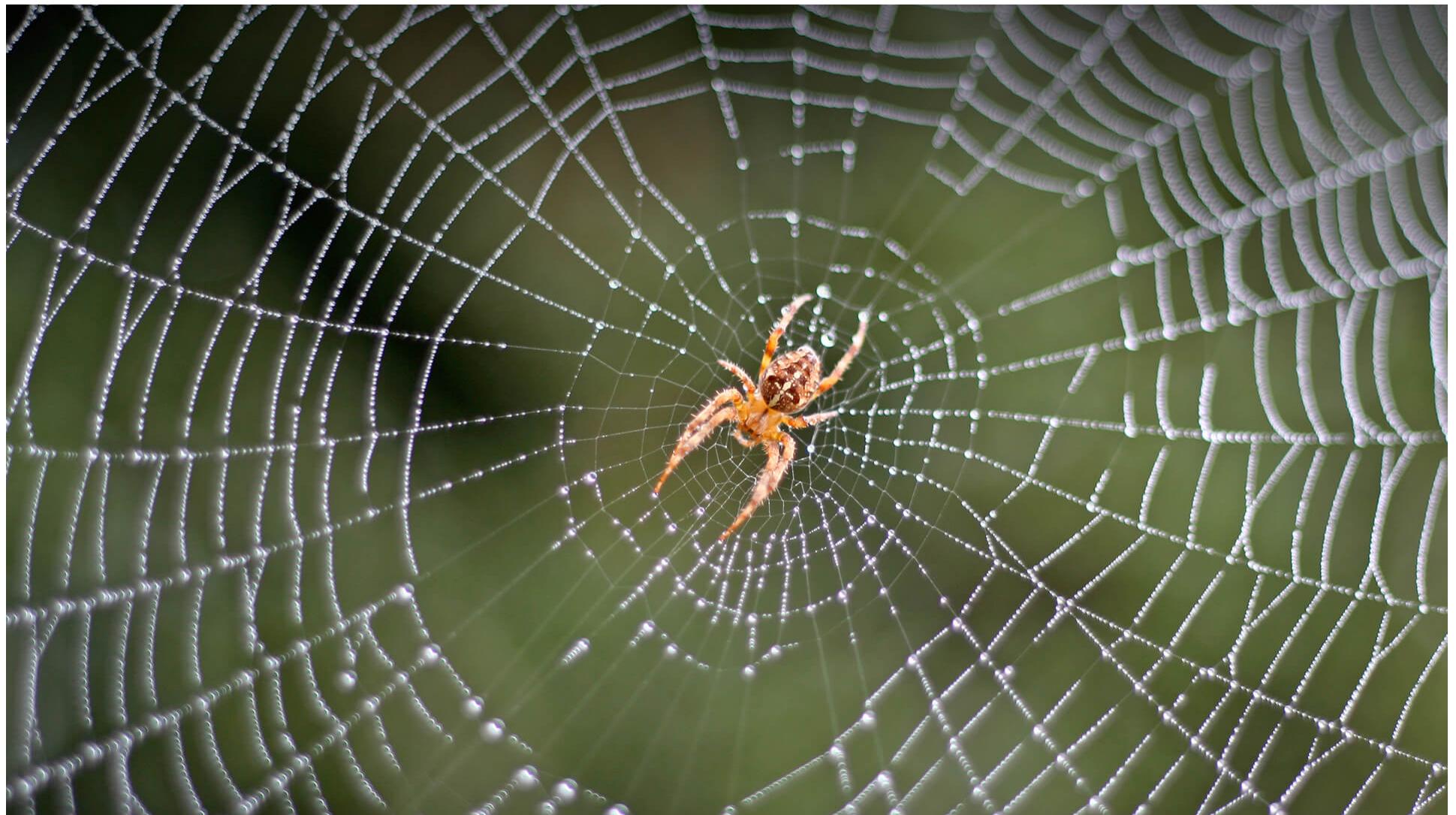






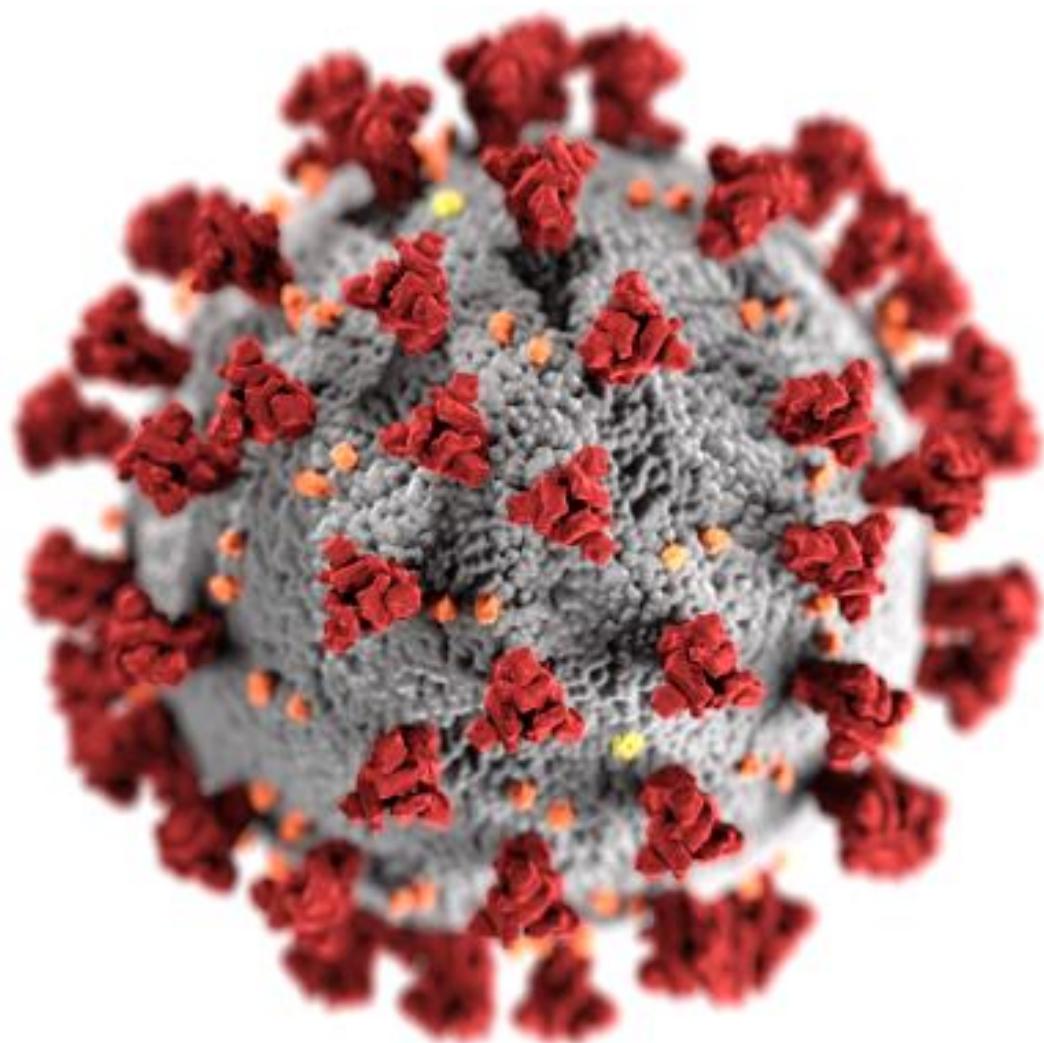












DEC. 25, 2023

# PERSON OF THE YEAR | TAYLOR SWIFT





What language stores the information of life?

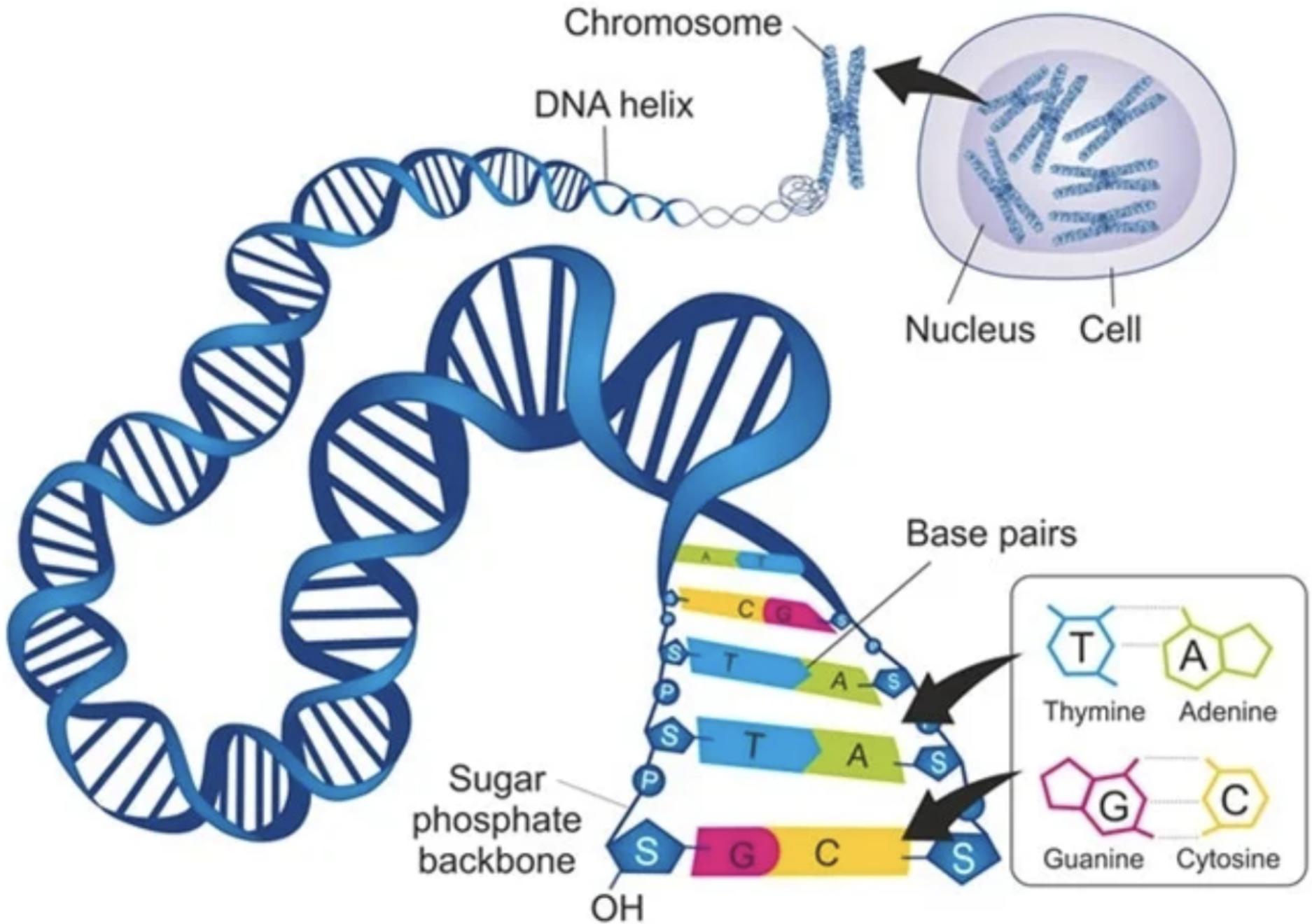
What is the alphabet of DNA?

- A, C, G, and T are the four bases (nucleotides) of DNA

- A, C, G, and T are the four bases (nucleotides) of DNA
- DNA is made up of two linked strands in a twisted ladder shape

- A, C, G, and T are the four bases (nucleotides) of DNA
- DNA is made up of two linked strands in a twisted ladder shape
- One base on each strand pairs with a base on the other strand

- A, C, G, and T are the four bases (nucleotides) of DNA
- DNA is made up of two linked strands in a twisted ladder shape
- One base on each strand pairs with a base on the other strand
- Which bases form pairs?

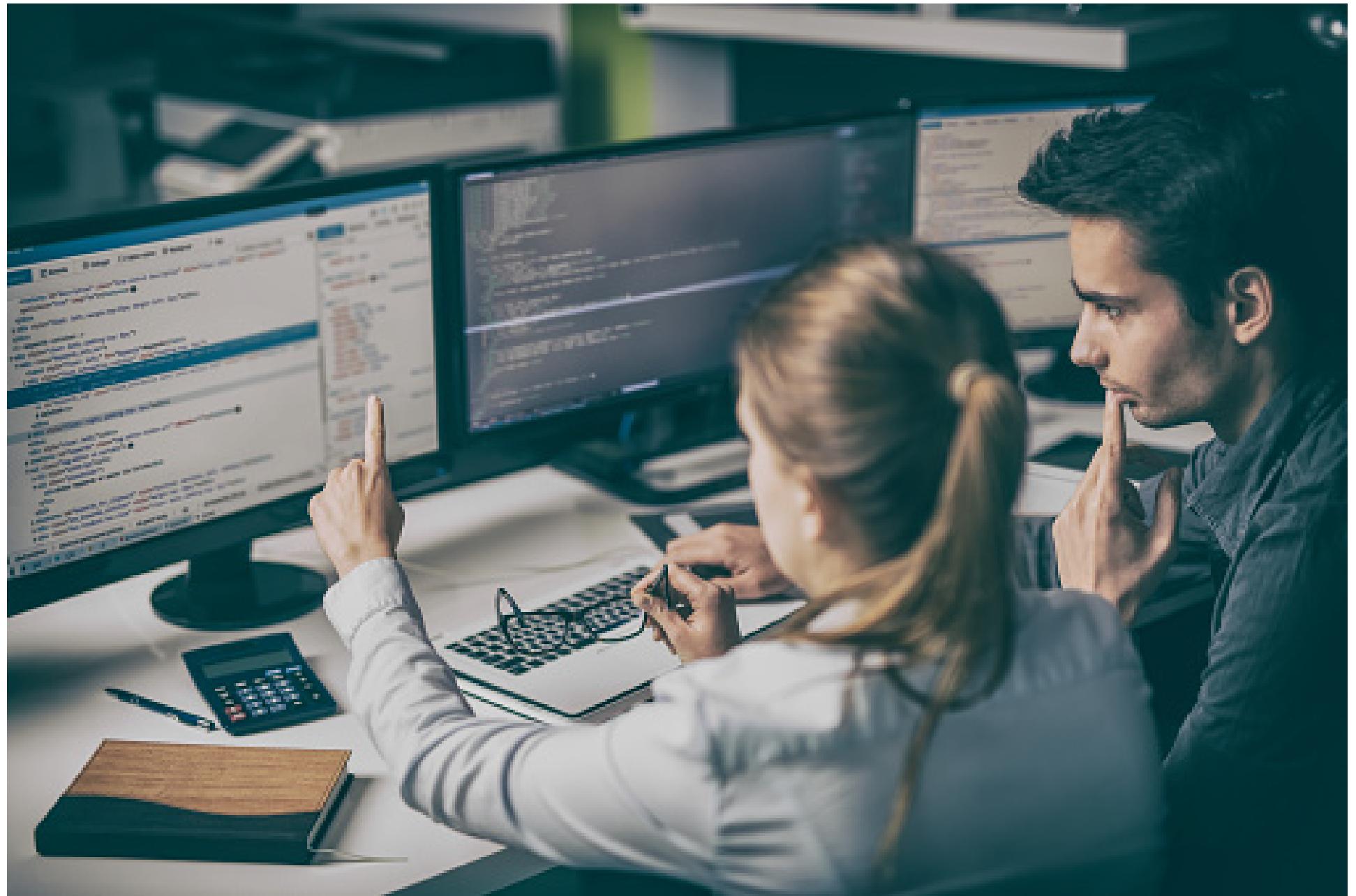


A C G T

GTAACCGTAAATAATAACTTTGAAGTCTAAGCTCATCATATCTATTCA  
TCCTTGATTCAAGACATTTTAAAAAATGCGCAATCACTATAAACCA  
TATCGATTAATGCGAATAACTATATTCTAGAACCTAGAAAAATCATTCC  
CACATCTTGAGAAAAACTTCCCCTAAATTGCTAGCGTGCATCTAACACGT  
GACTTCTTAATCTAACTTGGTAAAGTGCTGGTCTTGCCTCGTAGC  
TCAGCAGGATAGAGCGGTTGCCTCCTAACGCAGCAGGCCATGCCTCGAAT  
CGCATCGAGGACGATTTTGCCTTAACCTCTAAAGTACTAATTGCTT  
GTATCTGTGGTTACGTATTTAGCGATATTCTGTTGGTTCTGAAAA  
CTAGGTCCAGAAAGAAAATTATGAACCTCCCTCGCGATGCCTTCGCTAC  
ATGCATACGATAGCGAGCATCTGCAGGGCCGCACGTTCACGACTATTGG  
ATAAAAACCGGTTCCACCAAAACTGCAGGCATAGAAGTATCTCTAAATC  
ACAACAAAGTTCGCAGTTCAAACCTCGAGACTTCAAAATGCCATT  
TTCCATAGCAGCTAAAATGTTTCCCAGTACTTCTGACATGCGATTCC  
TAGTCGGAGATCCGACCTTACCATATAAAATATACTTCGGTGCCAAAG  
GCTGCTGCGTTGAAGAATGATTACAGTGGATGCTGATAAAGACATCCCC  
CTGCCAACGGTTCGACAAAGCAACCGCTTCCCTAACGTCAACGTATACAT  
CAGAAGATCGGGTTAGCTGAGGTTATAACCCATCCGCTTAAGTAACCT  
TGAACCGTCAAAGCAAGAGACAGGGTCAGGGACTTCTCTTCATAATGAAG  
TTCCTTACTGCCGTGCCTTGATCTTCCCCGTGTCCAGGATCTATAA  
ATATAACCTCACTGCGTCGTACACGCTGAGGAGGATTGGTGTGAGCA







# the field of bioinformatics

## Dictionary

Definitions from [Oxford Languages](#) · [Learn more](#)



**bi·o·in·for·mat·ics**

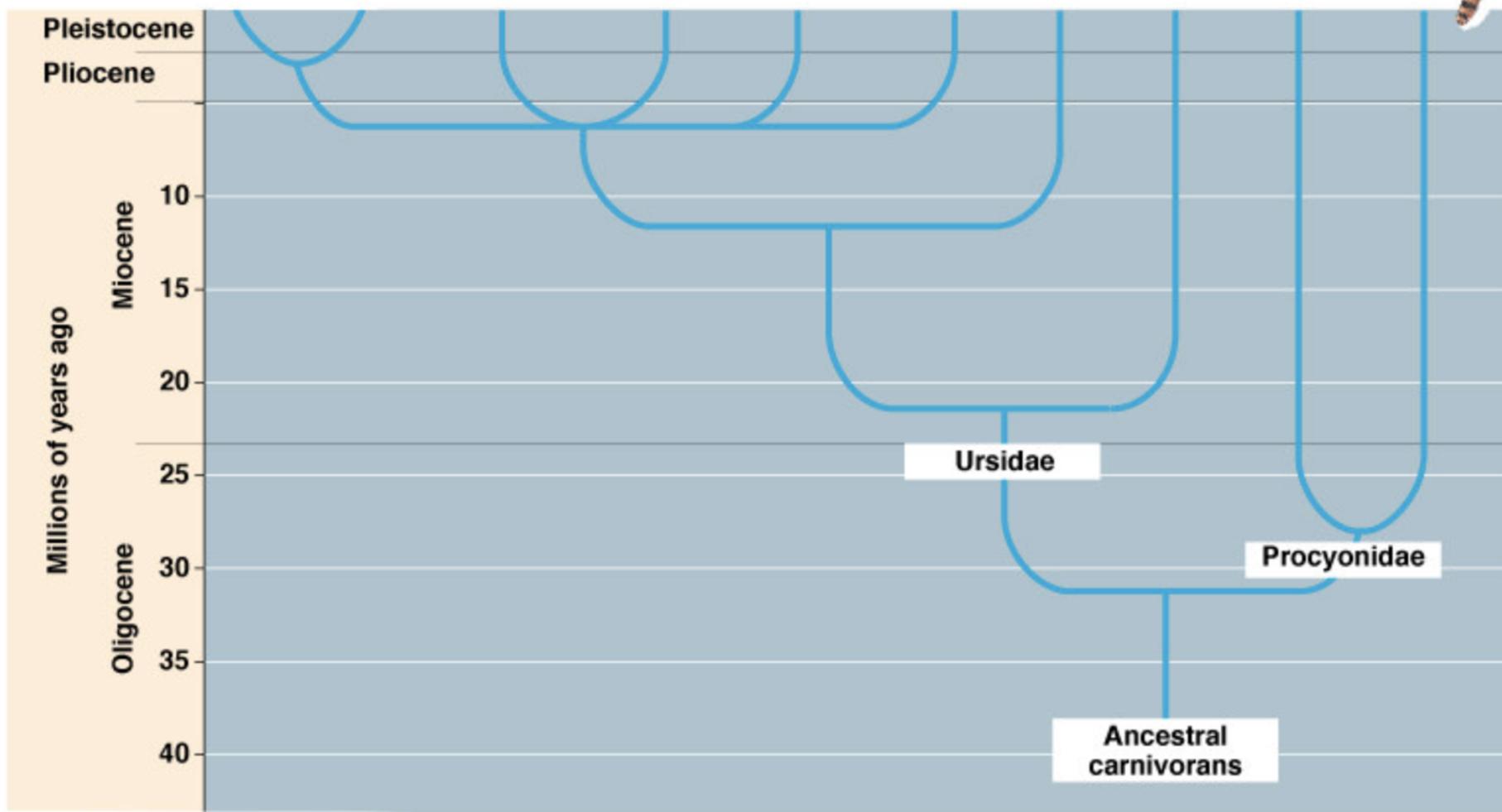
/,bīō,infər'madiks/

*noun*

the science of collecting and analyzing complex biological data such as genetic codes.







GTAACCGTAAATAATAACTTTGAAGTCTAAGCTCATCATATCTATTCA  
TCCTTGATTCAAGACATTTTAAAAAATGCGCAATCACTATAAACCA  
TATCGATTAATGCGAATAACTATATTCTAGAACCTAGAAAAATCATTCC  
CACATCTTGAGAAAAACTTCCCCTAAATTGCTAGCGTGCATCTAACACGT  
GACTTCTTAATCTAACTTGGTAAAGTGCTGGTCTTGCCTCGTAGC  
TCAGCAGGATAGAGCGGTTGCCTCCTAACGCAGCAGGCCATGCCTCGAAT  
CGCATCGAGGACGATTTTGCCTTAACCTCTAAAGTACTAATTGCTT  
GTATCTGTGGTTACGTATTTAGCGATATTCTGTTGGTTCTGAAAA  
CTAGGTCCAGAAAGAAAATTATGAACCTCCCTCGCGATGCCTTCGCTAC  
ATGCATACGATAGCGAGCATCTGCAGGGCCGCACGTTCACGACTATTGG  
ATAAAAACCGGTTCCACCAAAACTGCAGGCATAGAAGTATCTCTAAATC  
ACAACAAAGTTCGCAGTTCAAACCTCGAGACTTCAAAATGCCATT  
TTCCATAGCAGCTAAAATGTTTCCCAGTACTTCTGACATGCGATTCC  
TAGTCGGAGATCCGACCTTACCATATAAAATATACTTCGGTGCCAAAG  
GCTGCTGCGTTGAAGAATGATTACAGTGGATGCTGATAAAGACATCCCC  
CTGCCAACGGTTCGACAAAGCAACGCGTTCCCTAACGTCAACGTATACAT  
CAGAAGATCGGGTTAGCTGAGGTTATAACCCATCCGCTTAAGTAACCT  
TGAACCGTCAAAGCAAGAGACAGGGTCAGGGACTTCTCTTCATAATGAAG  
TTCCTTACTGCCGTGCCTTGATCTTCCCCGTGTCCAGGATCTATAA  
ATATAACCTCACTGCGTCGTACACGCTGAGGAGGATTGGTGTGAGCA

GTAACCGTAAATAATAACTTTGAAGTCTAAGCTCATCATATCTATTCA  
TCCTTGATTCAAGACATTTTAAAAAATGCGCAATCACTATAAACCA  
TATCGATTAATGCGAATAACTATATTCTAGAACCTAGAAAAATCATTCC  
CACATCTGAGAAAAACTTCCCCCTAAATTGCTAGCGTGCATCTAACACGT  
GACTTCTTAATCTAACTTGGTAAAGTGCTGGTCTTGCCTCGTAGC  
TCAGCAGGATAGAGCGGTTGCCTCCTAACAGCAGCAGGCC **ATGCGTTCGAAT**  
**CGCATCGAGGACGATTTTGCCTTAA**CTCCTAAAGTACTAATTGCTT  
GTATCTGTGGTTACGTATTTAGCGATATTCTGTTGGTTCTGAAAAA  
CTAGGTCCAGAAAGAAAATTATGAACCTCCCTCGGCGATGCCTTCGCTAC  
ATGCATACGATAGCGAGCATCTGCAGGGCCGCACGTTCACGACTATTGG  
ATAAAAAACCGCTTTCCACCAACACACACACACACACACACACACAC  
ACAACAAAG                   **genes** are important                   GCCATTTT  
TTCCATAGC                   **subsequences** of genomes                   TGCGATTCC  
TAGTCGGAG                   GTGCCAAAG  
GCTGCTGCCTTGAAGAATGATTACAGTGGATGCTGATAAAGACATCCCC  
CTGCCACGGTTCGACAAAGCAACCGCTTCCCTAACGTAAACGTATAACAT  
CAGAAGATCGGGTTAGCTGAGGTTATAACCCATCCGCTTAAGTAACCTT  
TGAACCGTCAAAGCAAGAGACAGGGTCAGGGACTTCTTCATAATGAAG  
TTCCTTACTGCCGTGCCTTGATCTTCCCCGTGTCCAGGATCTATAA  
ATATAACCTCACTGCGTCGTACACGCTGAGGAGGATTGGTGTGAGCA

**human**

**GCTTCGCTGC**

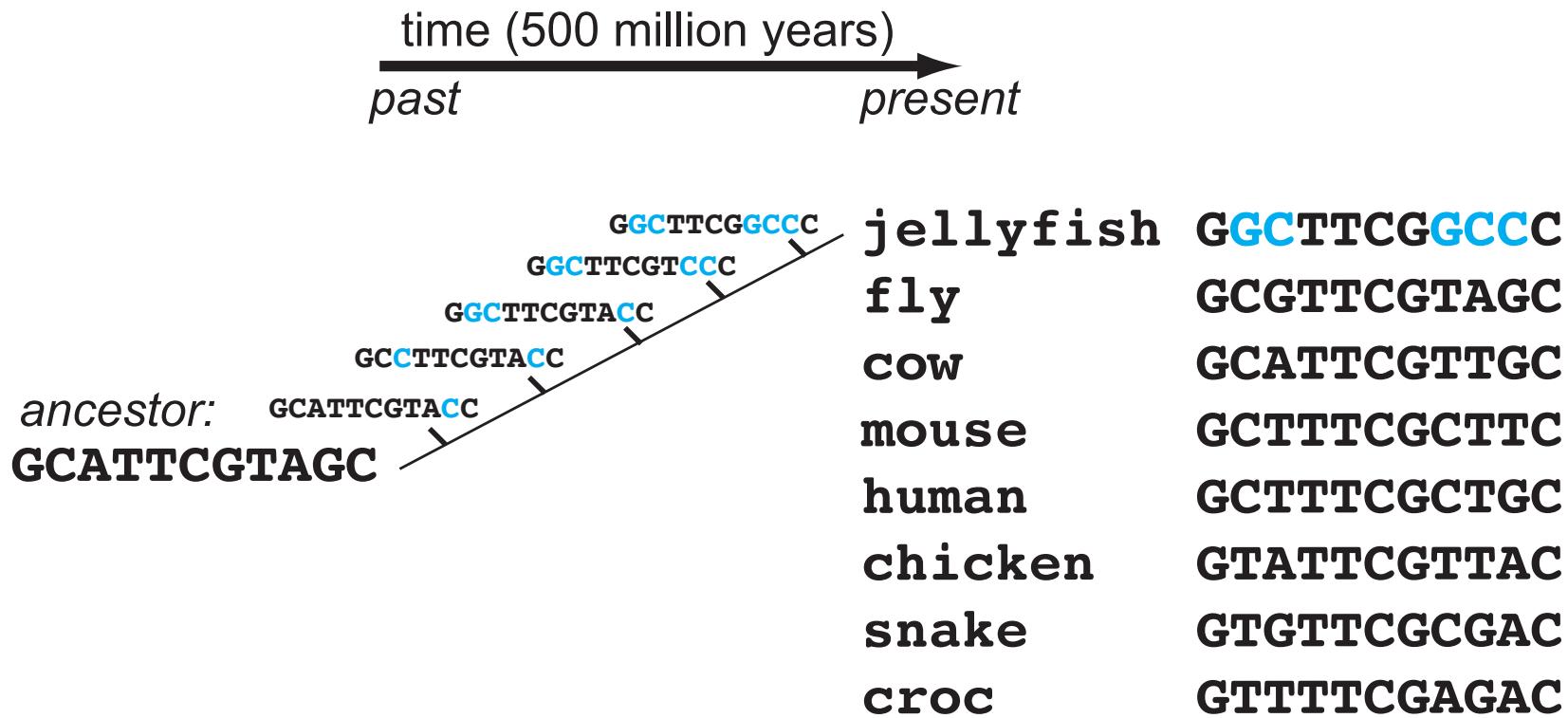
jellyfish	GGCTTCGGCCC
fly	GCCTTCGTAGC
cow	GCATTCGTTGC
mouse	GCTTCGCTTC
human	GCTTCGCTGC
chicken	GTATTCGTTAC
snake	GTGTTCGCGAC
croc	GTTTCGAGAC

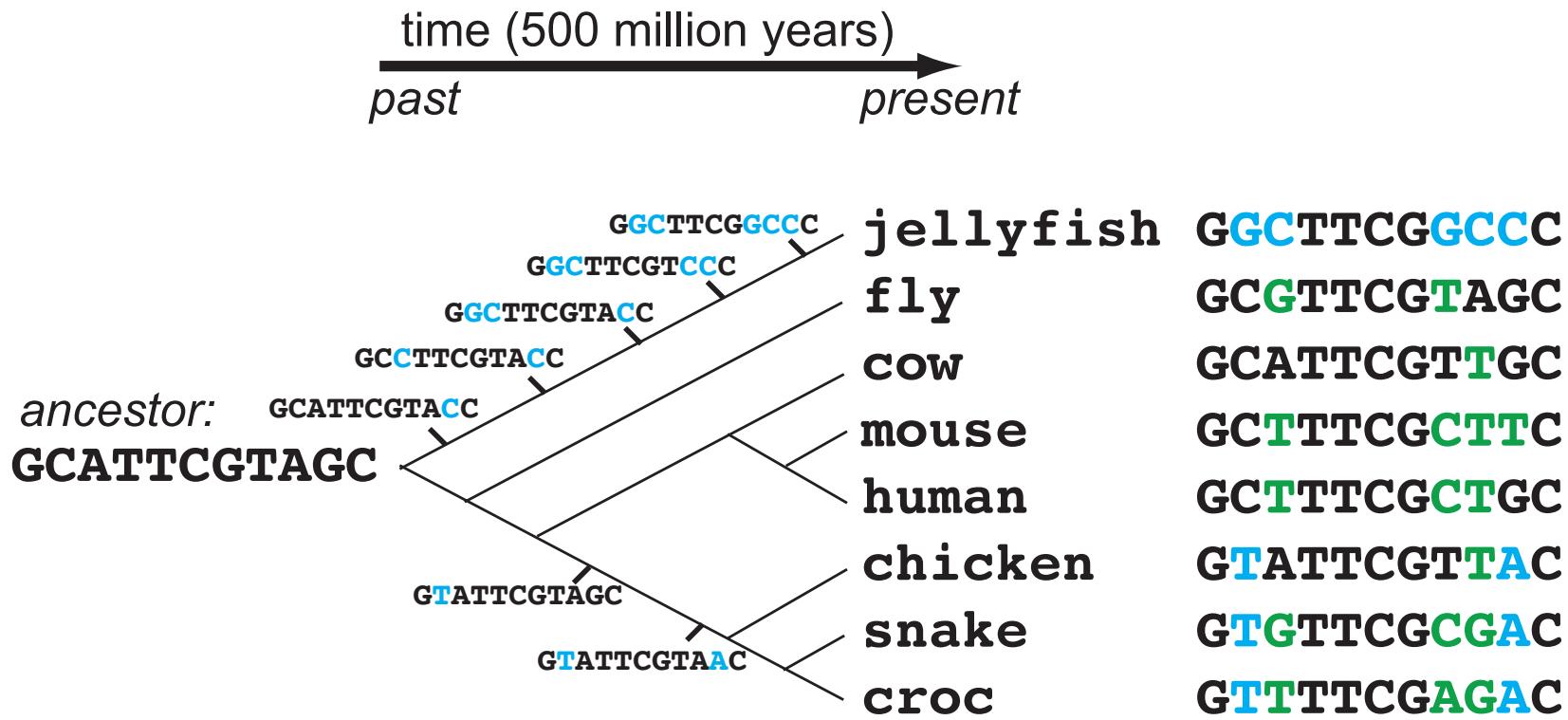
*ancestor:*

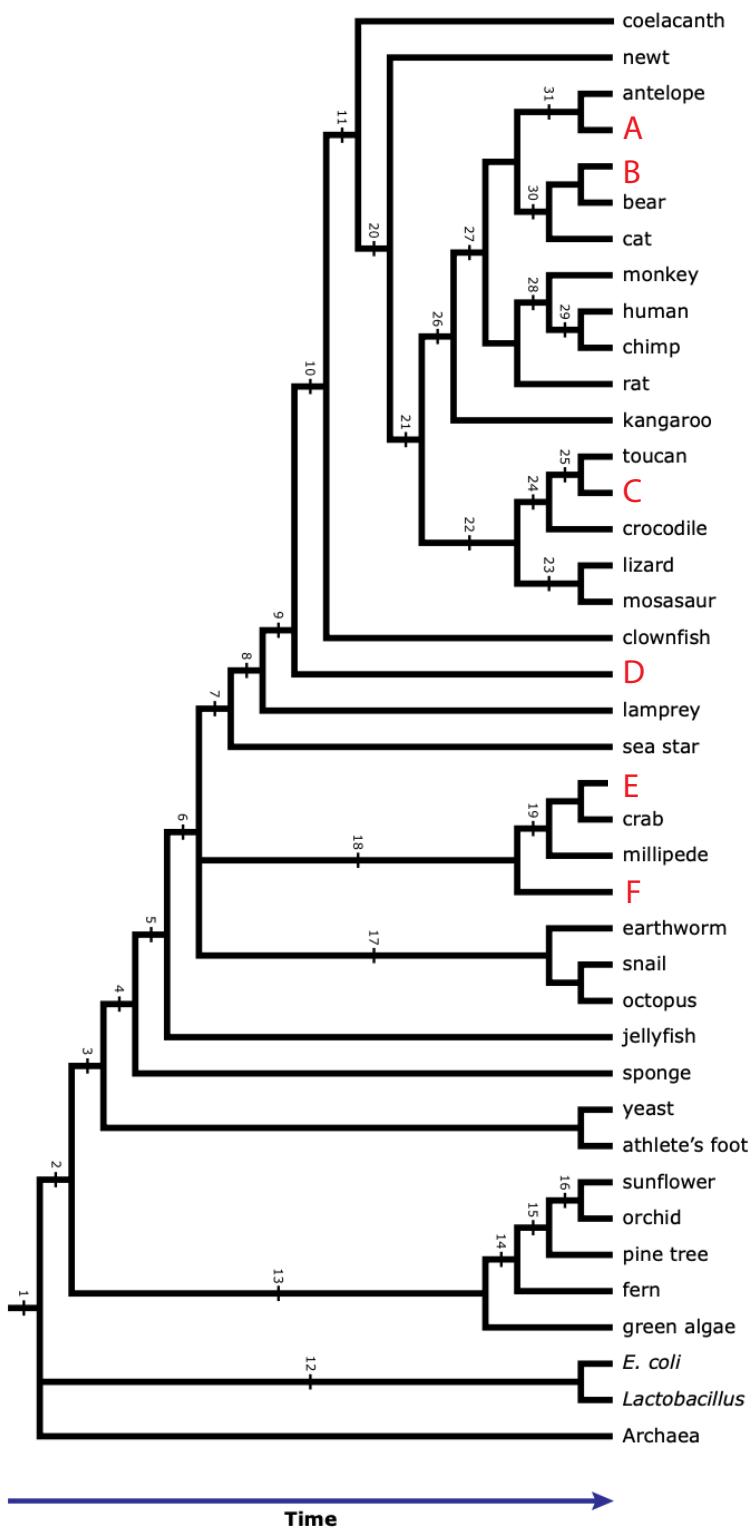
**GCATTCTGTAGC**

A horizontal arrow points from left to right, representing the progression of time. The word "past" is at the start of the arrow, and "present" is at the end. Above the arrow, the word "time" is written vertically. To the left of the arrow, the word "ancestor:" is followed by the DNA sequence "GCATTCTGTAGC". To the right of the arrow, a list of organisms and their corresponding DNA sequences are shown, starting with "jellyfish" and ending with "croc".

jellyfish	<b>GGCTTCGGCCC</b>
fly	<b>GCCTTCGTAGC</b>
cow	<b>GCATTCTGTTGC</b>
mouse	<b>GCTTTCGCTTC</b>
human	<b>GCTTTCGCTGC</b>
chicken	<b>GTATTCTGTTAC</b>
snake	<b>GTGTTCGCGAC</b>
croc	<b>GTTTTCGAGAC</b>

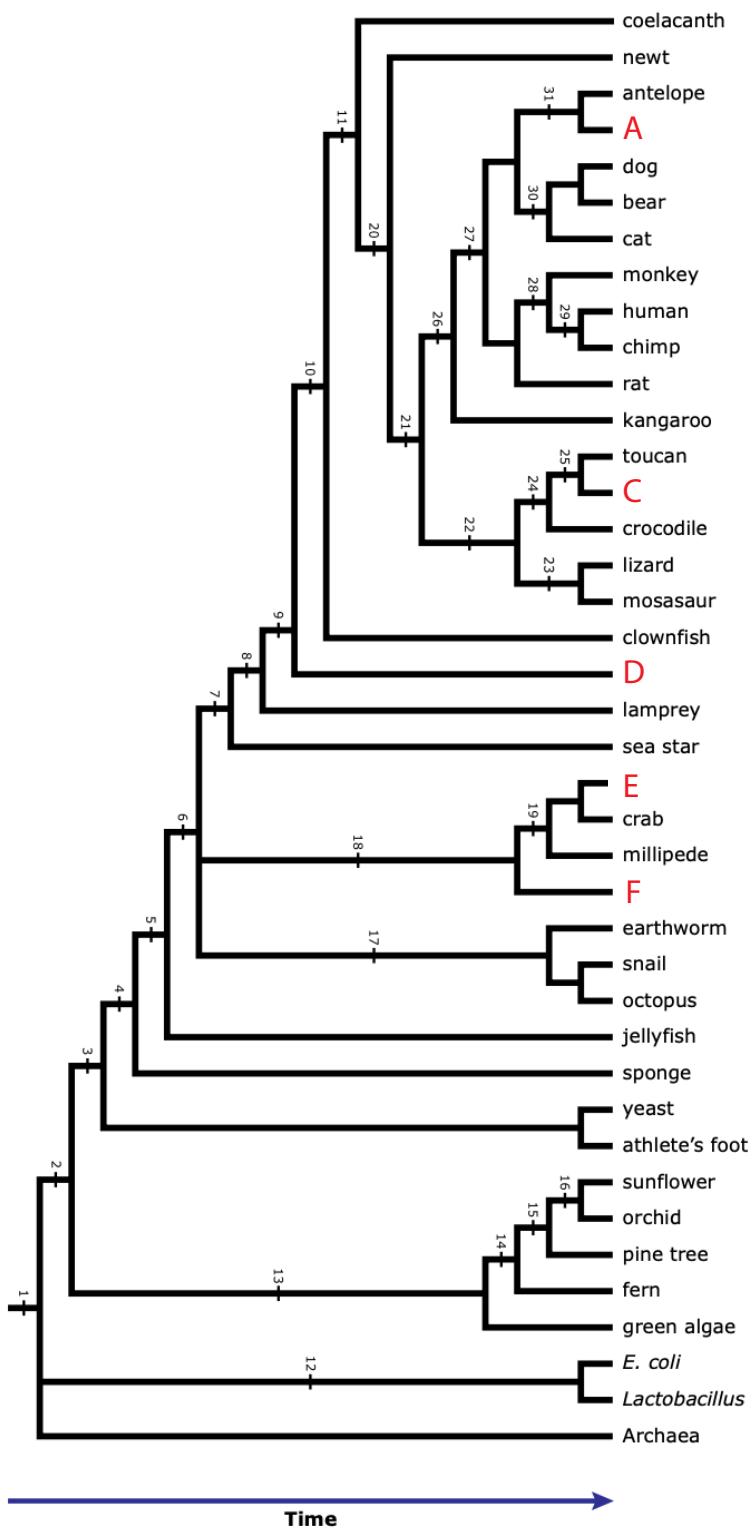






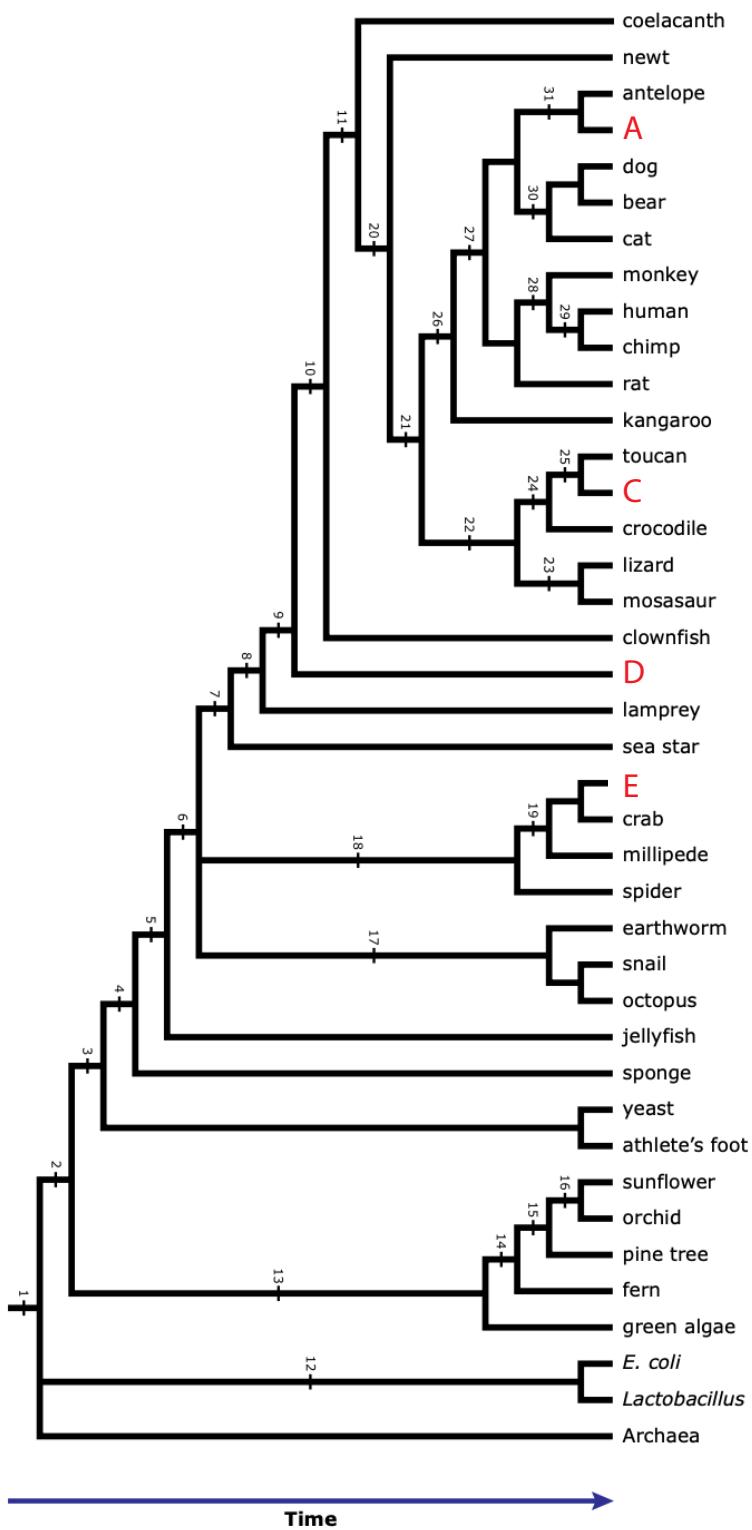
Where do the following organisms fit in the tree?

- dog?
- spider?
- T. rex?
- butterfly?
- whale?
- shark?



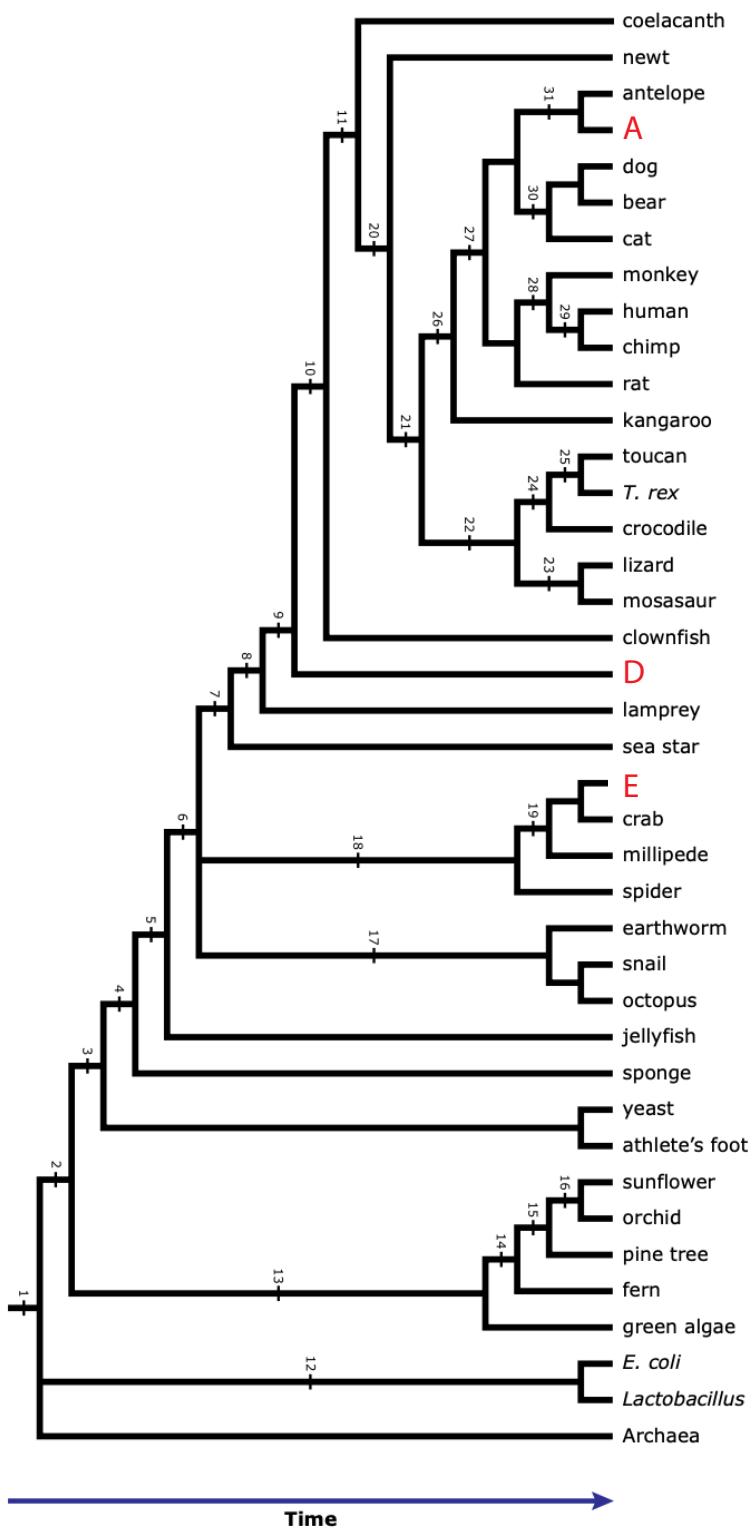
Where do the following organisms fit in the tree?

- dog?
- spider?
- T. rex?
- butterfly?
- whale?
- shark?



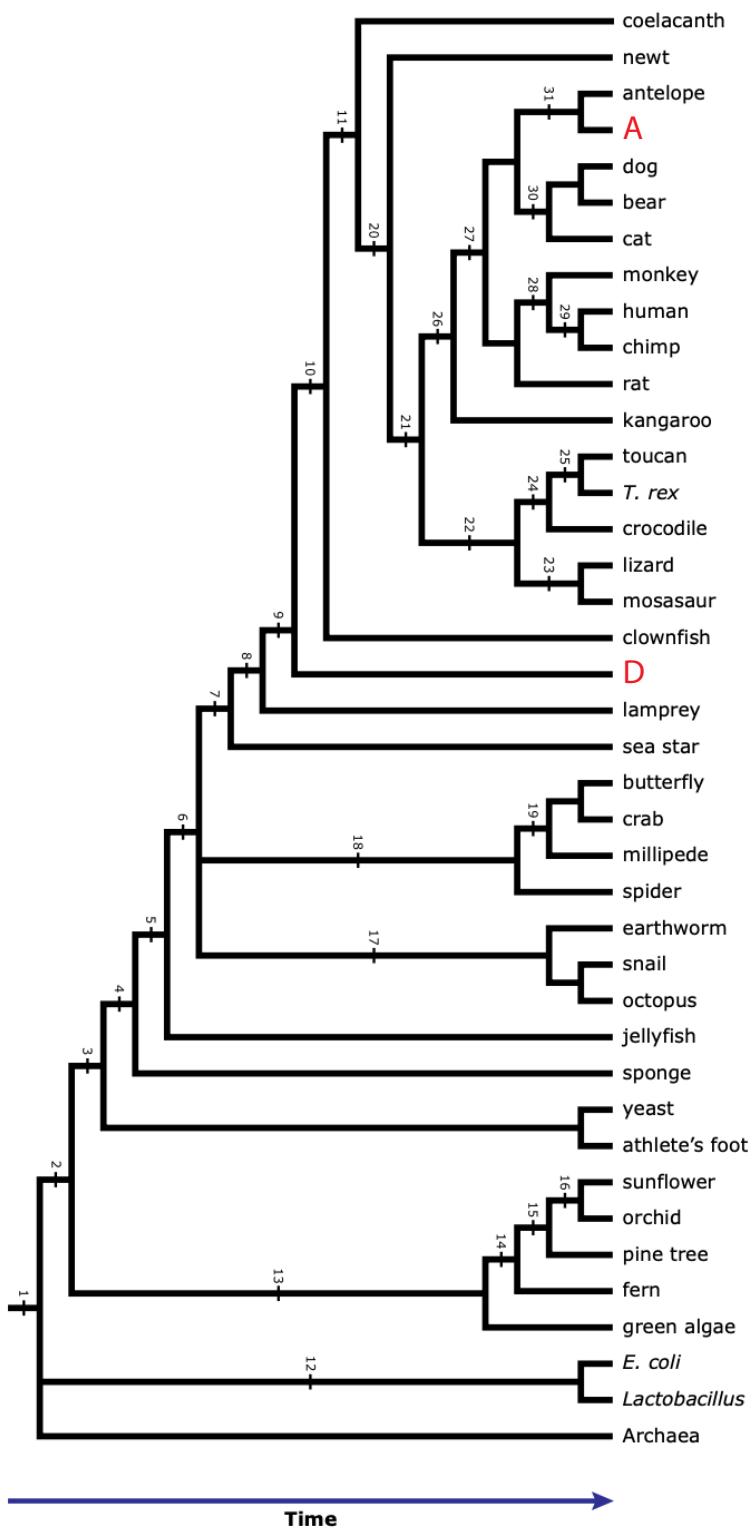
Where do the following organisms fit in the tree?

- dog?
- spider?
- T. rex?
- butterfly?
- whale?
- shark?



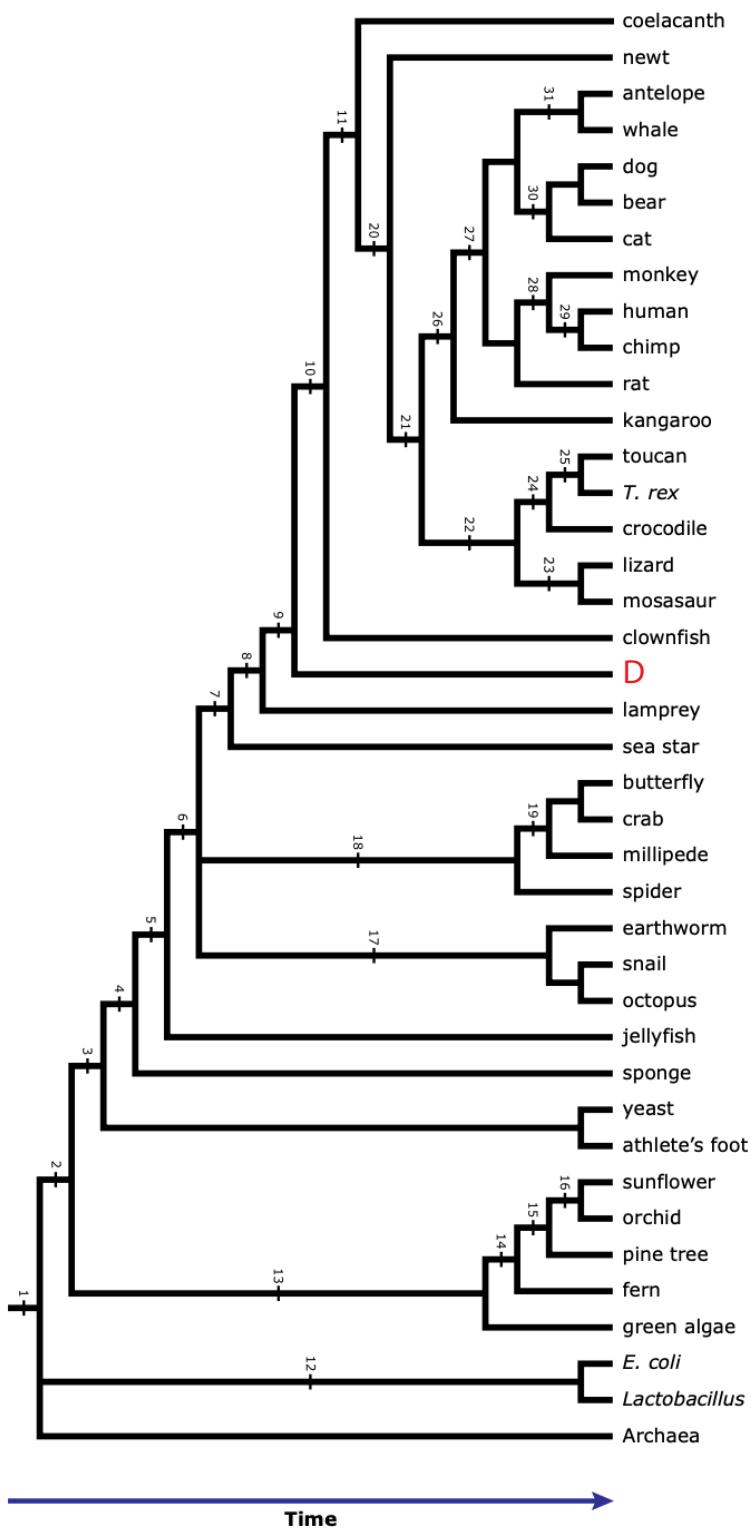
Where do the following organisms fit in the tree?

- dog?
- spider?
- *T. rex*?
- butterfly?
- whale?
- shark?



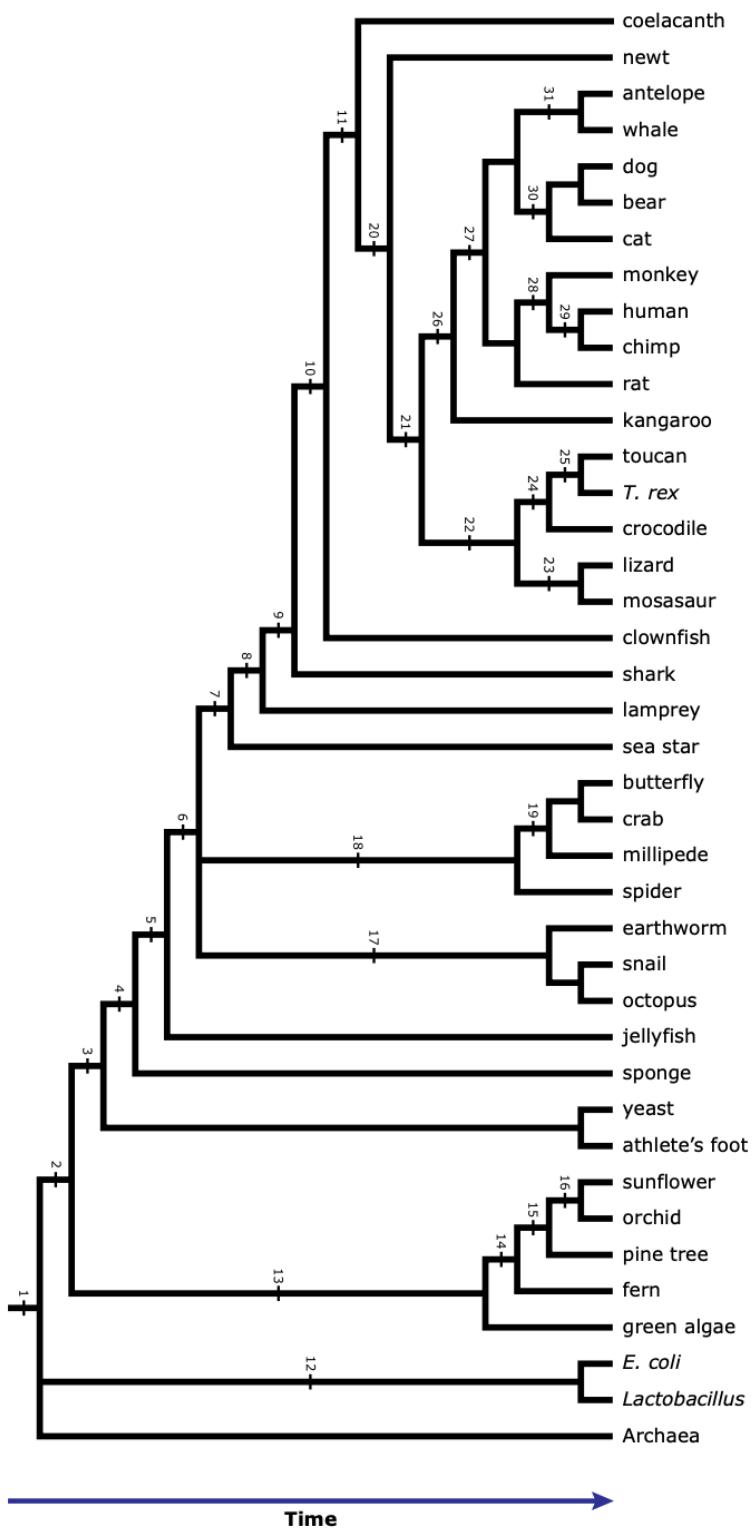
Where do the following organisms fit in the tree?

- dog?
- spider?
- *T. rex*?
- butterfly?
- whale?
- shark?



Where do the following organisms fit in the tree?

- dog?
- spider?
- *T. rex*?
- butterfly?
- whale?
- shark?



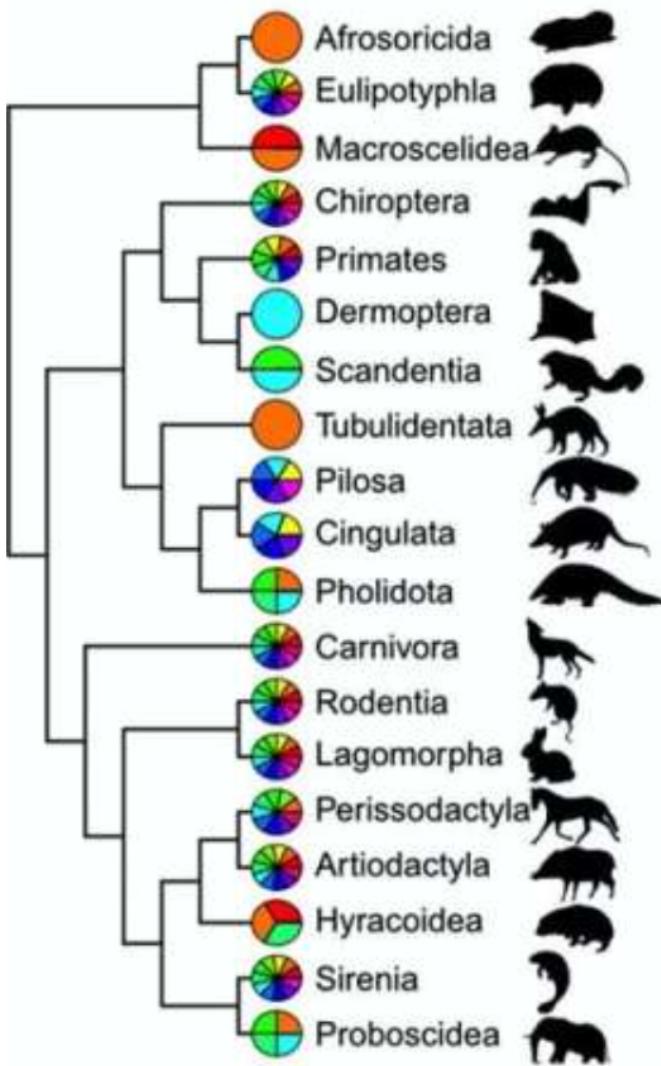
Where do the following organisms fit in the tree?

- dog?
- spider?
- T. rex?
- butterfly?
- whale?
- shark?

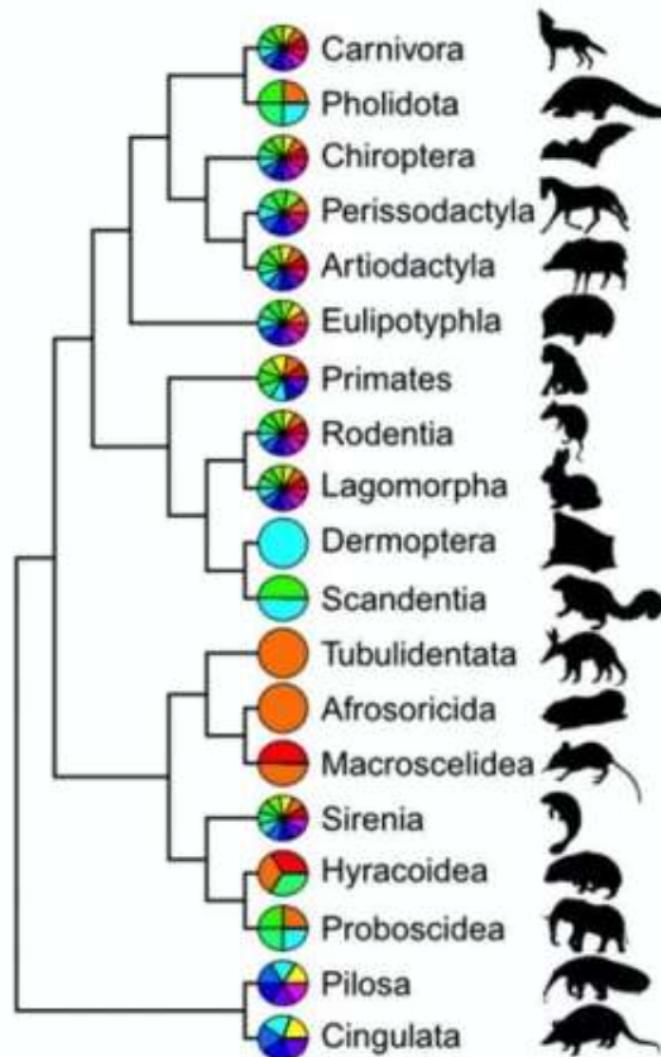
# Evolutionary tree of life: modern science is showing how we got so much wrong

by Matthew Wills, The Conversation

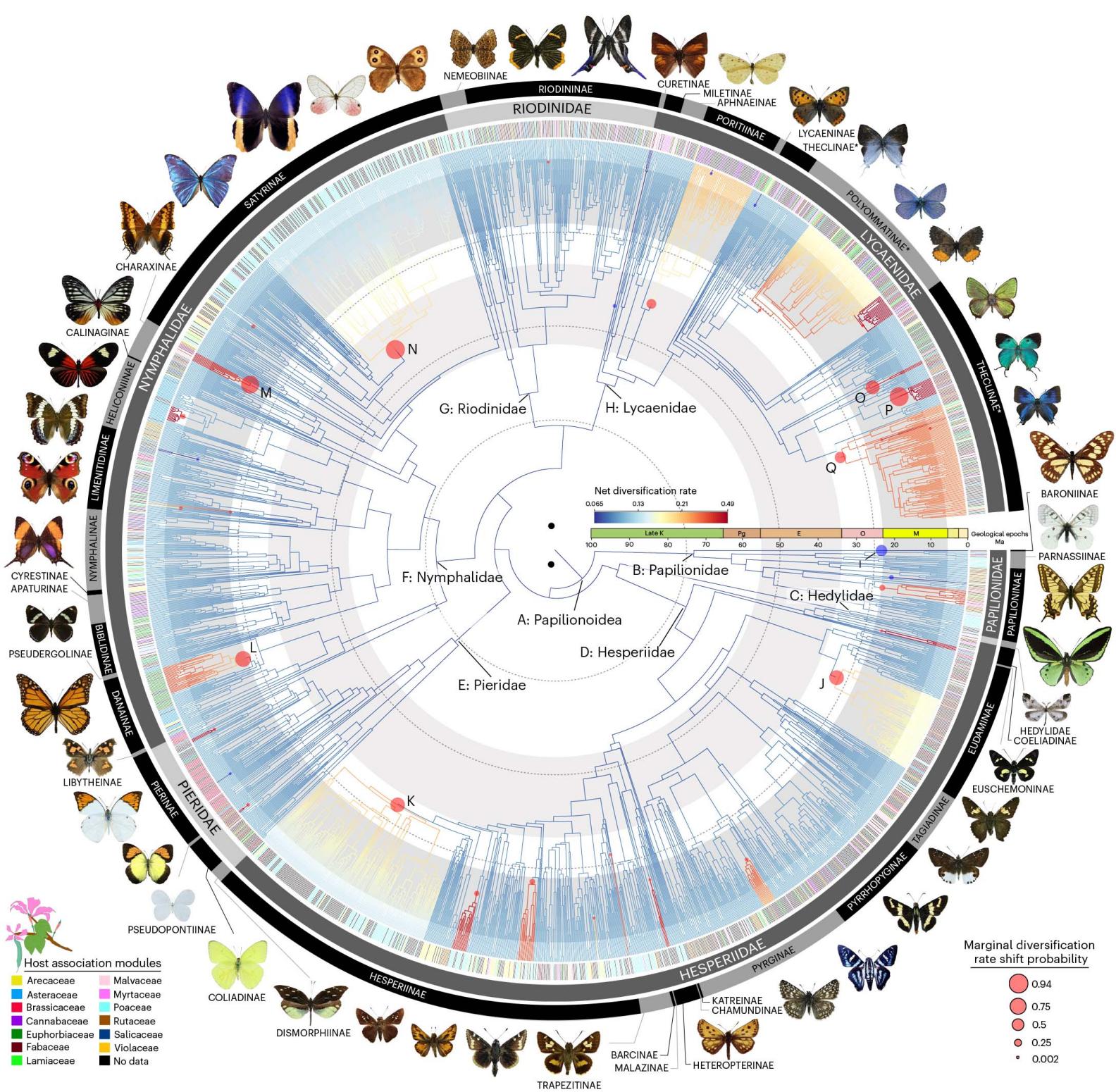
## Morphological



## Molecular



North Africa
Sub-Saharan Africa
Europe
North Asia
East Asia
West & Central Asia
South & Southeast Asia
Oceania
North America
Mesoamerica
South America
Carribean Islands



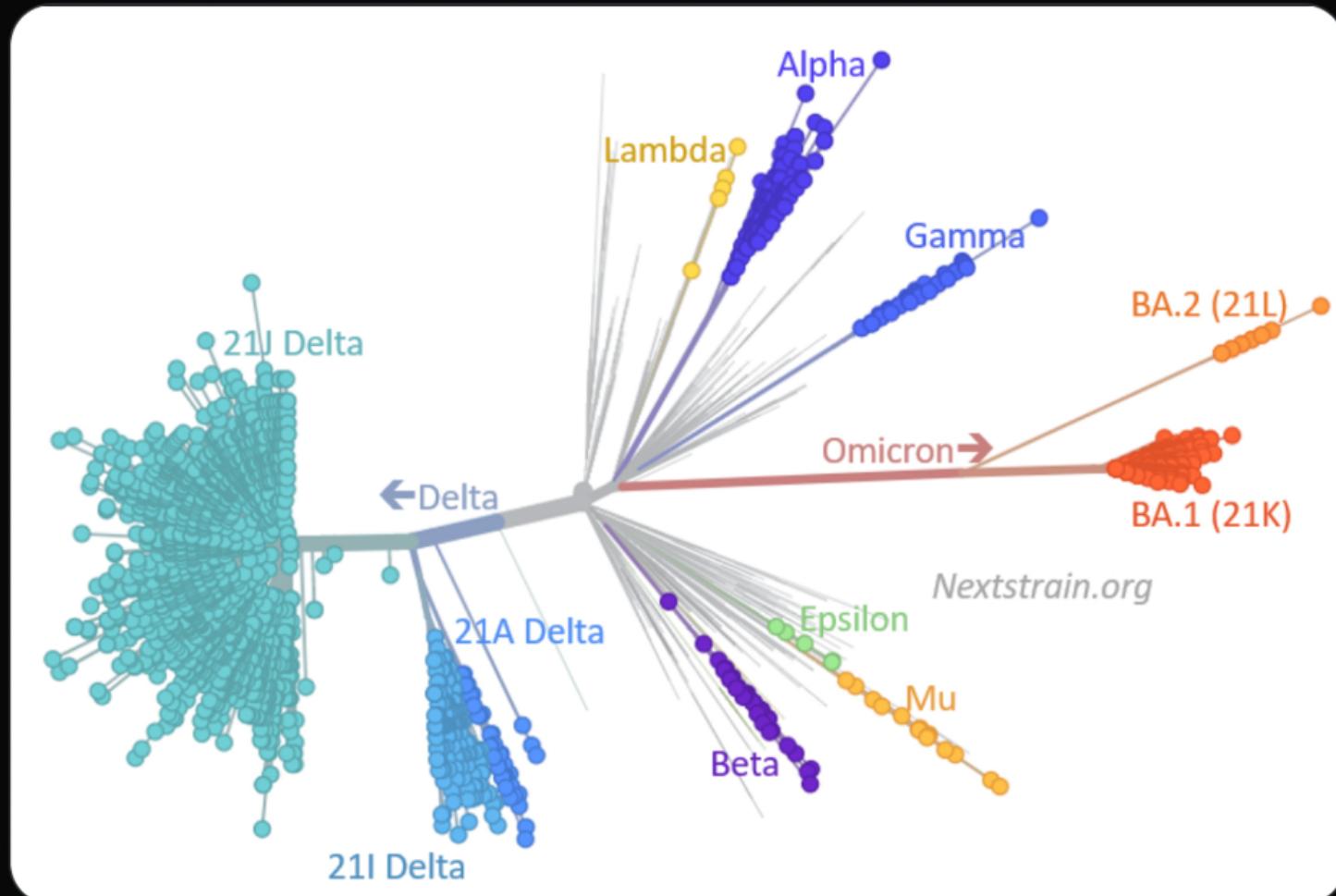


Dr Emma Hodcroft ✅ @firefox66 · Jan 27

...

This filtered @Nextstrain build gives a nice visual display of how distant the Omicron family is from everything else, & how different BA.1 (21K) & BA.2 (21L) are from each other. Distance is in mutations.

[nextstrain.org/ncov/gisaid/gl...](https://nextstrain.org/ncov/gisaid/gl...)



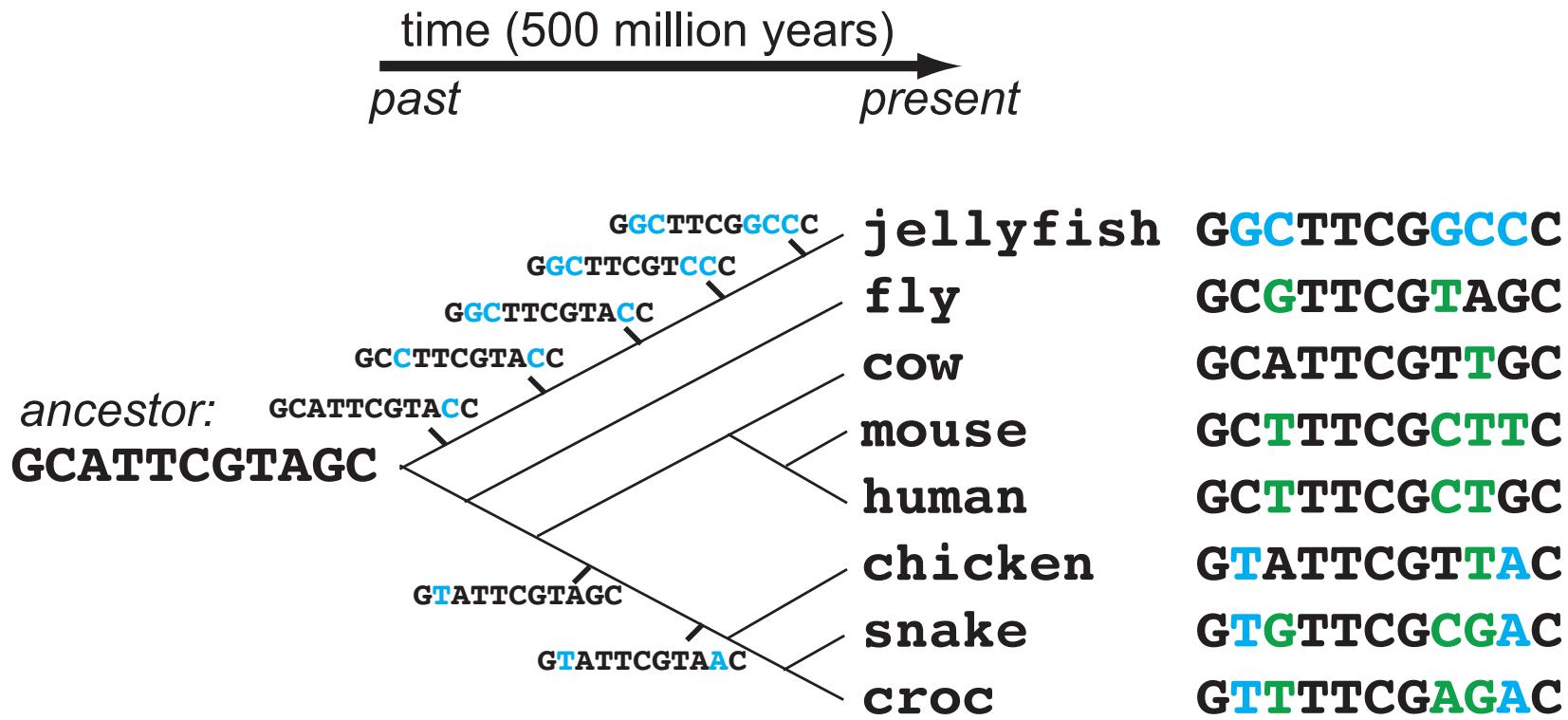
59

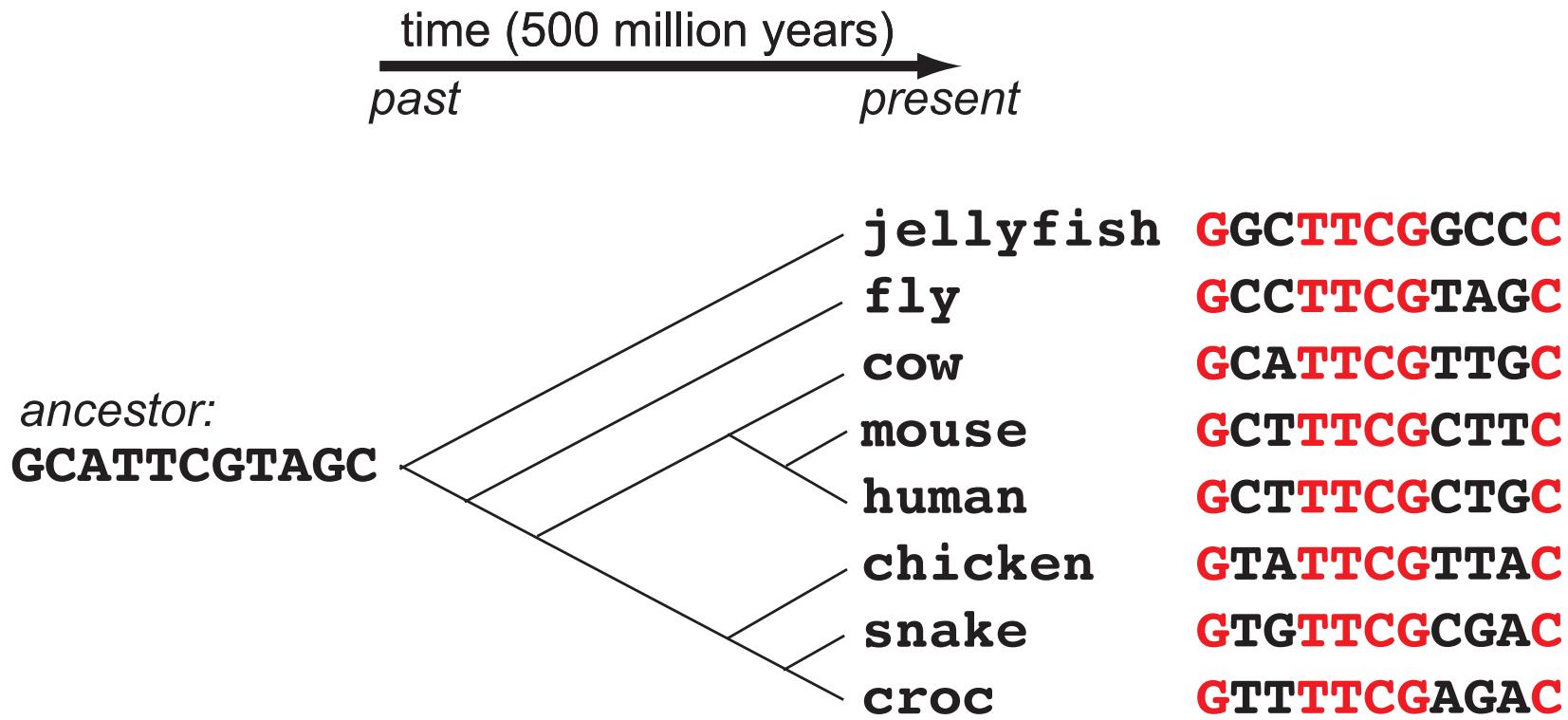
1.4K

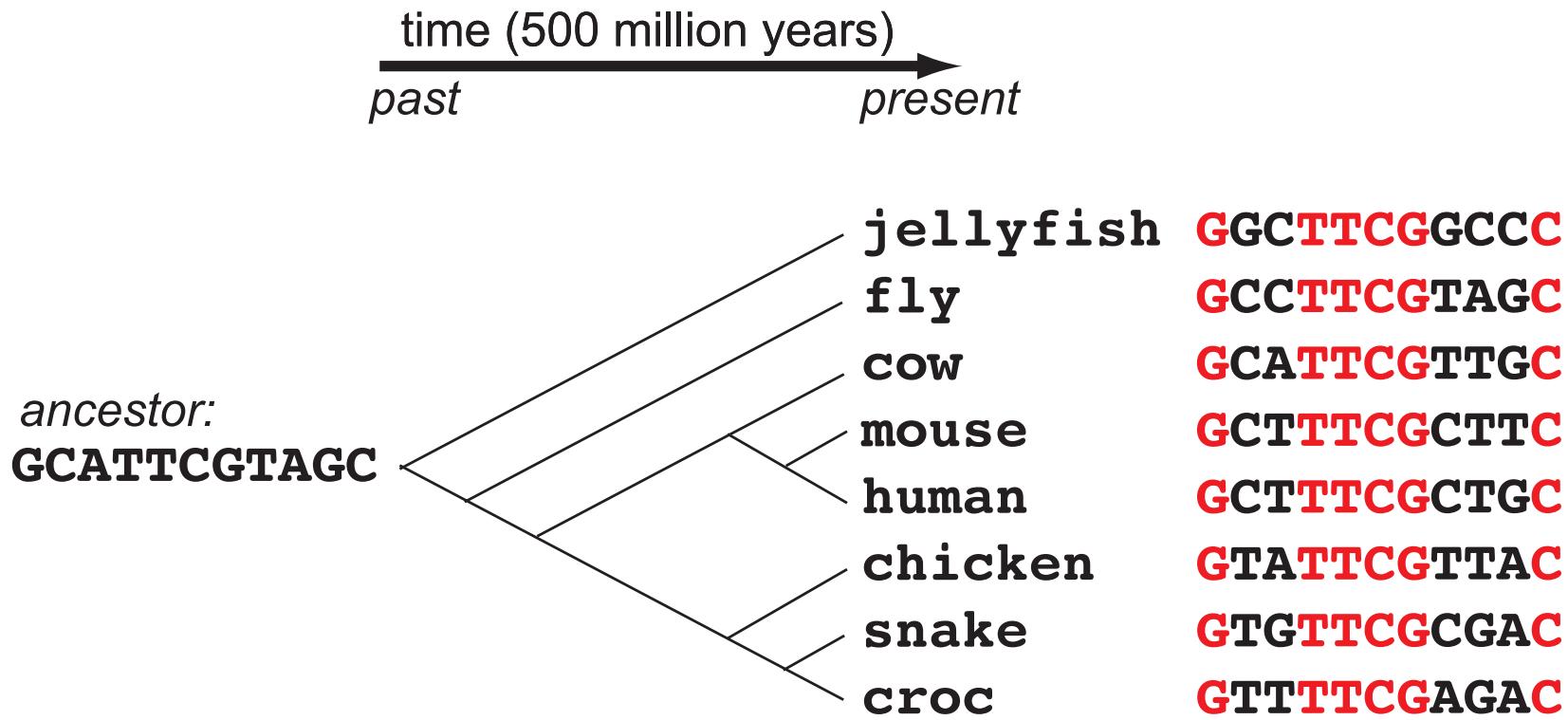
3K



1,999 × 1,763

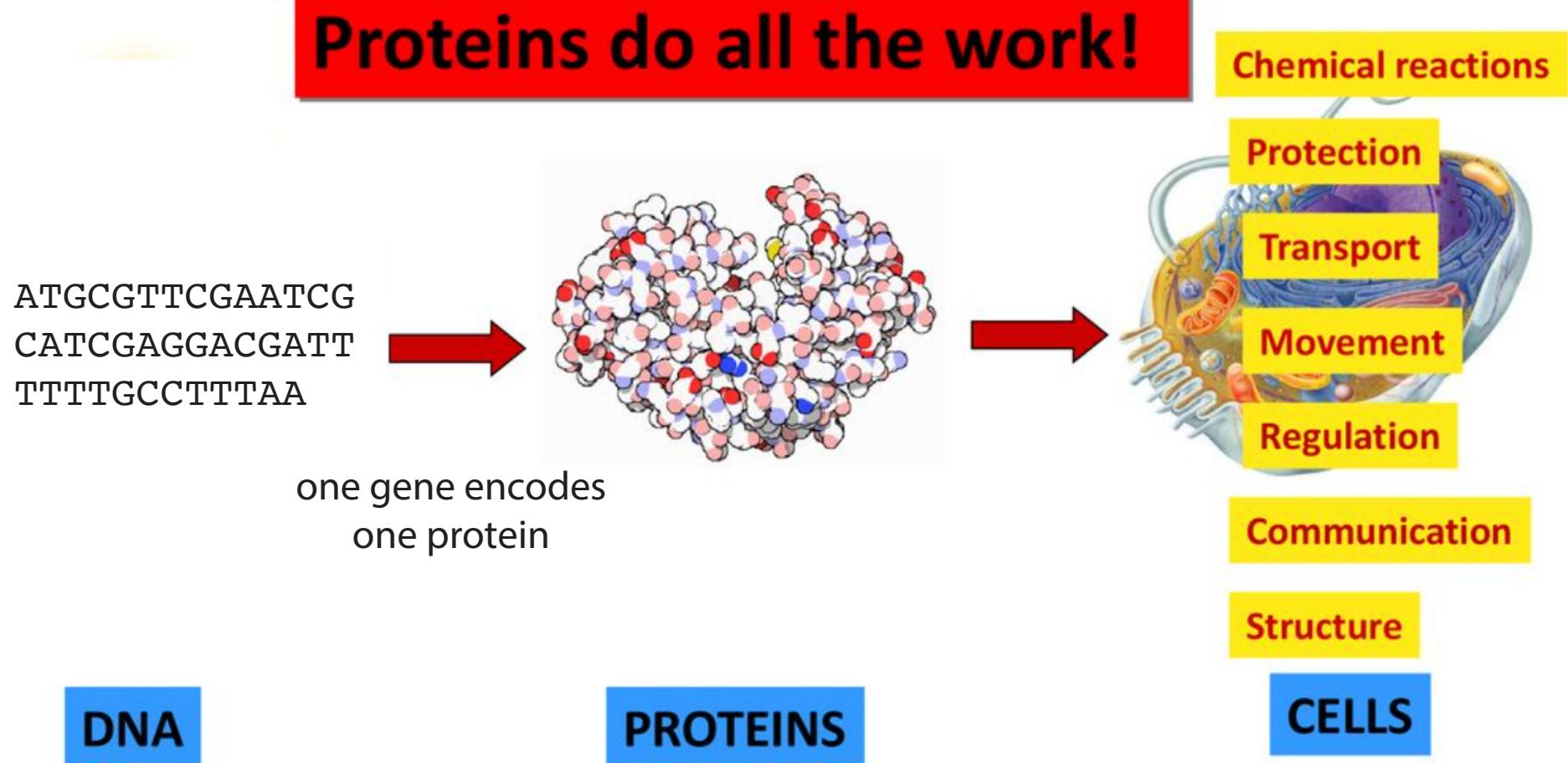






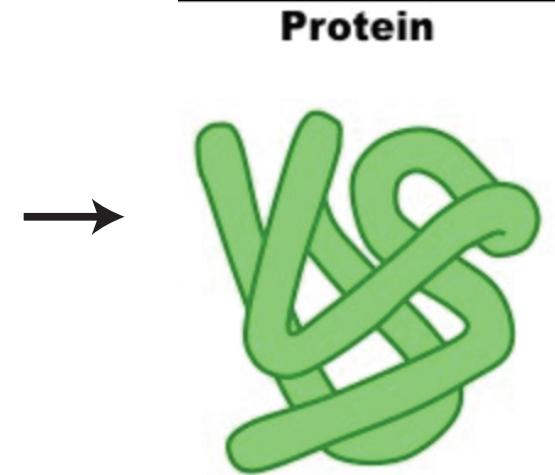
***Important sequences are conserved throughout evolution***

# Proteins are the workhorses of the cell.



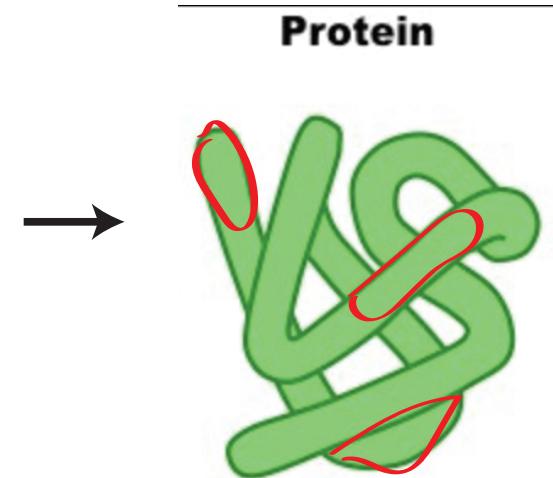
**One gene sequence contains the “instructions” for one protein**

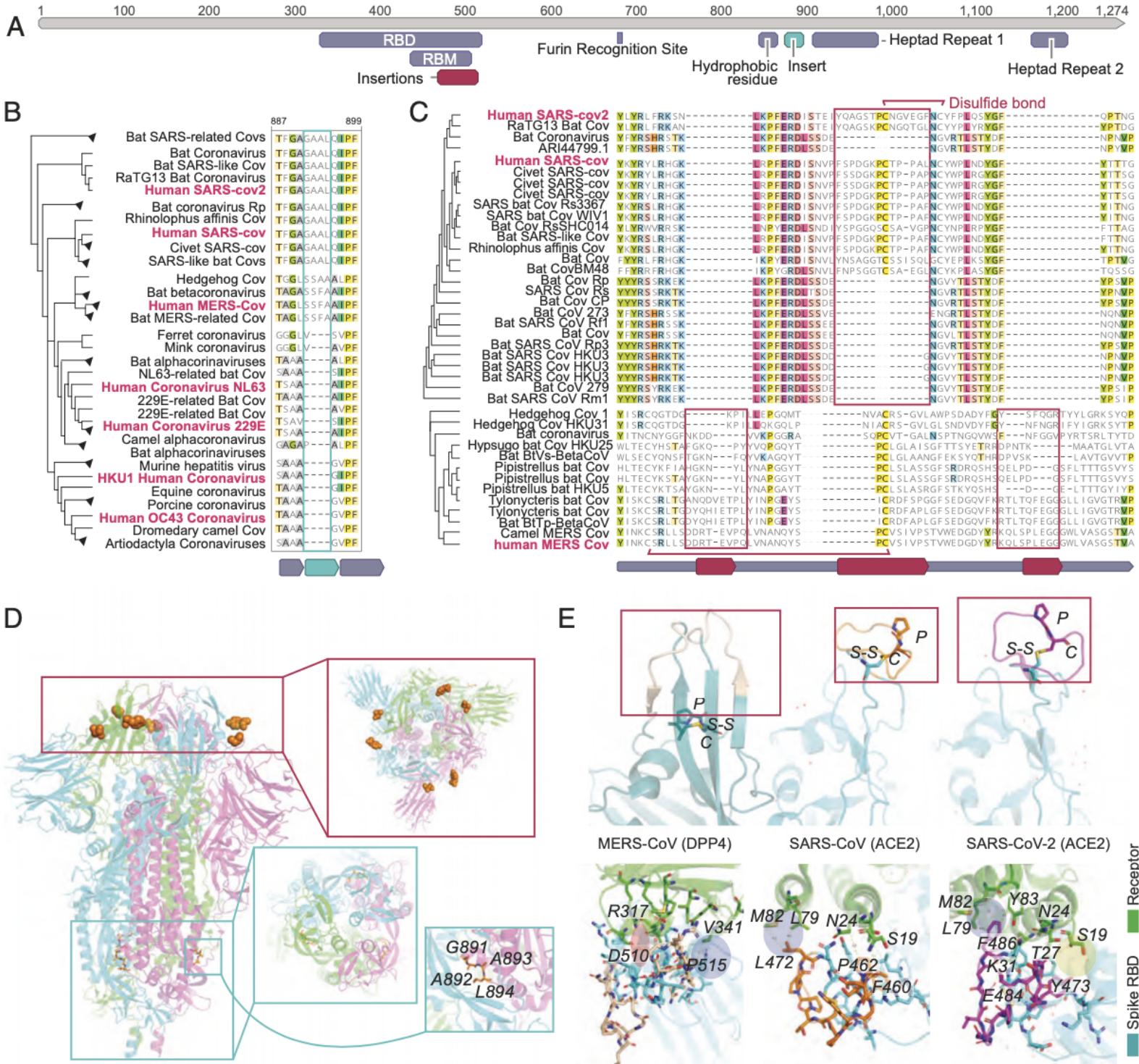
human ATGC GTT CGA ATCG C ATCG AGGG ACG ATT TTT GCCTTAA



# Knowing what parts of the sequence are conserved tells researchers what to focus on

****	*****	*****	
fly ATGCCTACGAA	TCGCAT-GGAAACCATT	TTTGTCAA	
cow ATGC GTTCGAA	TCGCATCGAGGACGAT	TTTGCCTTAA	
mouse ATGC CTCCGAA	TCGCAT-GAGGACCATT	TTTGTCTAGAG	
human ATGC GTTCGAA	TCGCATCGAGGACGAT	TTTGCCTTAA	
chicken ATGC TCTGGA-	TCGCAT-CTCTTAATT	TTTGC-TGTAA	
snake ATGC AGTC-A	TCGCATCTGGTACGAT	TTTGTCCCTTTC	





**What do I actually do all day?**

# I write code

```
File Edit Options Buffers Tools C Help
/* serial_loop():
 *
 * Read the sequence file one window of CM_MAX_RESIDUE_COUNT
 * residues (or one sequence, if seqlen < CM_MAX_RESIDUE_COUNT)
 * at a time. Search the top strand of the window, then revcomp it and
 * search the bottom strand.
 */
static int
serial_loop(WORKER_INFO *info, ESL_SQFILE *dbfp, int64_t *srcL)
{
    int      status;
    int      wstatus;
    int      prv_pli_ntophits; /* number of top hits before each cm_Pipeline() */
    int64_t  seq_idx = 0;
    ESL_SQ  *dbsq    = esl_sq_CreateDigital(info->cm->abc);

    wstatus = esl_sqio_ReadWindow(dbfp, info->pli->maxW, CM_MAX_RESIDUE_COUNT, dbsq);
    seq_idx++;

    while(wstatus == eslEOD) { /* this block is only necessary to chew up zero-length sequences */
        info->pli->nseqs++;
        esl_sq_Reuse(dbsq);
        wstatus = esl_sqio_ReadWindow(dbfp, info->pli->maxW, CM_MAX_RESIDUE_COUNT, dbsq);
        seq_idx++;
    }

    /*printf("SER just read seq %ld (%40s) %10ld..%10ld\n", seq_idx, dbsq->name, dbsq->start, dbsq->end);*/
    while (wstatus == eslOK ) {
        /* if this is the first window for this sequence, set dbsq->L */
        if(dbsq->start == 1) dbsq->L = srcL[seq_idx-1];

        cm_pli_NewSeq(info->pli, dbsq, seq_idx-1);

        if (info->pli->do_top) {
            prv_pli_ntophits = info->th->N;
            if((status = cm_Pipeline(info->pli, info->cm->offset, info->om, info->bg, info->p7_evparam, info->msvdata, dbsq, info->th, FALSE, /* FALSE: not in reverse complement */
                NULL, &(info->gm), &(info->Rgm), &(info->Lgm), &(info->Tgm), &(info->cm))) != eslOK cm_Fail("cm_pipeline() failed unexpected with status code %d\n%s\n", status, info->pli->errbuf);
            cm_pipeline_Reuse(info->pli); /* prepare for next search */

            /* subtract overlapping residues from previous window */
            if(dbsq->C > 0) cm_pli_AdjustNresForOverlaps(info->pli, dbsq->C, FALSE); /* 'FALSE': we're not on bottom strand */

            /* modify hit positions to account for the position of the window in the full sequence */
            cm_tophits_UpdateHitPositions(info->th, prv_pli_ntophits, dbsq->start, FALSE);
        }

        /* reverse complement */
        if (info->pli->do_bot && dbsq->abc->complement != NULL) {
-UU-----F1  cmsearch.c<src>  28% (796,0)  (C/*l Abbrev) -----
```

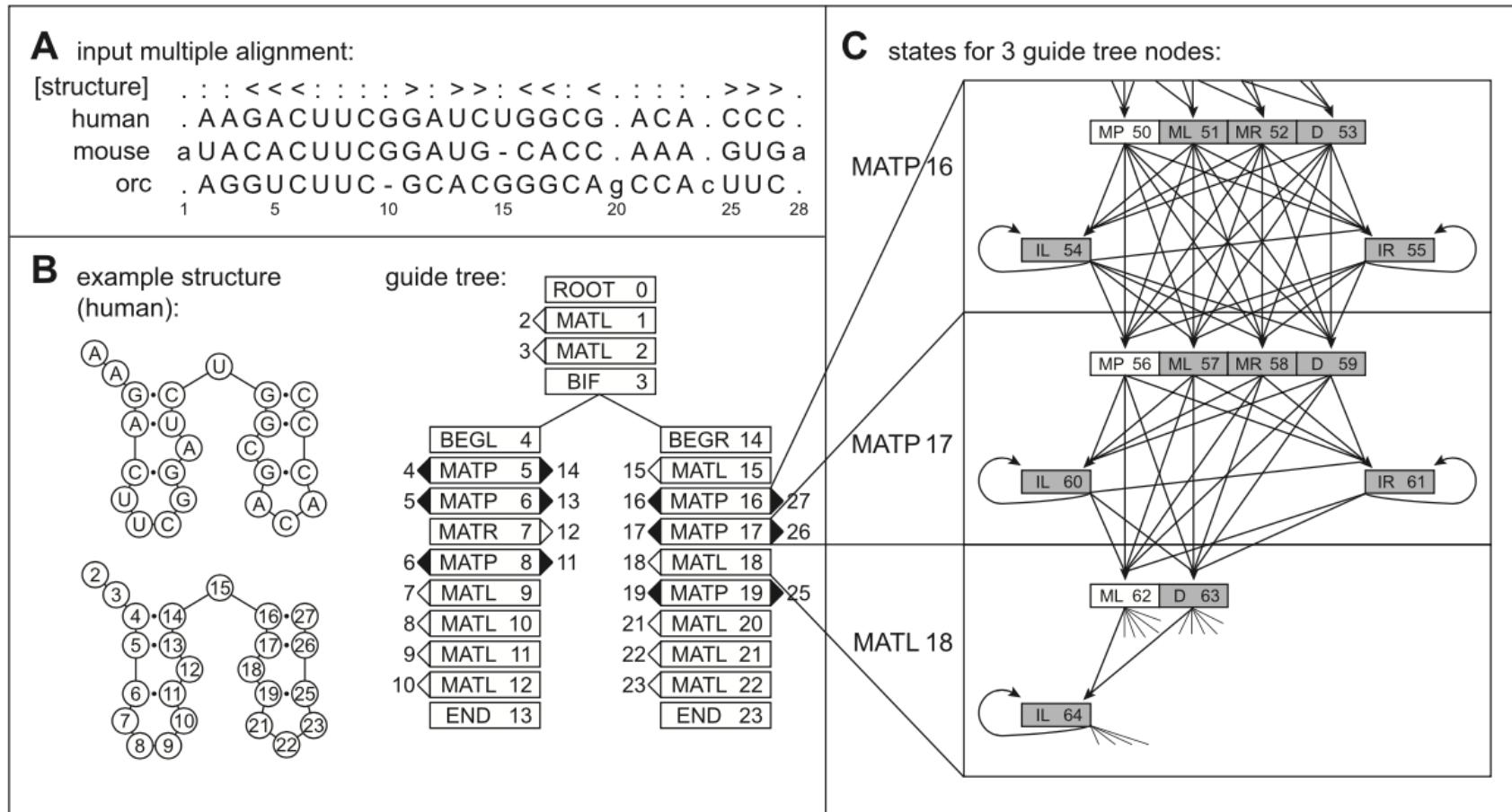
and develop and publish software

# Query-Dependent Banding (QDB) for Faster RNA Similarity Searches

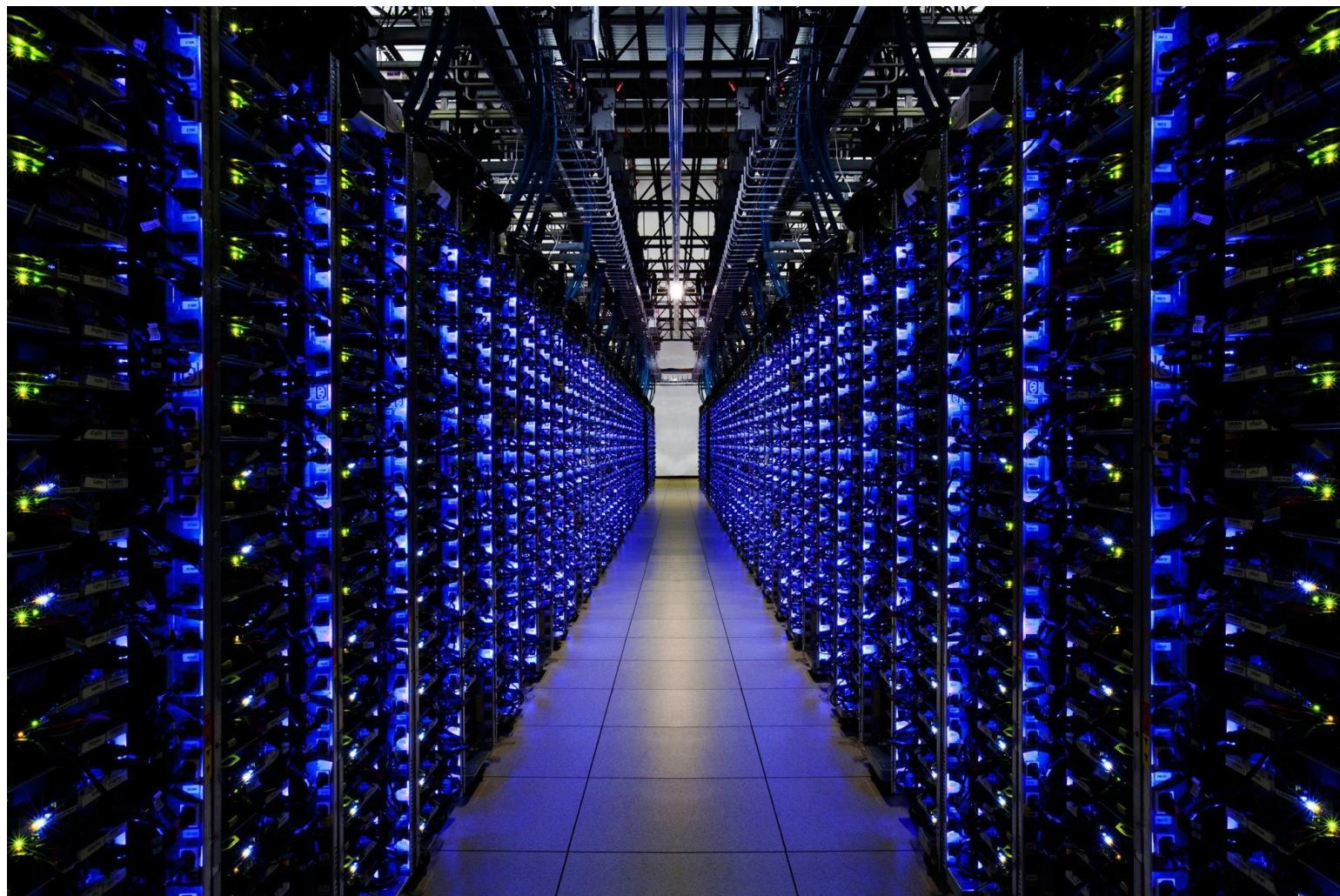


Eric P. Nawrocki, Sean R. Eddy\*

Howard Hughes Medical Institute, Janelia Farm Research Campus, Ashburn, Virginia, United States of America



**Figure 1.** An Example RNA Family and Corresponding CM



## Education

- College - University of Maryland
  - biology
  - computer science
- Graduate school - Washington University in St. Louis
  - Ph.D. in computational biology

# Biology careers

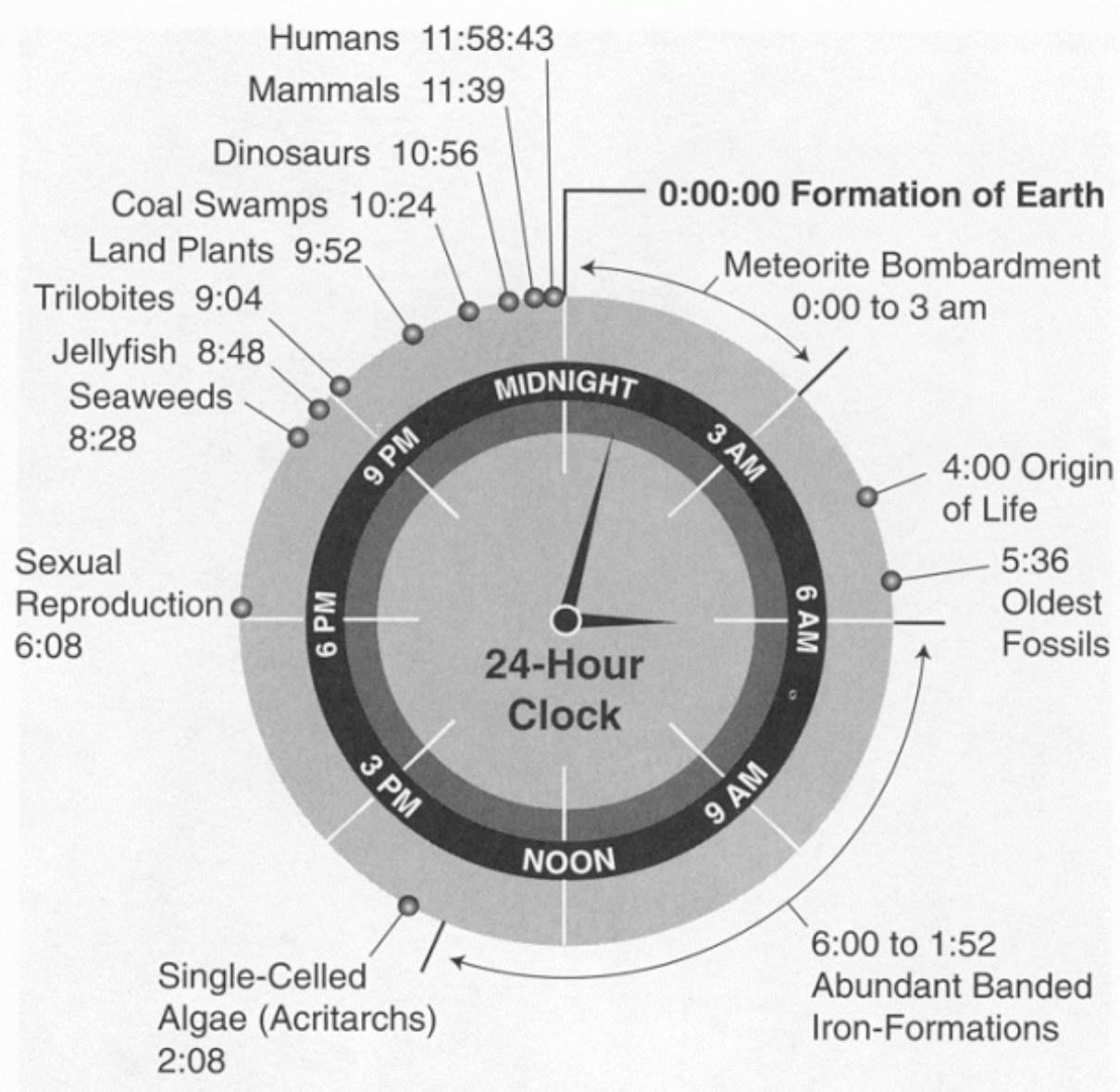
- Research
  - government (NIH)
  - academia - professor or research assistant
- Industry
  - pharmaceutical company
  - biotech company
- Medicine
  - doctor
  - nurse
  - veterinarian

## Science research job perks

- get paid to go to graduate school (at least for biology)
- very flexible; you can be your own boss
- get to follow your own interests
- conferences and collaborations are international
  - Spain (6 times)
  - England (5 times)
  - Sweden
  - Denmark

THANK Y<sup>O</sup>U





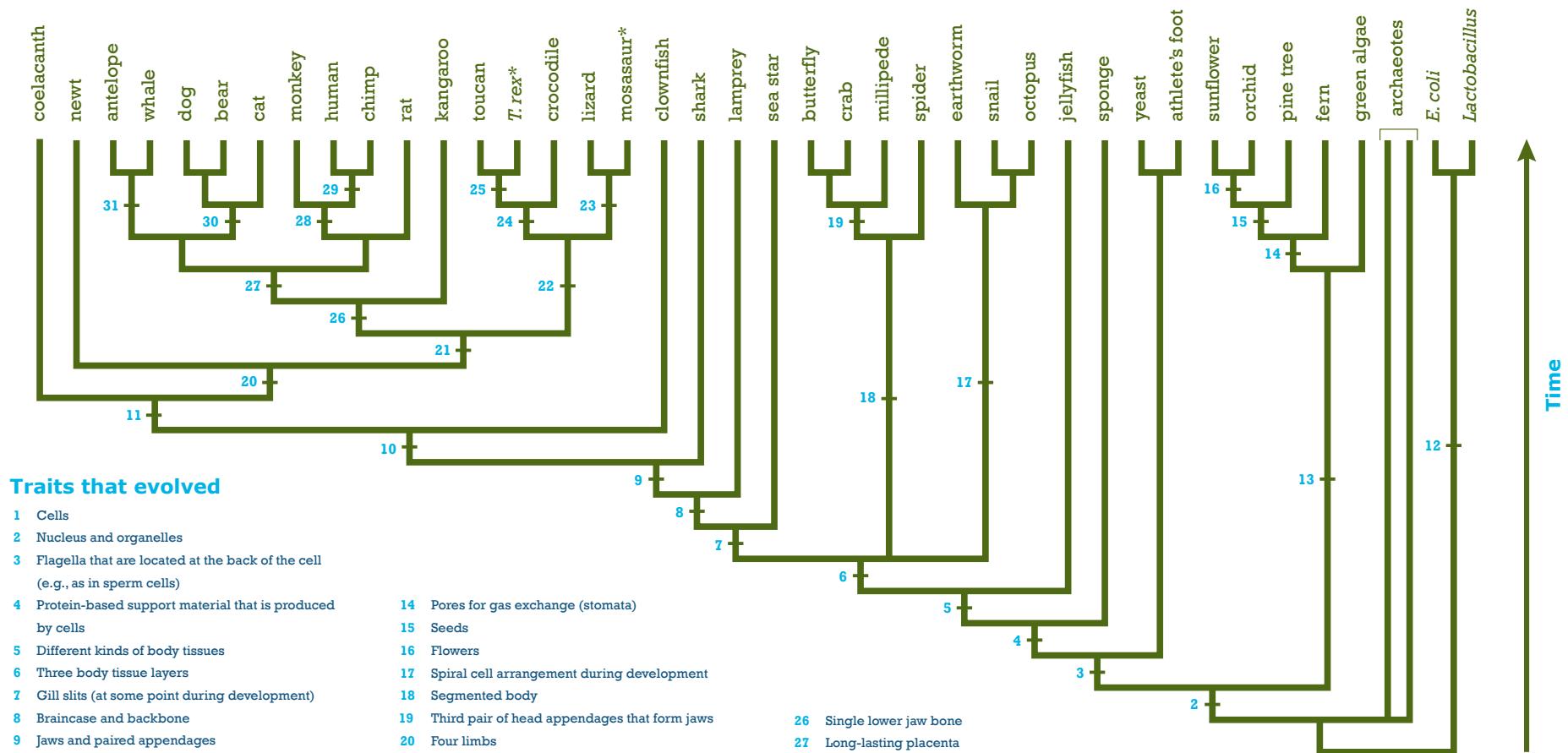


# The Tree of Life

Find your favorite organisms and follow the tree to find their common ancestor.

Visit [The Tree Room](http://www.treeroom.org) at [www.treeroom.org](http://www.treeroom.org)  
to learn more about evolutionary trees.

**UCMP** Understanding Evolution — <http://evolution.berkeley.edu>  
© 2015 The University of California Museum of Paleontology, Berkeley,  
and the Regents of the University of California

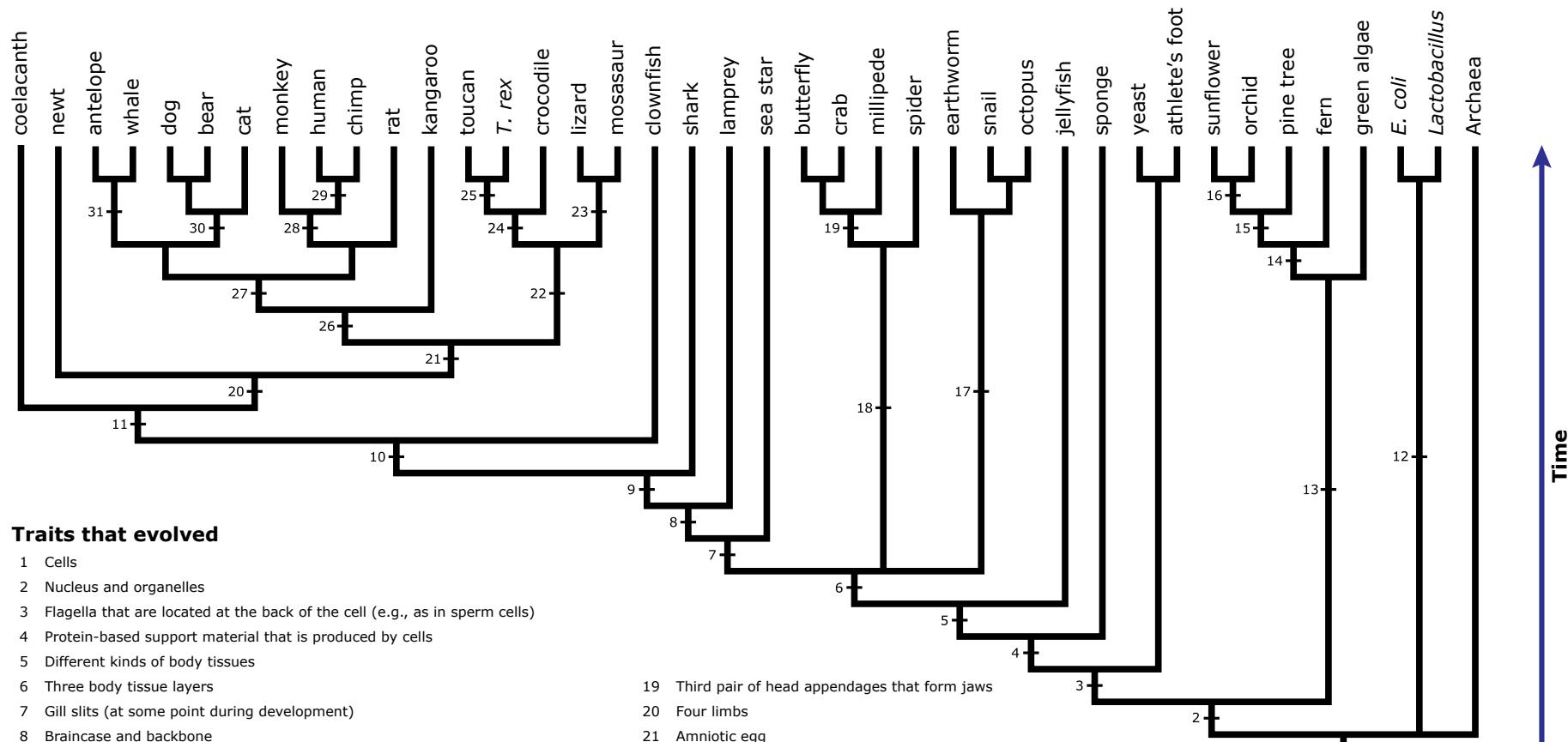


\*extinct

Modified from The Wellcome Tree of Life  
<http://www.wellcometreeoflife.org/interactive>



# The Tree of Life



## Traits that evolved

- 1 Cells
- 2 Nucleus and organelles
- 3 Flagella that are located at the back of the cell (e.g., as in sperm cells)
- 4 Protein-based support material that is produced by cells
- 5 Different kinds of body tissues
- 6 Three body tissue layers
- 7 Gill slits (at some point during development)
- 8 Braincase and backbone
- 9 Jaws and paired appendages
- 10 Bony skeleton
- 11 Appendage composed of a single bone (humerus) that articulates with the shoulder
- 12 Hollow flagella on cells
- 13 Cell wall made of cellulose
- 14 Pores for gas exchange (stomata)
- 15 Seeds
- 16 Flowers
- 17 Spiral cell arrangement during development
- 18 Segmented body
- 19 Third pair of head appendages that form jaws
- 20 Four limbs
- 21 Amniotic egg
- 22 Beta-keratin in the skin
- 23 Flexible bones at back of skull (cranial kinesis)
- 24 Bump on the rear of the ankle bone (calcaneus)
- 25 Wishbone or furcula (fused clavicle bones)
- 26 Single lower jaw bone
- 27 Long-lasting placenta
- 28 Fingernails instead of claws
- 29 Broad, shallow thorax
- 30 Specialized slicing teeth (large upper fourth premolar and lower first molar)
- 31 Double pulley ankle bone (astragalus)