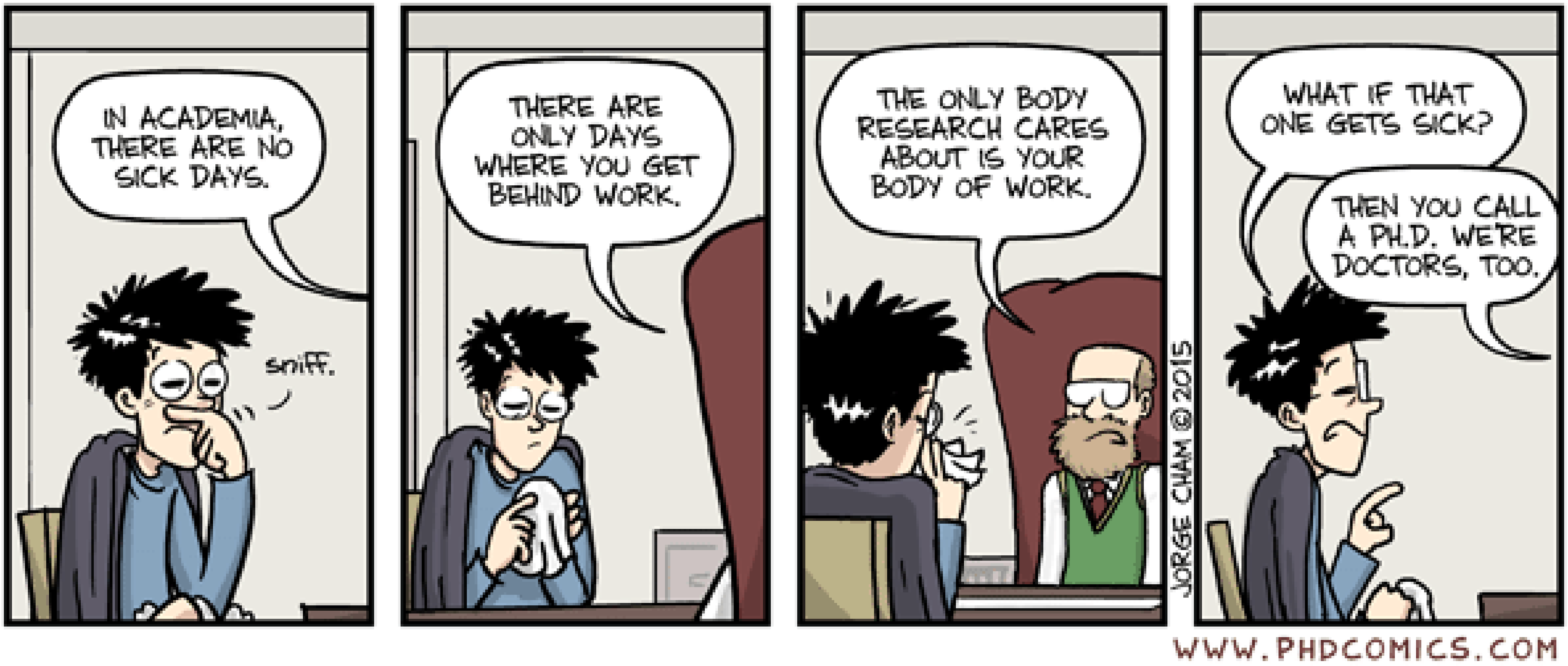
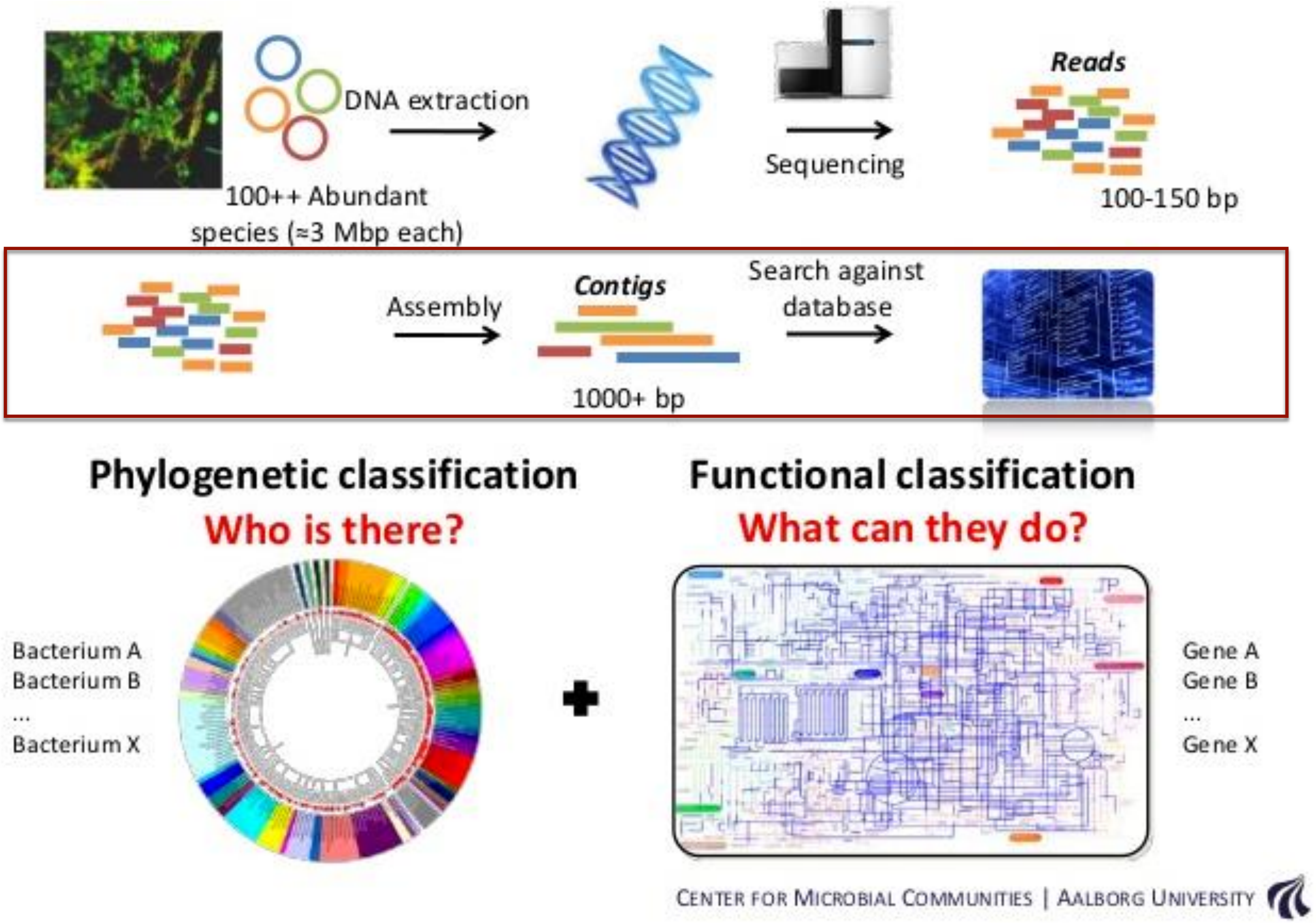
Assignment 8: Metagenomics


IN ACADEMIA, THERE ARE NO SICK DAYS.
sniff.
THERE ARE ONLY DAYS WHERE YOU GET BEHIND WORK.
THE ONLY BODY RESEARCH CARES ABOUT IS YOUR BODY OF WORK.
JORGE CHAM © 2015
WHAT IF THAT ONE GETS SICK?
THEN YOU CALL A PH.D. WE'RE DOCTORS, TOO.
WWW.PHDCOMICS.COM

Bio5488
Spring '16

Adapted from: http://phdcomics.com/comics/archive/phd031315s.gif

Extracting genomes from metagenomes



DNA extraction

100++ Abundant species (≈3 Mbp each)

Sequencing

Reads

100-150 bp

Assembly

Contigs

1000+ bp

Search against database

Phylogenetic classification

Who is there?

Bacterium A
Bacterium B
...
Bacterium X

Functional classification

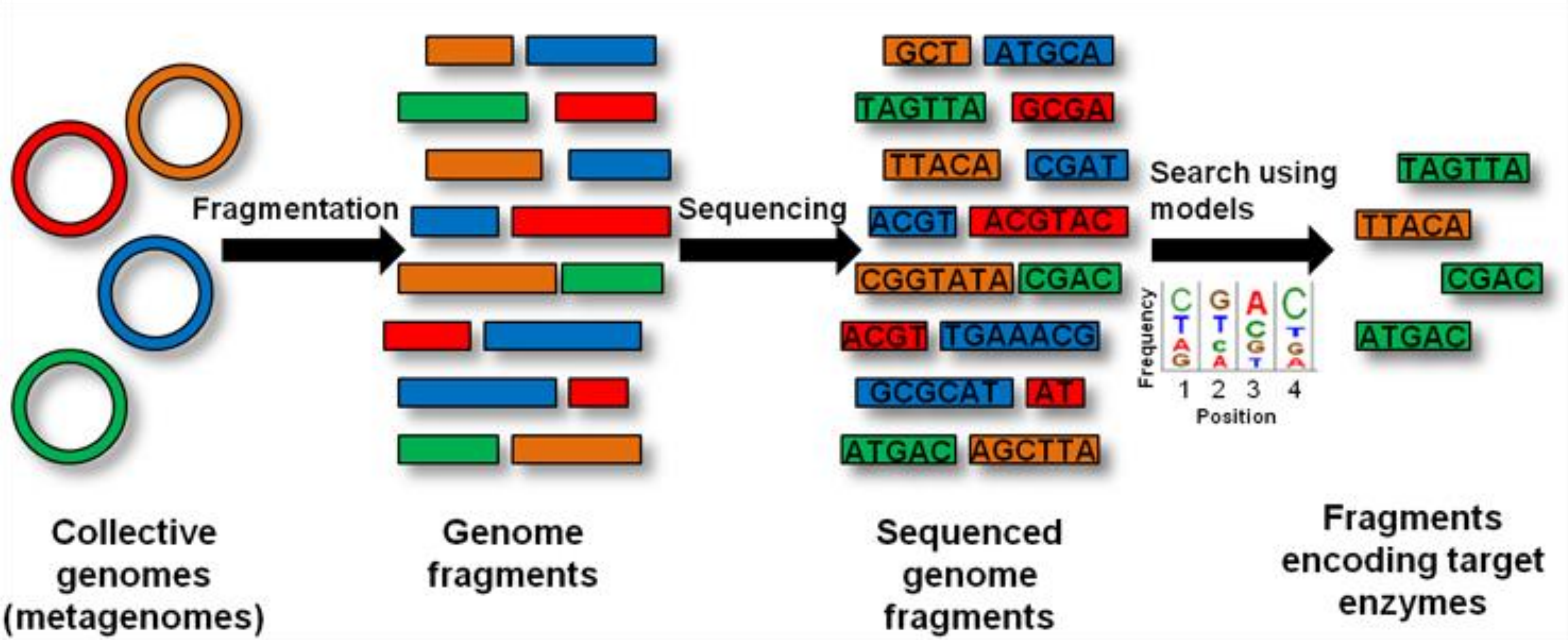
What can they do?

Gene A
Gene B
...
Gene X

CENTER FOR MICROBIAL COMMUNITIES | AALBORG UNIVERSITY

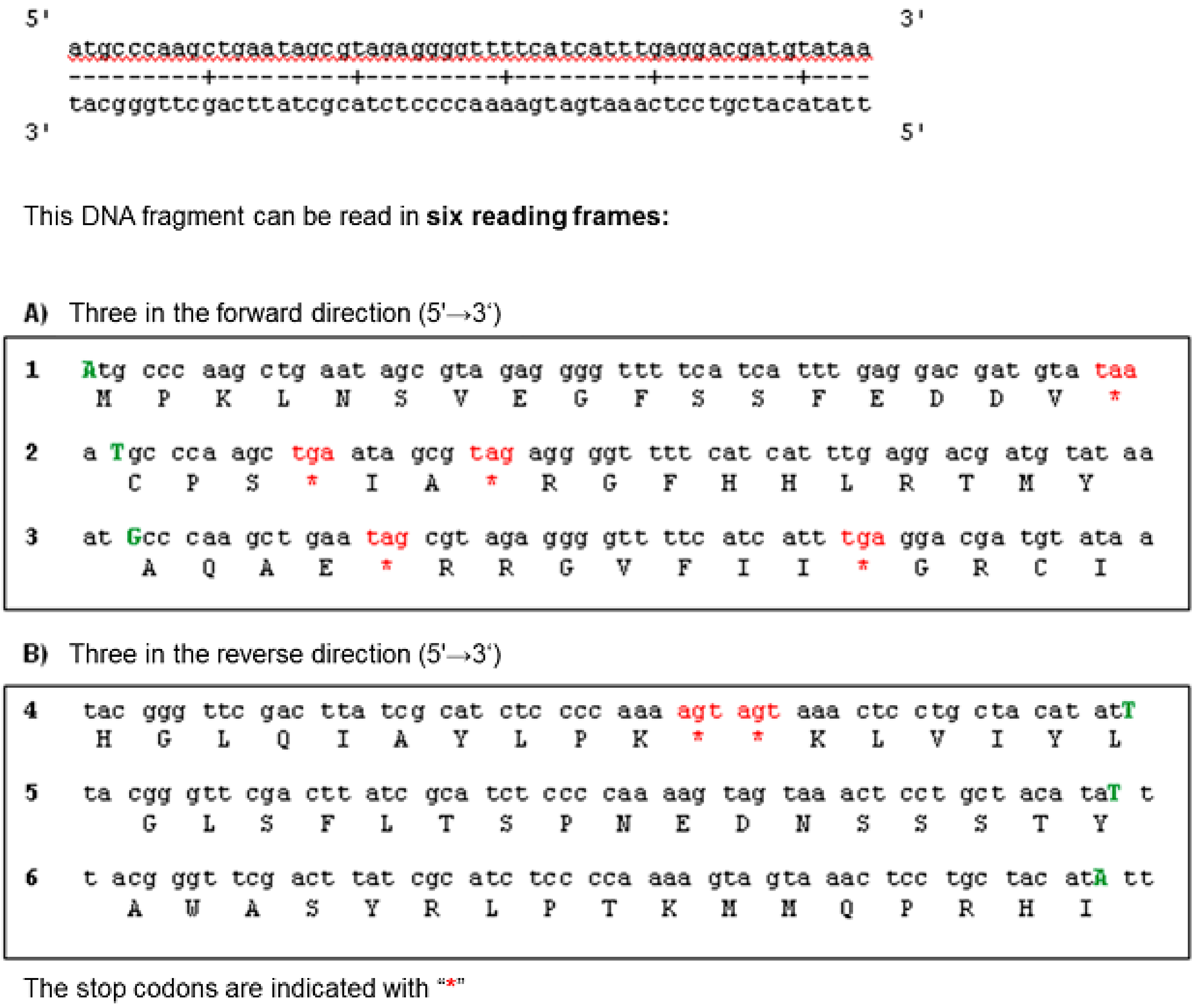
Adapted from: Slides of Mads Albertsen's talk, "Extracting genomes from metagenomes".

Metagenomics process



Adapted from: http://www.cebitec.uni-bielefeld.de/clib-gc/localmedia/short_desc/large/martha_zakrzewski.png

Six open reading frames (ORFs)



```
5'                                                              3'
   atgcccaagctgaatagcgtagaggggtttttcatcatttgaggacgatgtataa
   ---------+---------+---------+---------+---------+----
   tacgggttcgacttatcgcatctccccaaaagtagtaaactcctgctacatatt
3'                                                              5'
```

This DNA fragment can be read in **six reading frames:**

A) Three in the forward direction (5'→3')

```
1   Atg ccc aag ctg aat agc gta gag ggg ttt tca tca ttt gag gac gat gta taa
     M   P   K   L   N   S   V   E   G   F   S   S   F   E   D   D   V   *

2   a Tgc cca agc tga ata gcg tag agg ggt ttt cat cat ttg agg acg atg tat aa
       C   P   S   *   I   A   *   R   G   F   H   H   L   R   T   M   Y

3   at Gcc caa gct gaa tag cgt aga ggg gtt ttc atc att tga gga cga tgt ata a
        A   Q   A   E   *   R   R   G   V   F   I   I   *   G   R   C   I
```

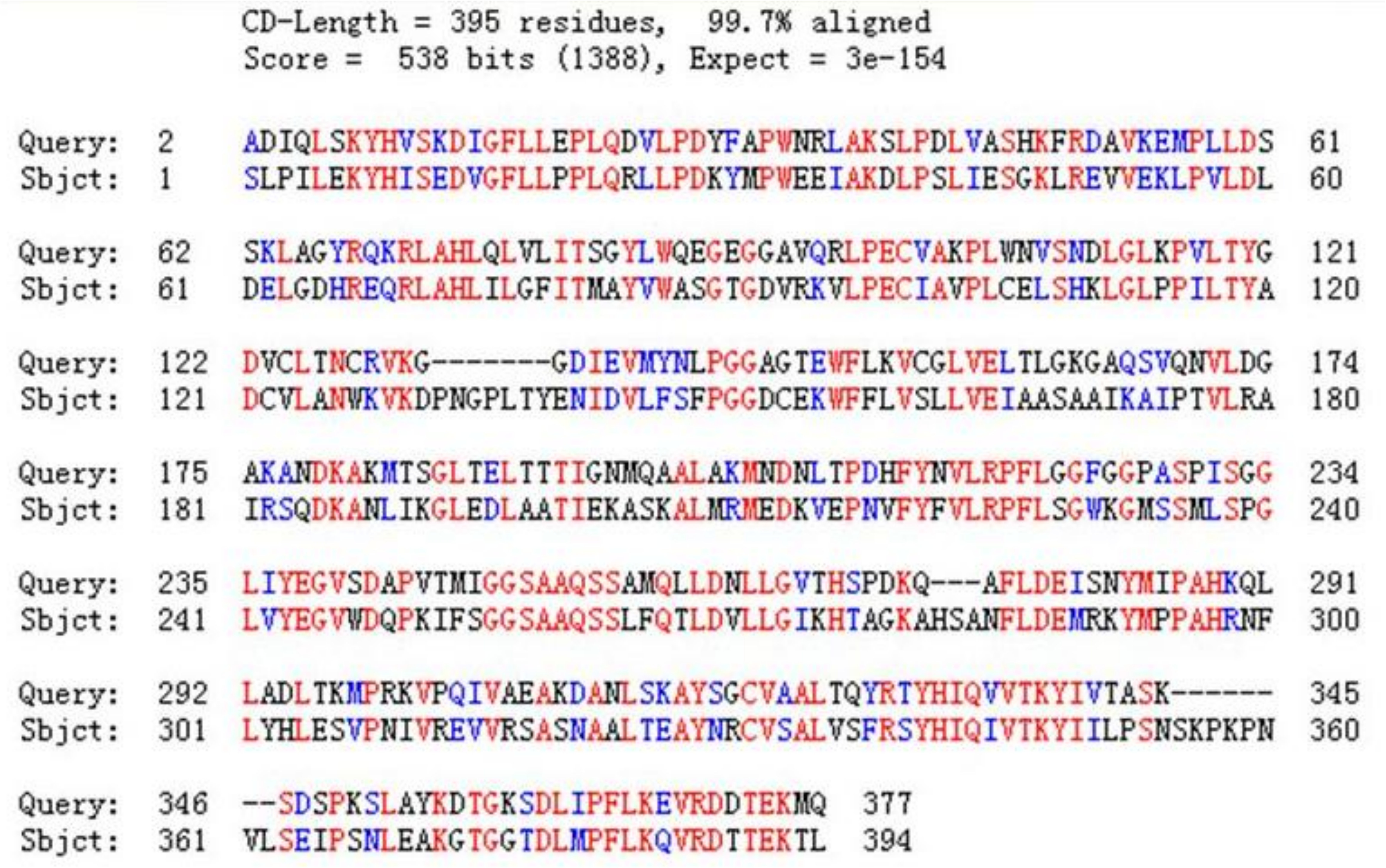
B) Three in the reverse direction (5'→3')

```
4   tac ggg ttc gac tta tcg cat ctc ccc aaa agt agt aaa ctc ctg cta cat atT
     H   G   L   Q   I   A   Y   L   P   K   *   *   K   L   V   I   Y   L

5   ta cgg gtt cga ctt atc gca tct ccc caa aag tag taa act cct gct aca taT t
        G   L   S   F   L   T   S   P   N   E   D   N   S   S   S   T   Y

6   t acg ggt tcg act tat cgc atc tcc cca aaa gta gta aac tcc tgc tac atA tt
       A   W   A   S   Y   R   L   P   T   K   M   M   Q   P   R   H   I
```

The stop codons are indicated with "*"

Adapted from: http://en.bioinformatica-na-escola.org/media/images/research3/orf.png

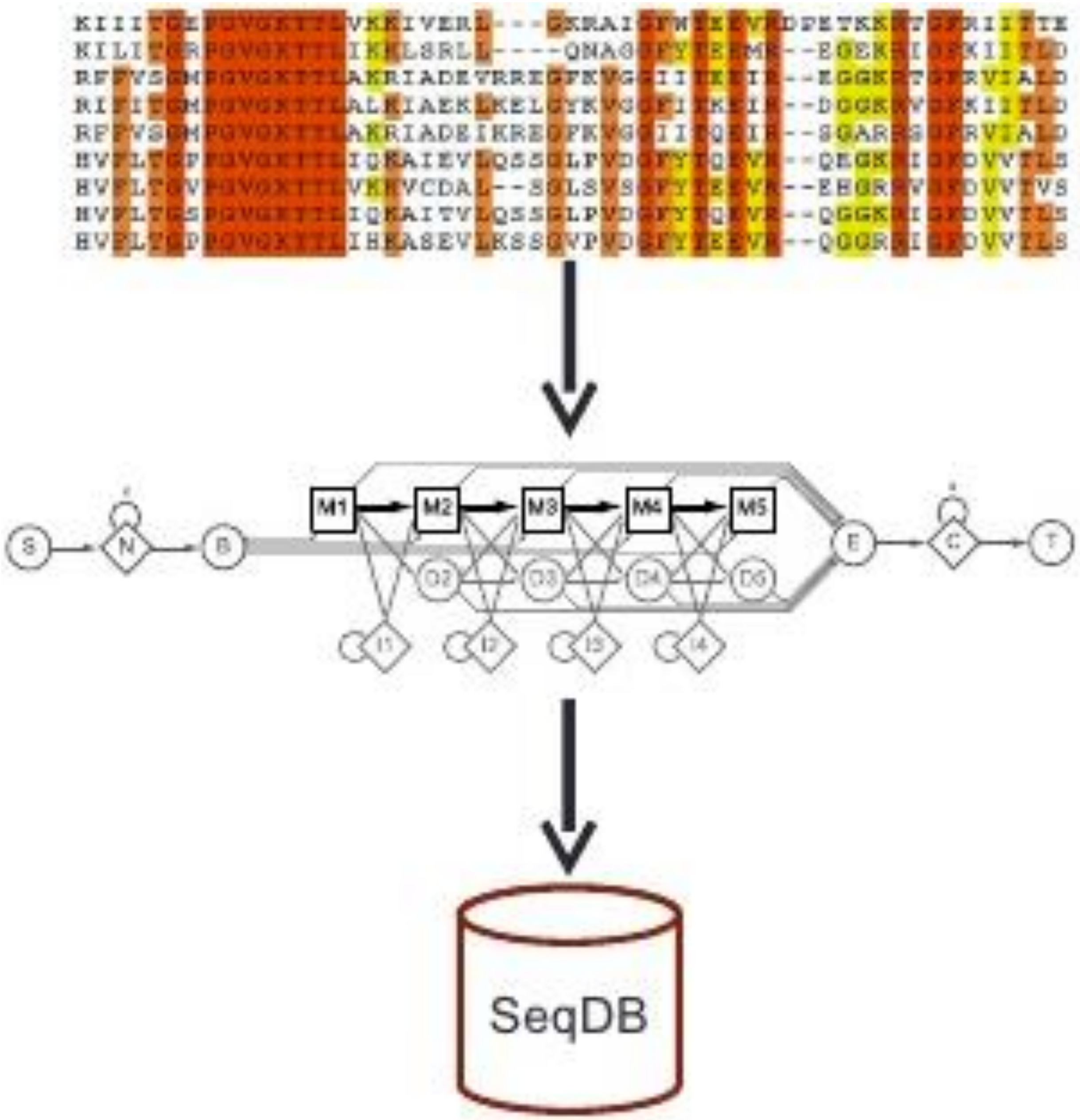
Search using blastp

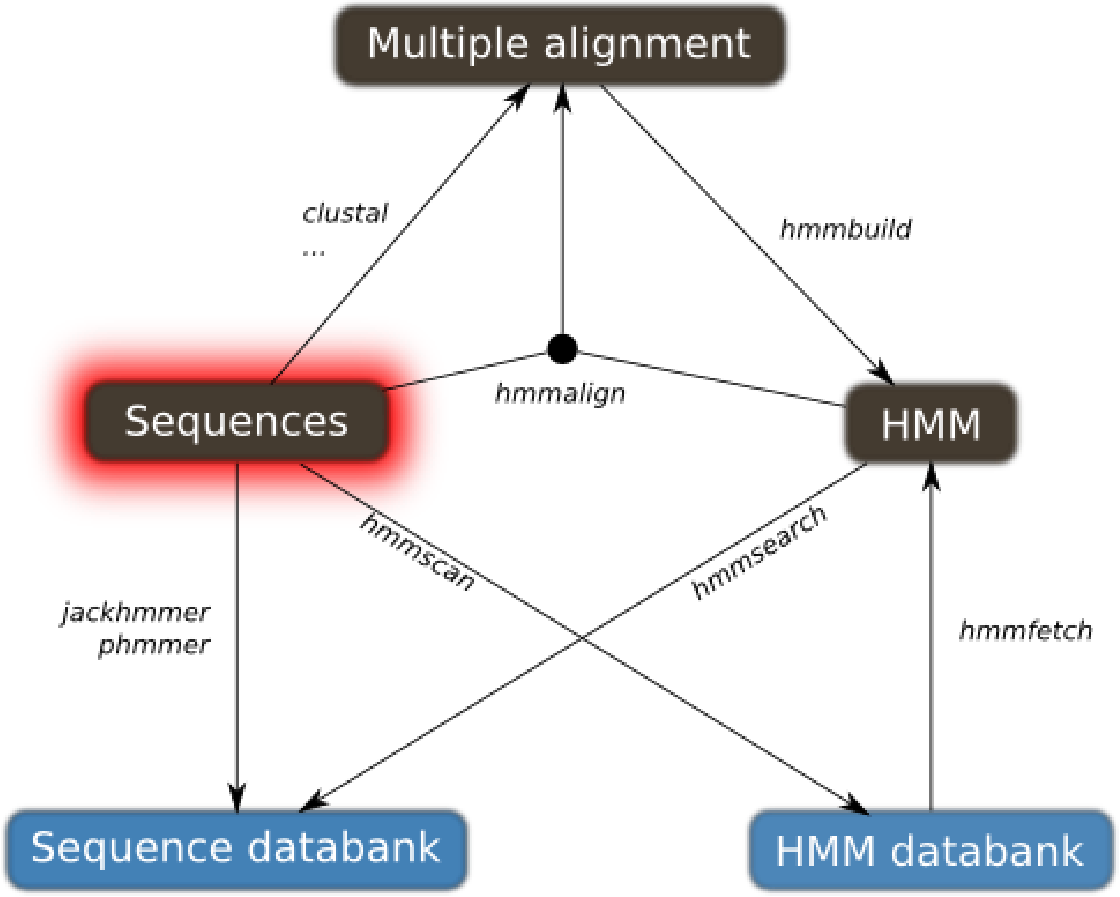


```
               CD-Length = 395 residues,  99.7% aligned
               Score =  538 bits (1388), Expect = 3e-154

Query:  2      ADIQLSKYHVSKDIGFLLEPLQDVLPDYFAPWNRLAKSLPDLVASHKFRDAVKEMPLLDS  61
Sbjct:  1      SLPILEKYHISEDVGFLLPPLQRLLPDKYMPWEEIAKDLPSLIESGKLREVVEKLPVLDL  60

Query:  62     SKLAGYRQKRLAHLQLVLITSGYLWQEGEGGAVQRLPECVAKPLWNVSNDLGLKPVLTYG  121
Sbjct:  61     DELGDHREQRLAHLILGFITMAYVWASGTGDVRKVLPECIAVPLCELSHKLGLPPILTYA  120

Query:  122    DVCLTNCRVKG--------GDIEVMYNLPGGAGTEWFLKVCGLVELTLGKGAQSVQNVLDG  174
Sbjct:  121    DCVLANWKVKDPNGPLTYENIDVLFSFPGGDCEKWFFLVSLLVEIAASAAIKAIPTVLRA  180

Query:  175    AKANDKAKMTSGLTELTTTIGNMQAALAKMNDNLTPDHFYNVLRPFLGGFGGPASPISGG  234
Sbjct:  181    IRSQDKANLIKGLEDLAATIEKASKALMRMEDKVEPNVFYFVLRPFLSGWKGMSSMLSPG  240

Query:  235    LIYEGVSDAPVTMIGGSAAQSSAMQLLDNLLGVTHSPDKQ---AFLDEISNYMIPAHKQL  291
Sbjct:  241    LVYEGVWDQPKIFSGGSAAQSSLFQTLDVLLGIKHTAGKAHSANFLDEMRKYMPPAHRNF  300

Query:  292    LADLTKMPRKVPQIVAEAKDANLSKAYSGCVAALTQYRTYHIQVVTKYIVTASK------  345
Sbjct:  301    LYHLESVPNIVREVVRSASNAALTEAYNRCVSALVSFRSYHIQIVTKYIILPSNSKPKPN  360

Query:  346    --SDSPKSLAYKDTGKSDLIPFLKEVRDDTEKMQ  377
Sbjct:  361    VLSEIPSNLEAKGTGGTDLMPFLKQVRDTTEKTL  394
```

Adapted from: Johnathan Bruce’s slides on, “What is BLAST?”

Search using superior Profile HMM models via HMMer



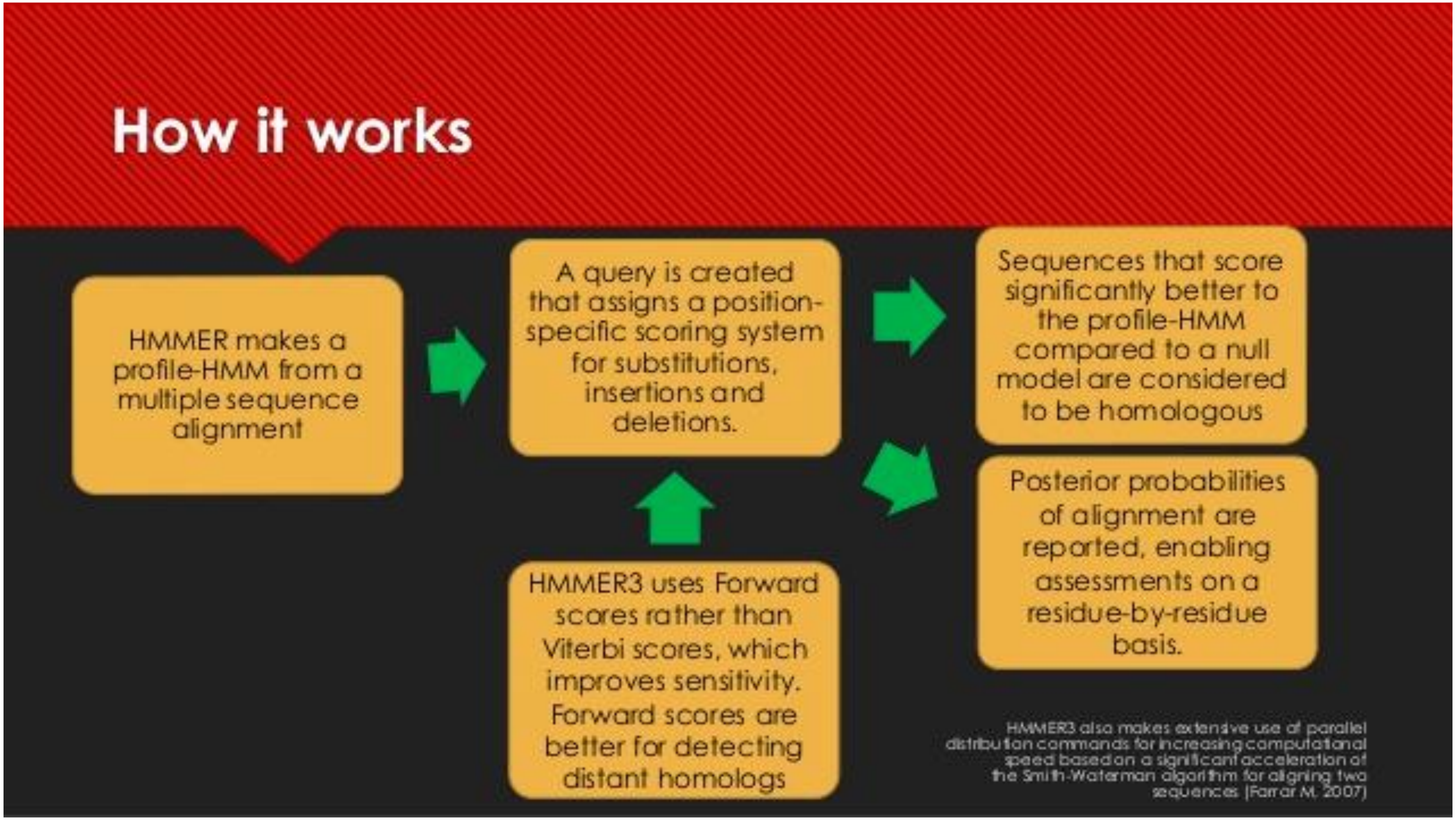


Adapted from: http://drmotifs.genouest.org/wp-content/uploads/2010/10/hmmer2.png

What is HMMer?

- HMMer is a open source program suite used to implement profile HMM for biological sequence analysis.
- Used to make the Pfam database of protein families
- Introduction to HMMer: http://pt.slideshare.net/anaxfotopoulos/introduction-to-hmmer-a-hidden-markov

…for those interested



How it works

HMMER makes a profile-HMM from a multiple sequence alignment

A query is created that assigns a position-specific scoring system for substitutions, insertions and deletions.

HMMER3 uses Forward scores rather than Viterbi scores, which improves sensitivity. Forward scores are better for detecting distant homologs

Sequences that score significantly better to the profile-HMM compared to a null model are considered to be homologous

Posterior probabilities of alignment are reported, enabling assessments on a residue-by-residue basis.

HMMER3 also makes extensive use of parallel distribution commands for increasing computational speed based on a significant acceleration of the Smith-Waterman algorithm for aligning two sequences (Farrar M. 2007)

Adapted from: http://pt.slideshare.net/anaxfotopoulos/introduction-to-hmmer-a-hidden-markov

Please turn in:

- A README.txt with the answers to the questions and the commands you used to answer the questions.
- A **commented** call_orfs.py to identify ORFs in a fasta of contigs
- A **commented** compare_orf_callers.py to compare MetaGeneMark output to your call_orfs.py output.
- A **commented** count_ar_genes_from_blast.py that filters your blast output.
- A **commented** count_ar_genes_from_resfams.py that counts the number of genes in a hmmscan output file.
- Optional: a **commented** gff_to_nt_aa.py that takes a MetaGeneMark GFF file and outputs the nucleotide and amino acid sequences from the predicted ORFs.
- All files created from the above scripts or commands: all_orfs.fna, all_proteins.faa, mgm_predictions.gff, mgm_orfs.fna, mgm_orfs.faa, mgm_predictions.gff, blast_to_card.txt, and resfams_annotations.txt
- **Due in 3 weeks (3/23/15) at 10 AM**

Questions

Adapted from: http://assets.illumina.com/content/dam/illumina-marketing/images/applications/microbiology/v2/intro-environmental-metagenomics.jpg

Good luck on your (metagenomic) journey! :)

Adapted from: http://mycor.nancy.inra.fr/blogGenomes/wp-content/uploads/2011/04/20110326_STD0011.jpg