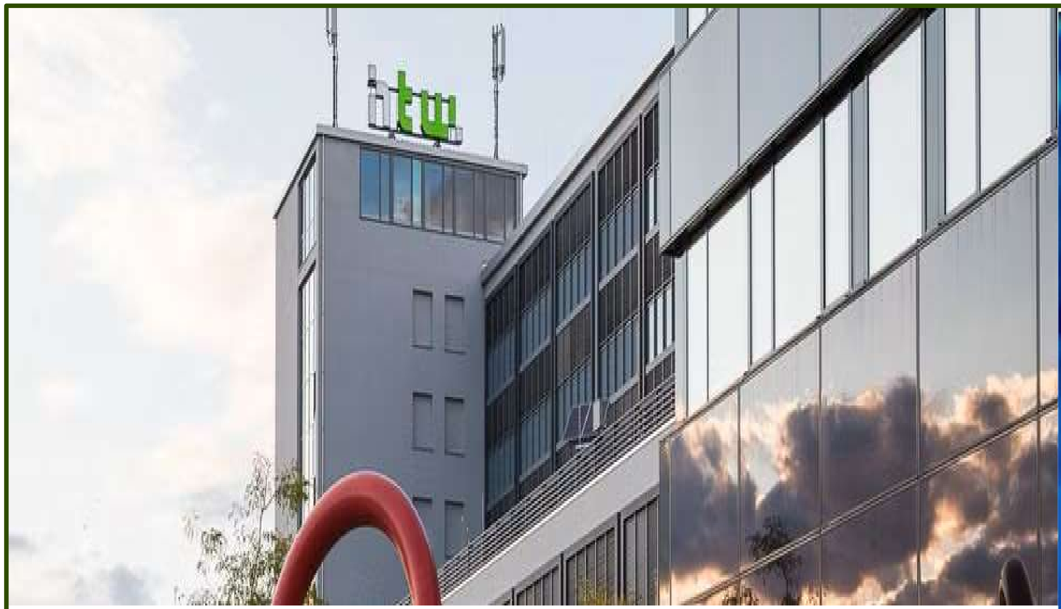
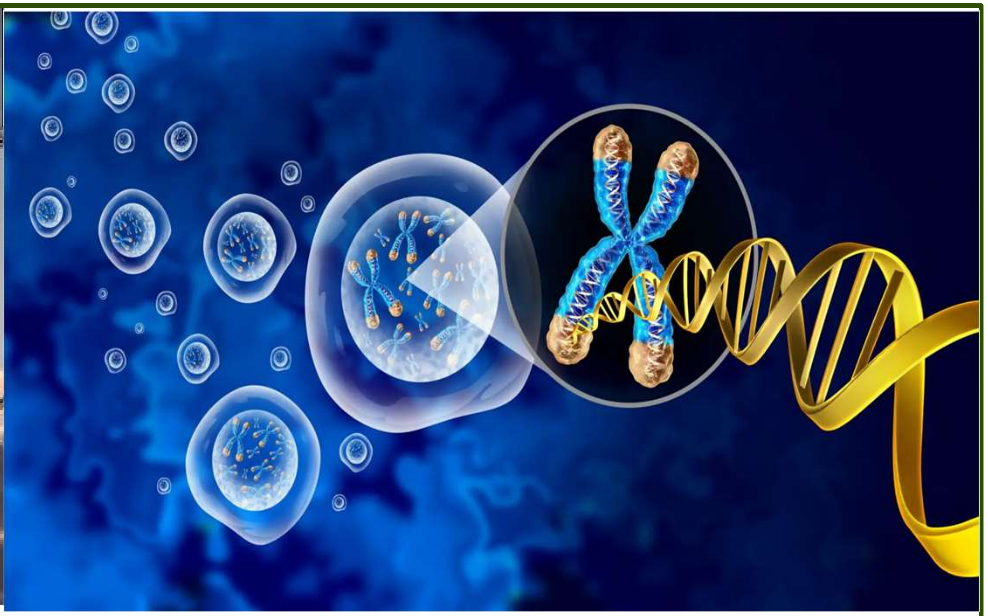




Finding patterns in Genome Data

Project Study - Summer Semester 2024

Presentation Date: 19th July 2024, Friday

Masters in Professional IT Business and Digitalization

Hochschule für Technik und Wirtschaft Berlin

University of Applied Sciences

Project Guidance

- Prof. Piotr Wojciech Dabrowski

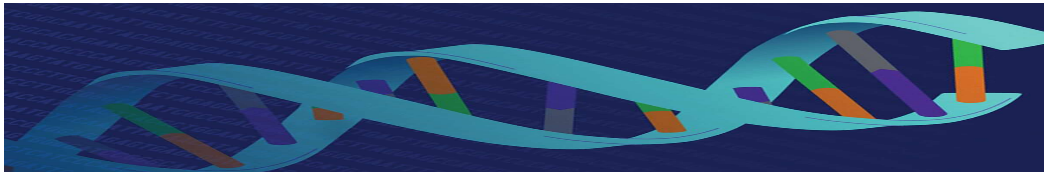
  Hochschule für Technik und Wirtschaft (HTW) Berlin

Team Members

- Kalyan Shencottah – [redacted]

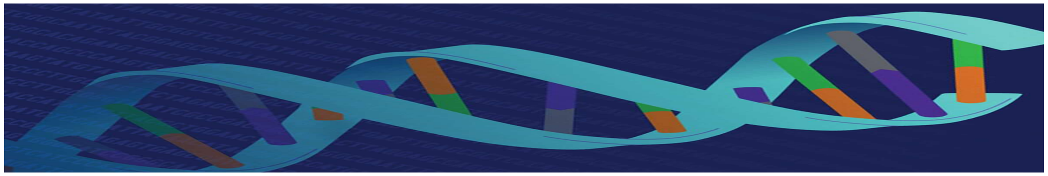
Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences



Agenda

- *Project Study Overview*
  - Motivation
  - Objective
- *The Approach*
  - Conceptual Understanding of Genome Data
  - Data Collection from Genome Databases
  - Genome Data Interpretation
  - Things Accomplished



- *Technical Implementation*
  - Architecture, Technical Components, Algorithms
  - Processing Coding Sequences
  - Processing Non-Coding Sequences
  - Generate clean data
- *Challenges and Resolutions*
  - Example Challenges and Resolutions
- *Conclusion and Next Steps*
- *Extras – Python Code*

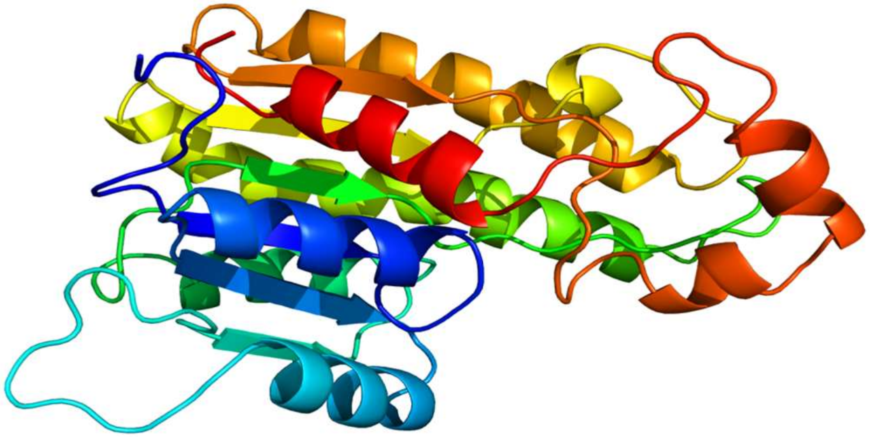
htw

Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Global Motivation

- ***Genome Data***
  - Evolving field of Bio-Informatics
  - Complex Data – DNA, Proteins
  - Discover Mutations, Cause of Diseases



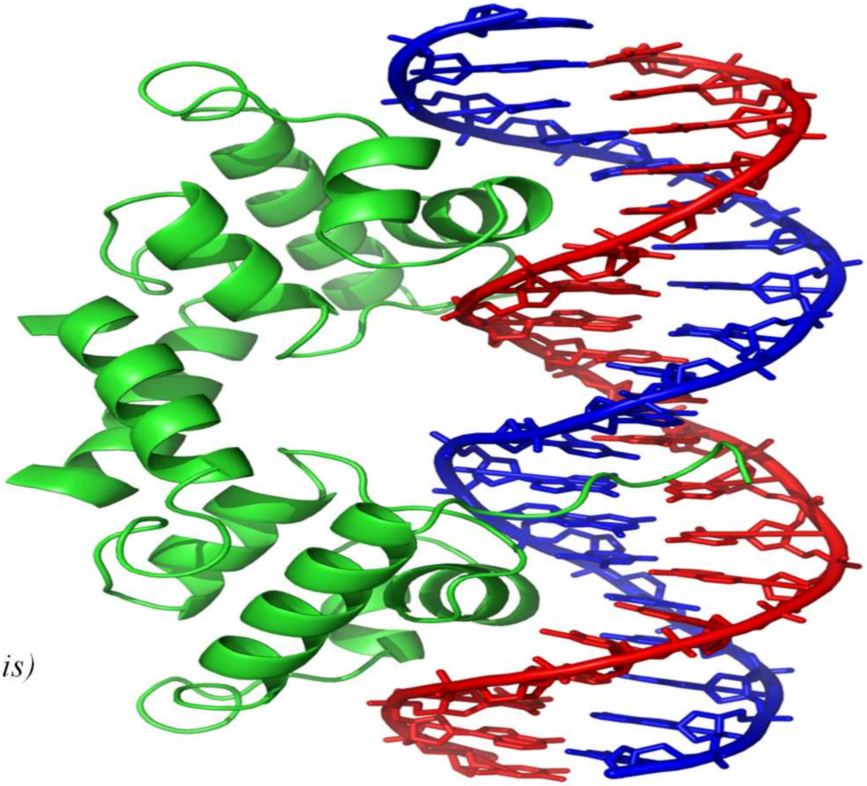
- ***Many ML Algorithm Possibilities***
  - Supervised and Unsupervised
  - Ex. Classification, Clustering, Neural Network

Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Personal Motivation

- ***To gain deeper understanding on BioInformatics / Computational Biology***
- ***Apply my Knowledge gained***
  - *Machine Intelligence Lab, Univ of Cincinnati, USA*
    - *ML, Pattern Recognition, Spatial Data Mining(Thesis)*
  - *HTW Course work*
- ***Python and Sequence Analysis***



Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Objective

- ***Research on Genome Data / Databases***
  - *NCBI / Ensembl Databases*
  - *Identify Organism (say, Bacteria)*
  - *RefSeq files (Sequence files / FASTA format)*
  - *Feature files (GFF files)*
- ***Build a Pipeline***
  - *Extract Sequence files from Genome Databases*
  - *Retrieve Data pertaining to Features (CDS, Gene)*
  - *Identify Coding and Non-Coding Regions*
  - *Store Coding and Non-Coding regions in FASTA files*
  - *Preparation of Training and Test Data to be fed to NN*

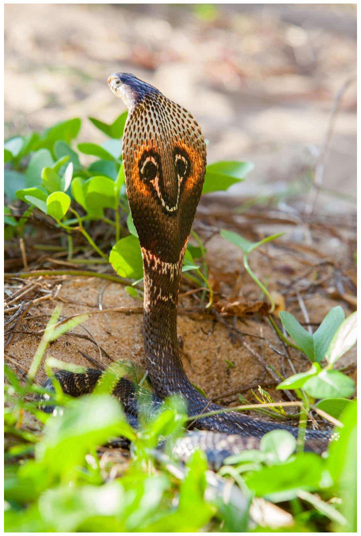
htw

Hochschule für Technik
und Wirtschaft Berlin

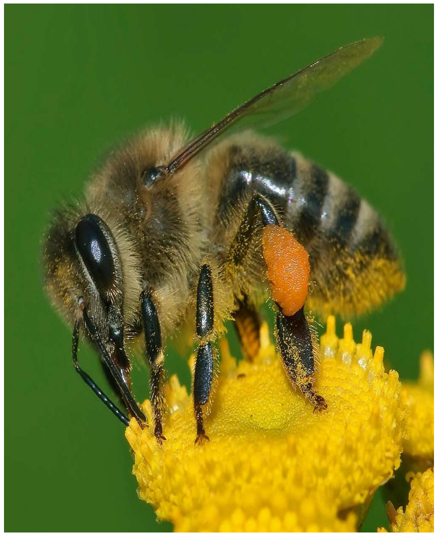
University of Applied Sciences

Conceptual Understanding of Genome Data

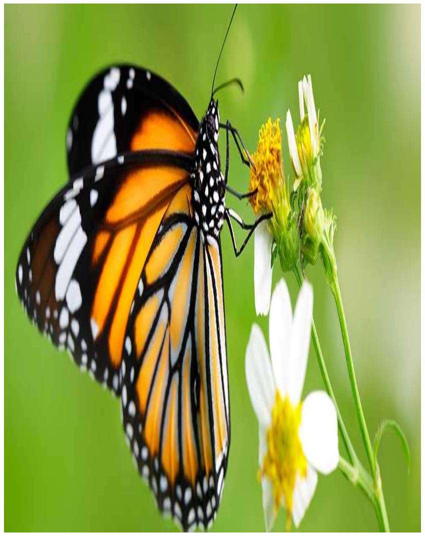
COUPLED



NON-LINEAR

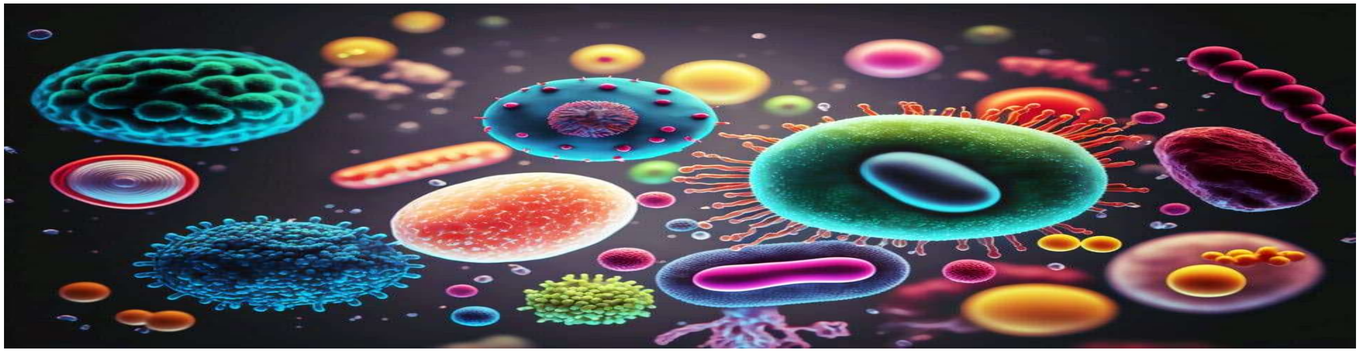


COMPLEXITY



PLASTICITY

NON-EQUILIBRIUM

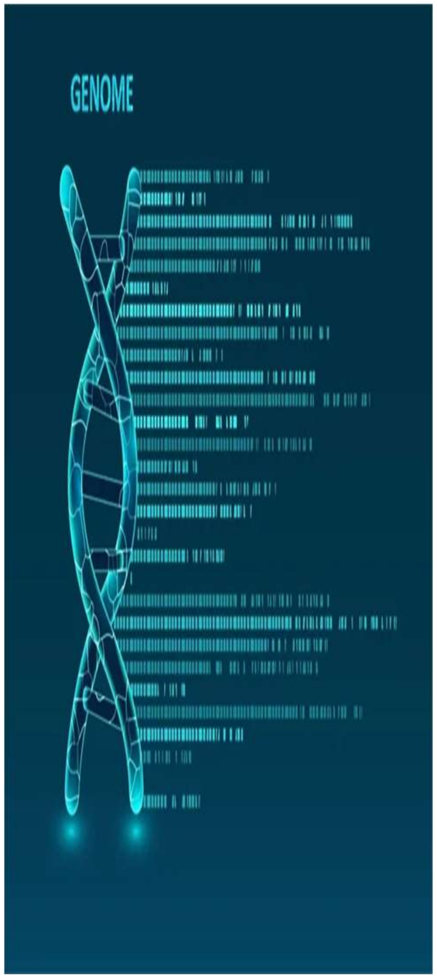


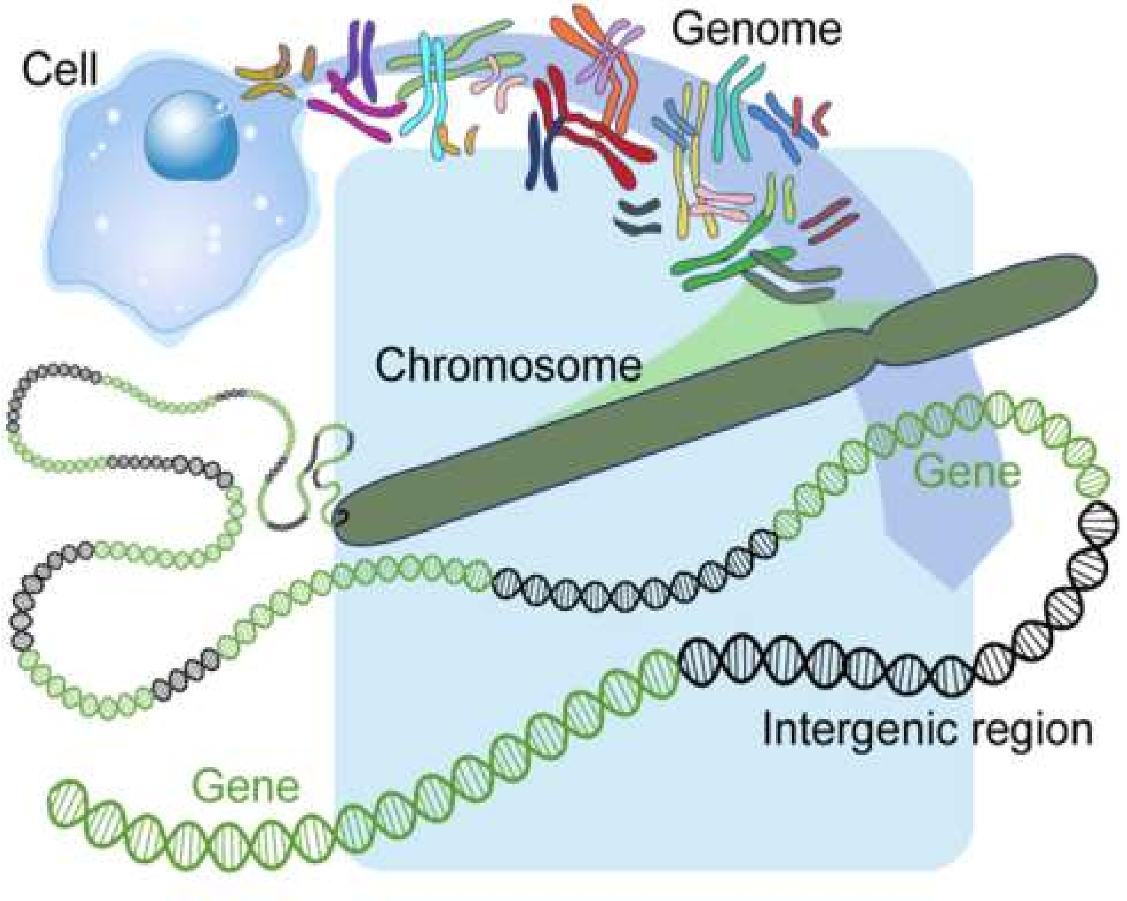
COMPLEX PHENOTYPES

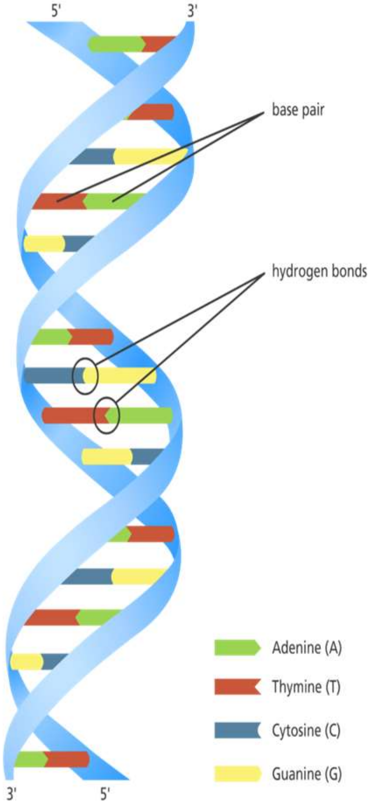
Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Conceptual understanding of Genome





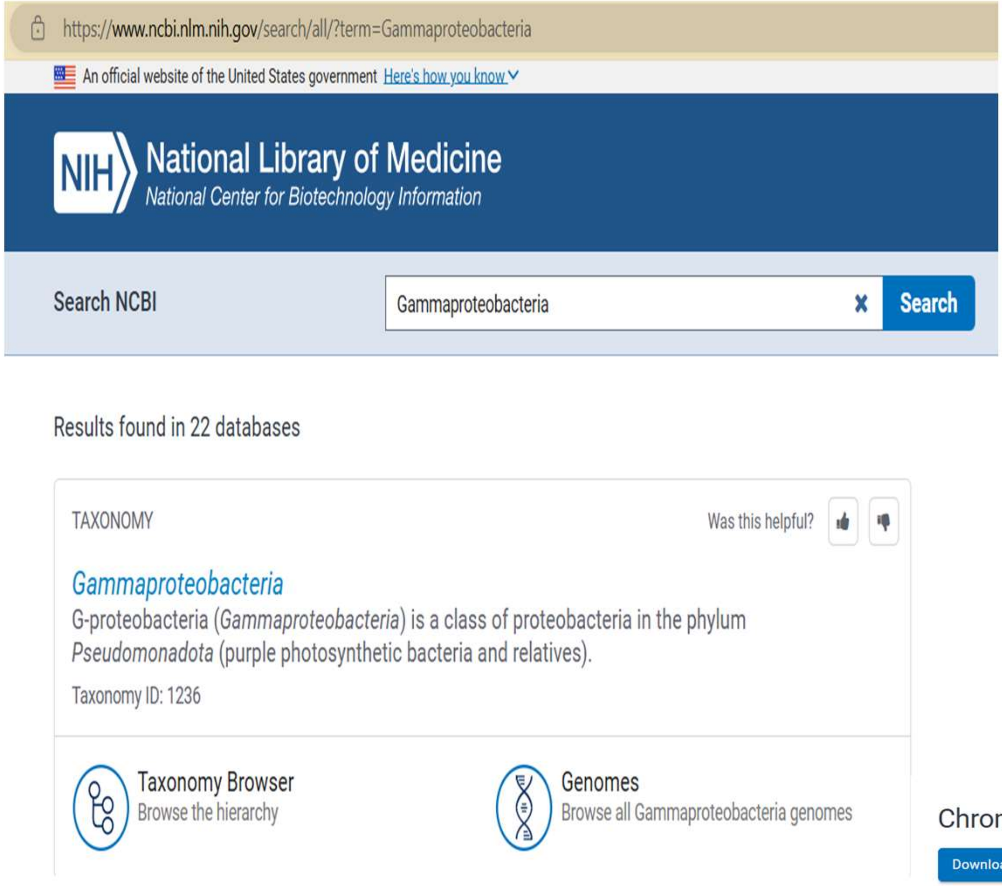


Two types of Organisms: Prokaryotes(Bacteria), Eukaryotes (Humans, Animals, Plants)

Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Data Collection from Genome Databases



https://www.ncbi.nlm.nih.gov/search/all/?term=Gammaproteobacteria

An official website of the United States government Here's how you know

NIH National Library of Medicine
National Center for Biotechnology Information

Search NCBI: Gammaproteobacteria ✖ Search

Results found in 22 databases

TAXONOMY — Was this helpful?

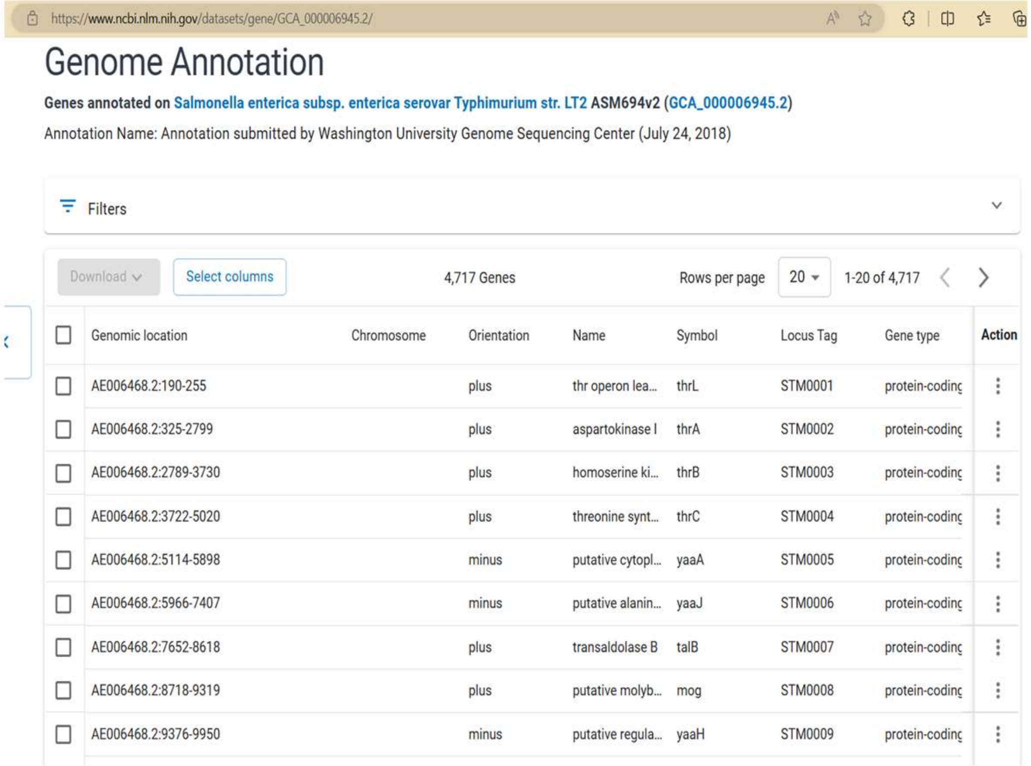
Gammaproteobacteria

G-proteobacteria (*Gammaproteobacteria*) is a class of proteobacteria in the phylum *Pseudomonadota* (purple photosynthetic bacteria and relatives).

Taxonomy ID: 1236

Taxonomy Browser
Browse the hierarchy

Genomes
Browse all Gammaproteobacteria genomes



https://www.ncbi.nlm.nih.gov/datasets/gene/GCA_000006945.2/

Genome Annotation

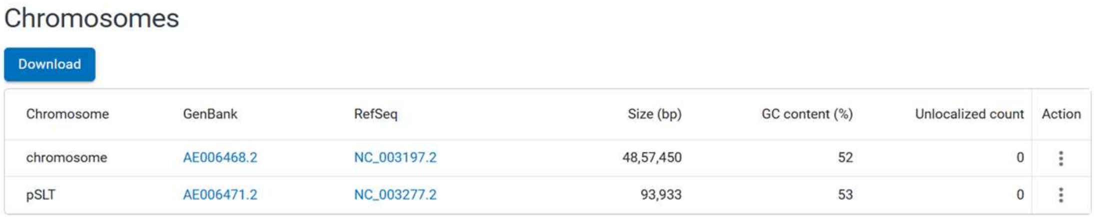
Genes annotated on Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 ASM694v2 (GCA_000006945.2)

Annotation Name: Annotation submitted by Washington University Genome Sequencing Center (July 24, 2018)

Filters

Download | Select columns | 4,717 Genes | Rows per page 20 | 1-20 of 4,717

□	Genomic location	Chromosome	Orientation	Name	Symbol	Locus Tag	Gene type	Action
□	AE006468.2:190-255		plus	thr operon lea...	thrL	STM0001	protein-codinç	⋮
□	AE006468.2:325-2799		plus	aspartokinase I	thrA	STM0002	protein-codinç	⋮
□	AE006468.2:2789-3730		plus	homoserine ki...	thrB	STM0003	protein-codinç	⋮
□	AE006468.2:3722-5020		plus	threonine synt...	thrC	STM0004	protein-codinç	⋮
□	AE006468.2:5114-5898		minus	putative cytopl...	yaaA	STM0005	protein-codinç	⋮
□	AE006468.2:5966-7407		minus	putative alanin...	yaaJ	STM0006	protein-codinç	⋮
□	AE006468.2:7652-8618		plus	transaldolase B	talB	STM0007	protein-codinç	⋮
□	AE006468.2:8718-9319		plus	putative molyb...	mog	STM0008	protein-codinç	⋮
□	AE006468.2:9376-9950		minus	putative regula...	yaaH	STM0009	protein-codinç	⋮

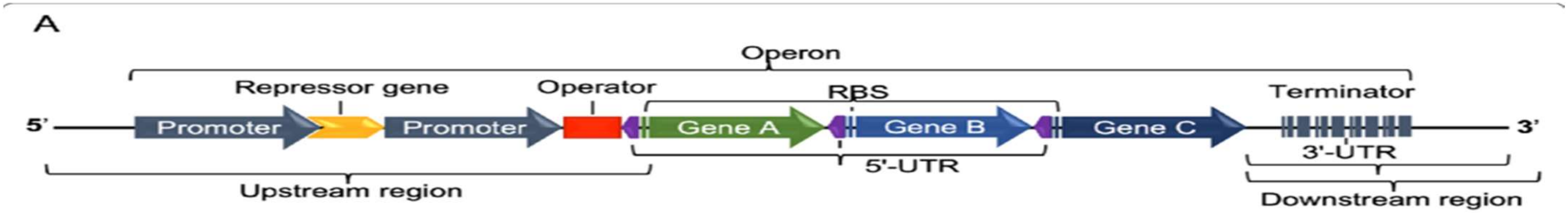


Chromosomes

Download

Chromosome	GenBank	RefSeq	Size (bp)	GC content (%)	Unlocalized count	Action
chromosome	AE006468.2	NC_003197.2	48,57,450	52	0	⋮
pSLT	AE006471.2	NC_003277.2	93,933	53	0	⋮

Interpreting Genome Data (Prokaryotes-Bacteria)



UTR – Untranslated Regions

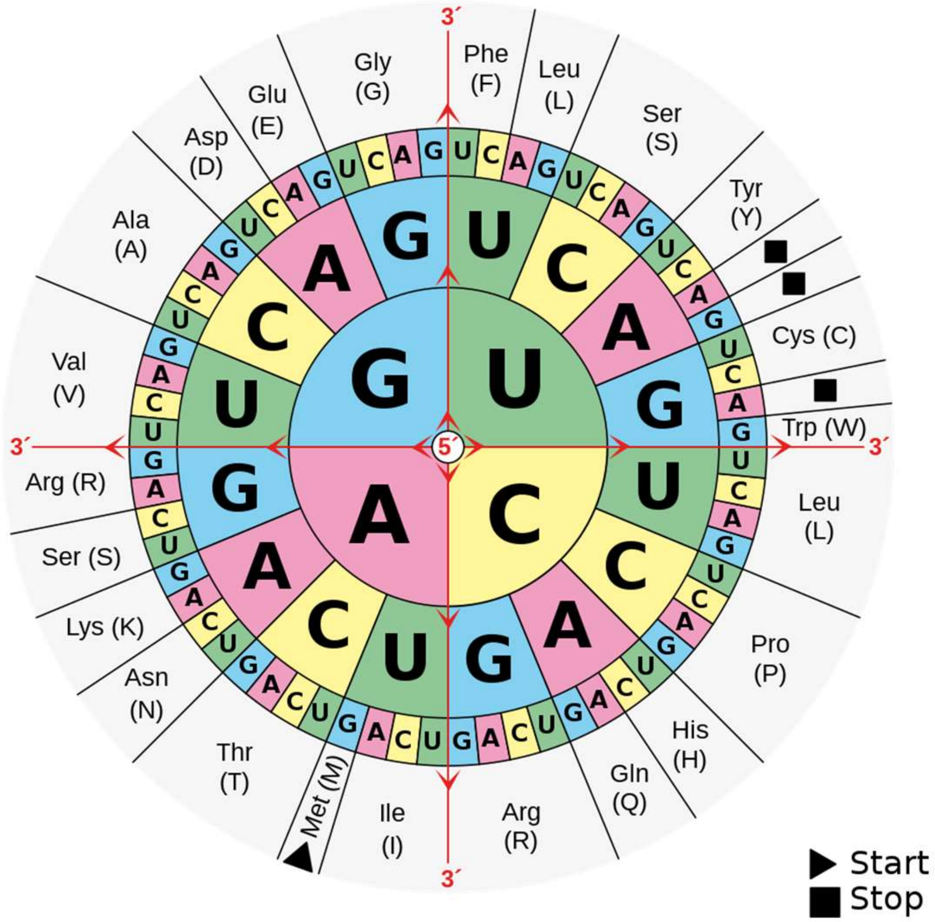
RBS – Ribosome Binding Sites
Codons – 3 Nucleotides – START and STOP
ORF – Open Reading Frame

Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Source: Wikipedia

Interpreting Genome Data – Mapping Codon Table



Start Codon – AUG (ATG)

Initiates translation process

Stop Codon – UAA, UAG, UGA (TAA, TAG, TGA)

Initiates the termination

DNA	**(m)RNA**
A – Adenine	A
T – Thymine	T
C - Cytosine	U - Uracil
G - Guanine	G

htw

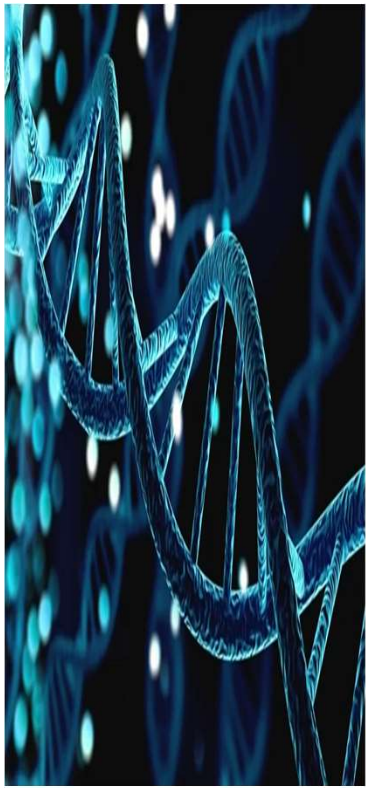
Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Source: Wikipedia

Things Accomplished

- **Download Data from NCBI Database**
  - *RefSeq/GenBank (FASTA files, GFF files)*
  - *FASTA – Sequence files, GFF – Features*
- **Processing of Fasta Sequences**
  - *Create separate files for Genes (say, AE006468.2, AE006471.2)*
  - *Cleanup unwanted characters(say, N), spaces*
- **Processing of GFF Files (Features)**
  - *Create Features DB (SQLLite)*
  - *Create different GFF files for Genes (say, AE006468.2, AE006471.2)*
- **For the Sequences, for given feature (say, CDS)**
  - *Identify Coding and Non-Coding Sequences*
  - *Process Forward (+) Strands and Reverse (-) Strand. Reverse complement (-) strands*
  - *Generate Coding Sequences in FASTA format, Non-Coding Sequences in FASTA format*
  - *Finalize Training and Test Data to be fed to NN*

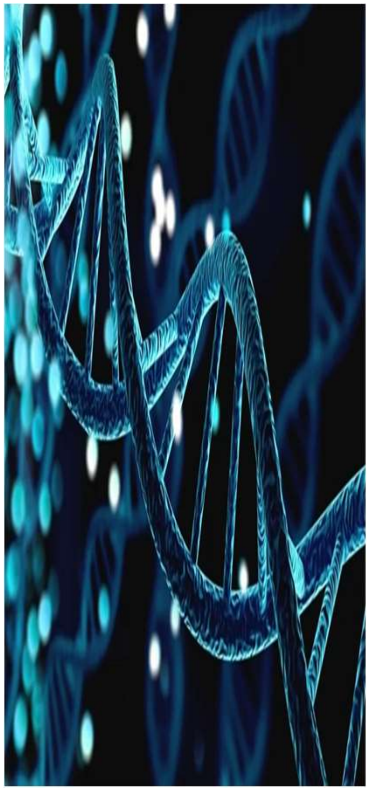


htw

Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Technical Implementation



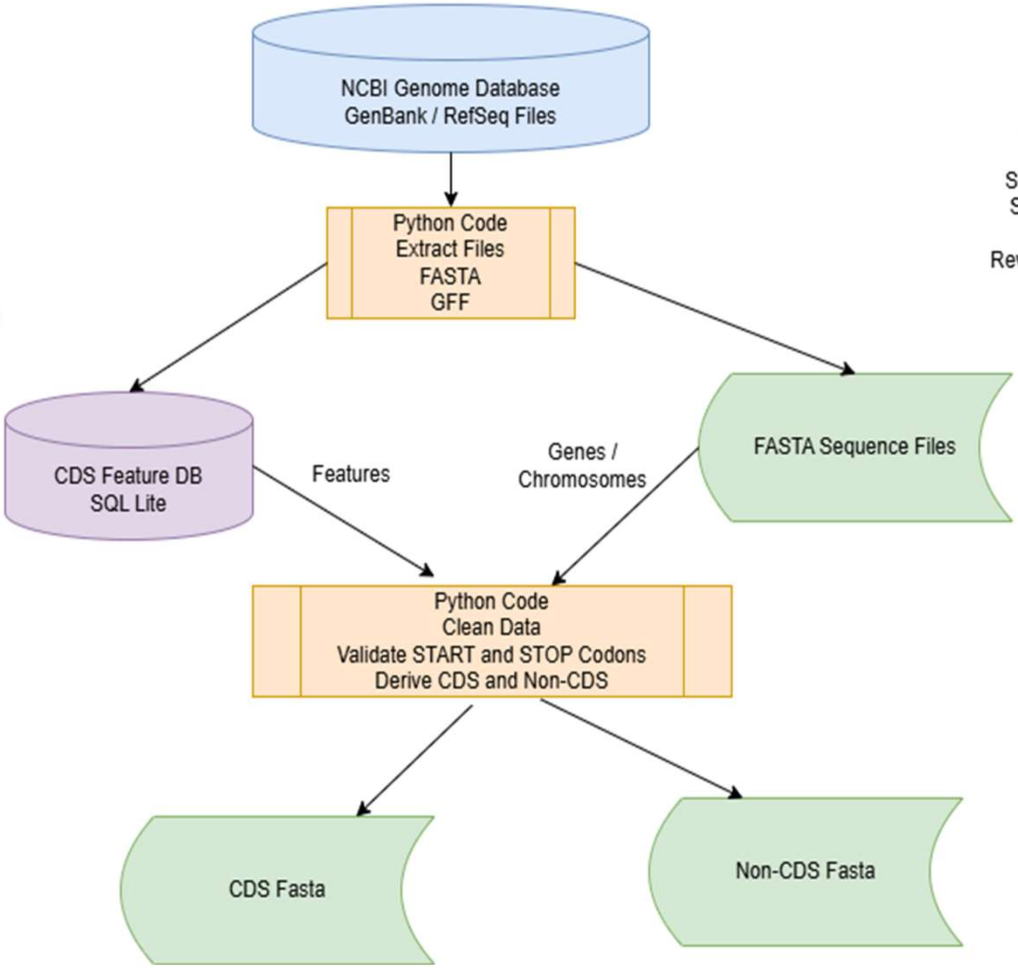
Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Architecture For Finding Patterns in Genome Data

Technical Components:

Python
BioPython
SQL Lite DB
FASTA Files
GFF Features Files



Algorithm

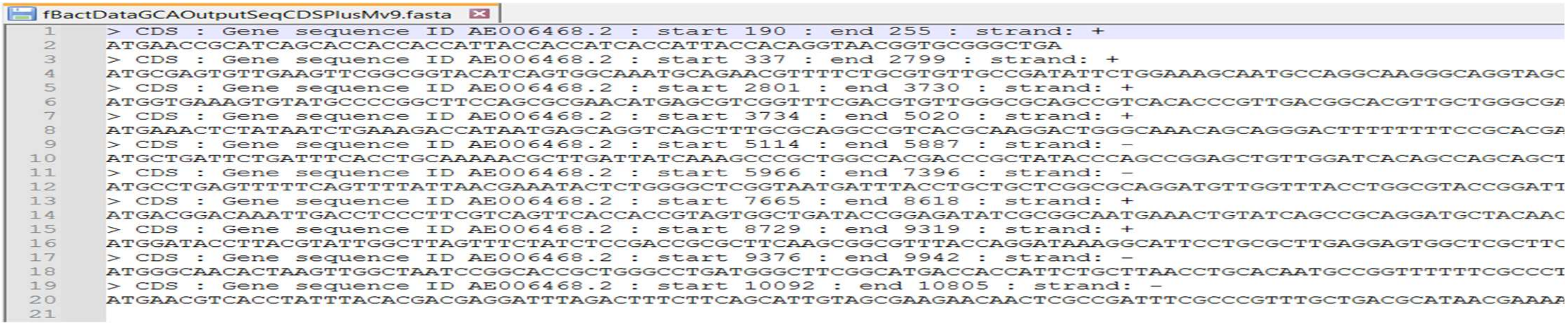
Sequence Parsing
Sequence Slicing
Indexing
Reverse Complement

Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Generate Output in FASTA – CDS and Non-CDS

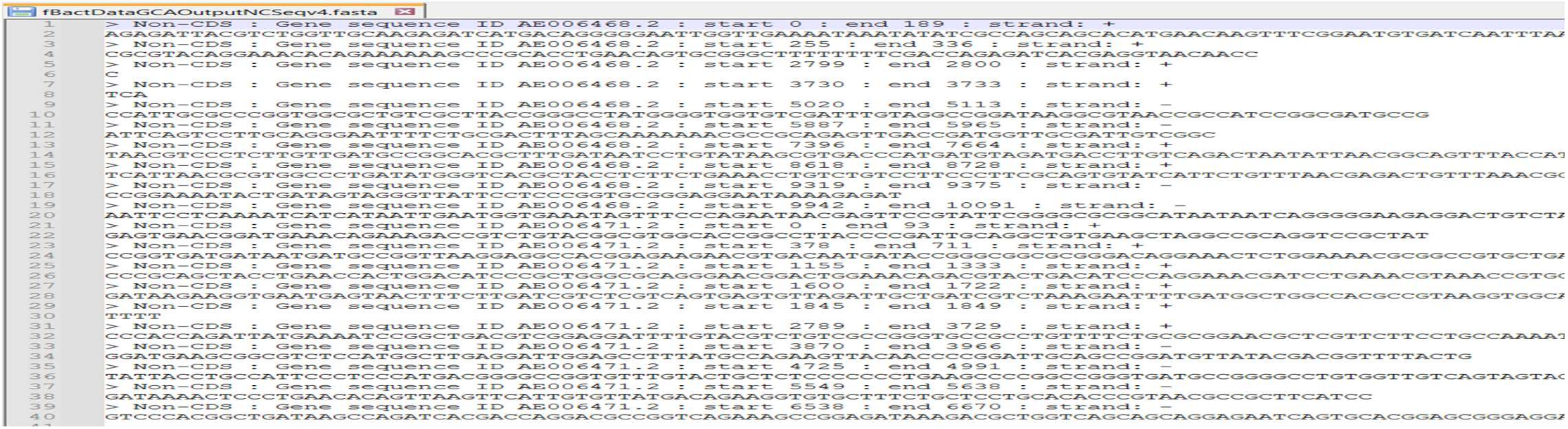
CDS – Coding Sequences FASTA Output



fBactDataGCAOutputSeqCDSPlusMv9.fasta

```
> CDS : Gene sequence ID AE006468.2 : start 190 : end 255 : strand: +
ATGAACCGCATCAGCACCACCACCATTACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGA
> CDS : Gene sequence ID AE006468.2 : start 337 : end 2799 : strand: +
ATGCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTGGAAAGCAATGCCAGGCAAGGGCAGGTAGC
> CDS : Gene sequence ID AE006468.2 : start 2801 : end 3730 : strand: +
ATGGTGAAAGTGTATGCCCCGGCTTCCAGCGCGAACATGAGCGTCGGTTTCGACGTGTTGGGCGCAGCCGTCACACCCGTTGACGGCACGTTGCTGGGCGA
> CDS : Gene sequence ID AE006468.2 : start 3734 : end 5020 : strand: +
ATGAAACTCTATAATCTGAAAGACCATAATGAGCAGGTCAGCTTTGCGCAGGCCGTCACGCAAGGACTGGGCAAACAGCAGGGACTTTTTTTTCCGCACGA
> CDS : Gene sequence ID AE006468.2 : start 5114 : end 5887 : strand: –
ATGCTGATTCTGATTTCACCTGCAAAAACGCTTGATTATCAAAGCCCGCTGGCCACGACCCGCTATACCCAGCCGGAGCTGTTGGATCACAGCCAGCAGCT
> CDS : Gene sequence ID AE006468.2 : start 5966 : end 7396 : strand: –
ATGCCTGAGTTTTTCAGTTTTATTAACGAAATACTCTGGGGCTCGGTAATGATTTACCTGCTGCTCGGCGCAGGATGTTGGTTTACCTGGCGTACCGGATT
> CDS : Gene sequence ID AE006468.2 : start 7665 : end 8618 : strand: +
ATGACGGACAAATTGACCTCCCTTCGTCAGTTCACCACCGTAGTGGCTGATACCGGAGATATCGCGGCAATGAAACTGTATCAGCCGCAGGATGCTACAAC
> CDS : Gene sequence ID AE006468.2 : start 8729 : end 9319 : strand: +
ATGGATACCTTACGTATTGGCTTAGTTTCTATCTCCGACCGCGCTTCAAGCGGCGTTTACCAGGATAAAGGCATTCCTGCGCTTGAGGAGTGGCTCGCTTC
> CDS : Gene sequence ID AE006468.2 : start 9376 : end 9942 : strand: –
ATGGGCAACACTAAGTTGGCTAATCCGGCACCGCTGGGCCTGATGGGCTTCGGCATGACCACCATTCTGCTTAACCTGCACAATGCCGGTTTTTTCGCCCT
> CDS : Gene sequence ID AE006468.2 : start 10092 : end 10805 : strand: –
ATGAACGTCACCTATTTACACGACGAGGATTTAGACTTTCTTCAGCATTGTAGCGAAGAACAACTCGCCGATTTCGCCCGTTTGCTGACGCATAACGAAAA
```

Non-CDS – Non-Coding Sequences FASTA Output

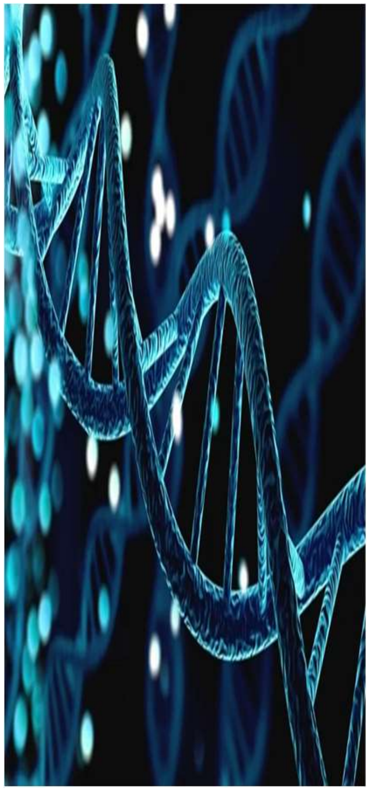


fBactDataGCAOutputNCSeqv4.fasta

```
> Non-CDS : Gene sequence ID AE006468.2 : start 0 : end 189 : strand: +
AGAGATTACGTCTGGTTGCAAGAGATCATGACAGGGGGAATTGGTTGAAAATAAATATATCGCCAGCAGCACATGAACAAGTTTCGGAATGTGATCAATTTAA
> Non-CDS : Gene sequence ID AE006468.2 : start 255 : end 336 : strand: +
CGCGTACAGGAAACACAGAAAAAAAGCCCGCACCTGAACAGTGCGGGCTTTTTTTTTCGACCAGAGATCACGAGGTAACAACC
> Non-CDS : Gene sequence ID AE006468.2 : start 2799 : end 2800 : strand: +
C
> Non-CDS : Gene sequence ID AE006468.2 : start 3730 : end 3733 : strand: +
TCA
> Non-CDS : Gene sequence ID AE006468.2 : start 5020 : end 5113 : strand: –
CCATTGCGCCCGGTGGCGCTGTCGCTTACCGGGCCTATGGGGTGGTGTCGATTTGTAGGCCGGATAAGGCGTAACCGCCATCCGGCGATGCCG
> Non-CDS : Gene sequence ID AE006468.2 : start 5887 : end 5965 : strand: –
ATTCAGTCCTTGCAGGGAATTTTCTGCGACTTTAGCAAAAAAACGCCGCAGAGTTGACCGATGGTTGCGATTGTCGGC
> Non-CDS : Gene sequence ID AE006468.2 : start 7396 : end 7664 : strand: +
TAACGTCCCTCTTGTTGATGCCGGCACGCTTTGATAATCCTGTATAAGCGTGACCCATGATGTAGATGACCTTGTCAGACTAATATTAACGGCAGTTTACCAT
> Non-CDS : Gene sequence ID AE006468.2 : start 8618 : end 8728 : strand: +
TCATTAACGCGTGGCCCTGATATGGGTCACGCTACCTCTTCTGAAACCTGTCTGTCCTTCCCTTCGCAGTGTATCATTCTGTTTAACGAGACTGTTTAAACGC
> Non-CDS : Gene sequence ID AE006468.2 : start 9319 : end 9375 : strand: –
CCGGAAAATACTGATAGTAGGGTTATTCCTCCCGGTGCGGGAGGAATAAAAGAGAT
> Non-CDS : Gene sequence ID AE006468.2 : start 9942 : end 10091 : strand: –
AATTCCTCAAAATCATCATAATTGAATGGTGAAATAGTTTCCCAGAATAACGAGTTCCGTATTCGGGGCGCGGCATAATAATCAGGGGGAAGAGGACTGTCTA
> Non-CDS : Gene sequence ID AE006471.2 : start 0 : end 93 : strand: +
GAGTGAACGGATGAAACAGAAAGACCGTCTGTACGGCGTGGCACCGGCCTTACCCCGATTGCAGGCTGTGAAGCTAGGCCGCAGGTCCGCTAT
> Non-CDS : Gene sequence ID AE006471.2 : start 378 : end 711 : strand: +
CCGGTGATGATAATGATGCCGGTTAAGGAGGCCACGGAGAAGAACGTGACAATGATACCGGGCGGCGCGGGACAGGAAACTCTGGAAAACGCGGCCGTGCTGA
> Non-CDS : Gene sequence ID AE006471.2 : start 1155 : end 1333 : strand: +
CCCGCAGCTACCTGAACCACTGGACATCCCGCTGGGCGCAGGGAACGGACTGGAAACAGACGTACTGACATCCCAGGAAACGATCCTGAAACGTAAACCGTGC
> Non-CDS : Gene sequence ID AE006471.2 : start 1600 : end 1722 : strand: +
GATAAGAAGGTGAATGAGTAACTTTCTTGATCGTCTCGTCAGTGAGTGTTAGATTGCTGATCGTCTAAAGAATTTTGATGGCTGGCCACGCCGTAAGGTGGCA
> Non-CDS : Gene sequence ID AE006471.2 : start 1845 : end 1849 : strand: +
TTTT
> Non-CDS : Gene sequence ID AE006471.2 : start 2789 : end 3729 : strand: +
CCCACCAGATTATGAAAATCCGGCTGACGTCGGAGGATTTTGTACGTCTGTCGCCGGGTGCCGCCTGTTTTCTGCGCGGAACGCTCGTTCTTCCTGCCAAAAT
> Non-CDS : Gene sequence ID AE006471.2 : start 3870 : end 3966 : strand: –
GGATGAAGCGGCGTCTCCATGGCTTGAGGATTGGAGCCTTTATGCCAGAAGTTACAACCCCGGATTGCAGCCGGATGTTATACGACGGTTTTACTG
> Non-CDS : Gene sequence ID AE006471.2 : start 4725 : end 4991 : strand: –
TATTACCTGCCATTCCCTCCCATGACGGGGCCGGTGTTTGTACTGCTCTCCCCCCCTGAAGCCCCGGCCGGGTGATGCCGGGGCCTGTGGTTGTCAGTAGTAC
> Non-CDS : Gene sequence ID AE006471.2 : start 5549 : end 5638 : strand: –
GATAAAACTCCCTGAACACAGTTAAGTTCATTGTGTTATGACAGAAGGTGTGCTTTCTGCTCCTGCACACCCGTAACGCCGCTTCATCC
> Non-CDS : Gene sequence ID AE006471.2 : start 6538 : end 6670 : strand: –
GTCCCACGGCTGATAAGCCAGATCACGACCAGGACGCCGGTCAGAAAGCCGGAGATAAAGACGCTGGTCAGCAGCAGGAGAATCAGTGCACGGAGCGGGAGGA
```

Challenges and Resolutions

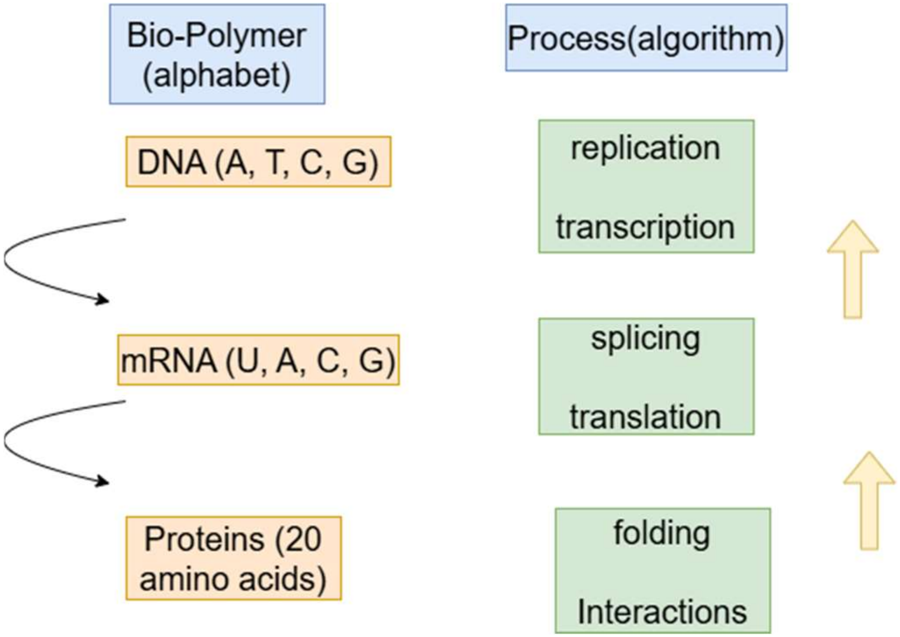
- **Finding an Organism to work with**
  - *Choose Bacteria. Understand complexity, clear goal on what data to work with*
- **Memory Issues in handling Datasets**
  - *Reduce the Data to execute Logic.*
- **Understanding Sequences, Features**
  - *Gain knowledge on the Features, its Data Structure*



Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Next Steps



Problem -> Algorithms

Sequencing -> Fragment assembly problem -> The shortest superstring problem

Gene Finding -> Hidden Markov Models, Pattern Recognition Methods

Sequence comparison -> pairwise and multiple sequence alignments -> Dynamic programming algorithm, Heuristic Methods

Finding Unknown Patterns -> CNN

Source: folding.chmcc.org, Intro to BioInformatics, Professor Jarek Meller

Hochschule für Technik und Wirtschaft Berlin

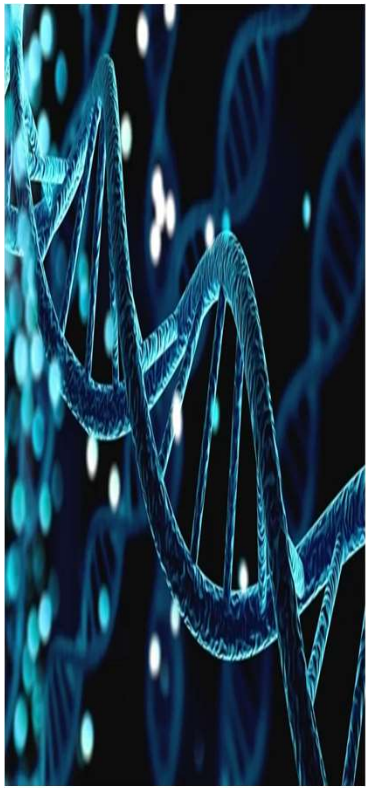
University of Applied Sciences

References

- Deep Modeling of DNA Sequences with Python and Keras, Martin Preusse, Gokcen Eraslan(Researchers at Helmholtz Munich)
- Intro to BioInformatics , Prof. Jarek Meller, Cincinnati Childrens Hospital Research(www.chmcc.org)
- Demystify DNA Sequencing with Machine Learning and Python, The AI Dream, Oct 26, 2020

Python Coding

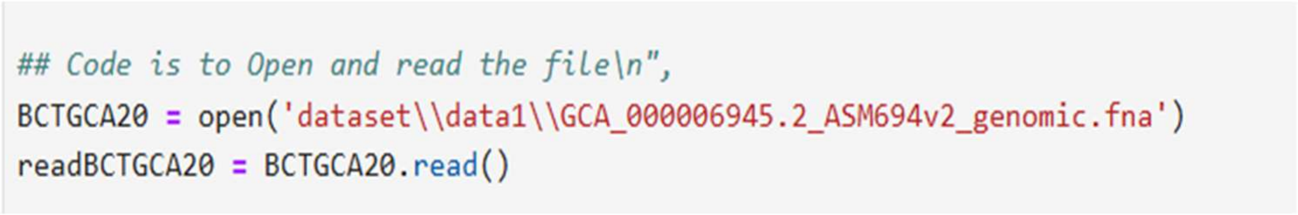
(Extra, if time permits)



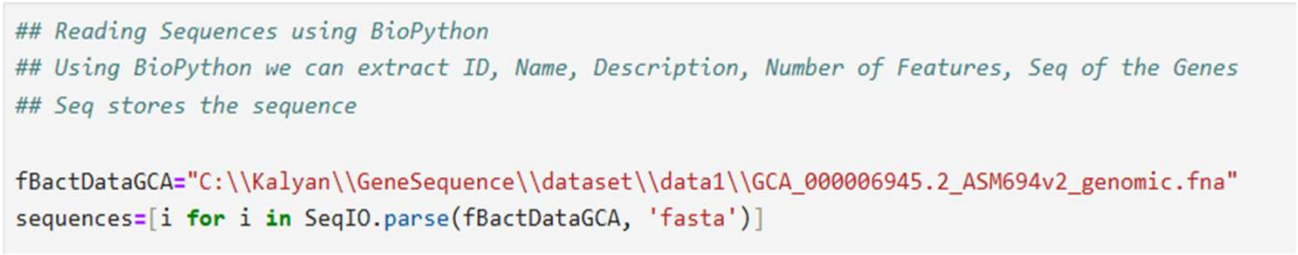
Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

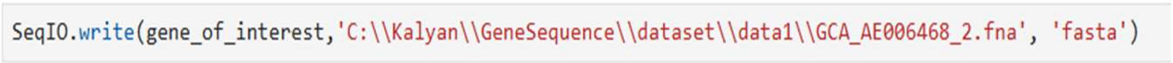
Processing Sequence Files – Python Code



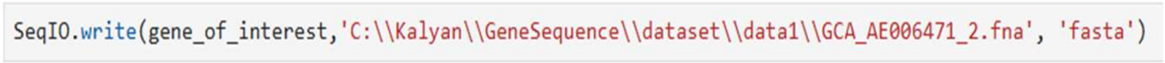
```
## Code is to Open and read the file\n",
BCTGCA20 = open('dataset\\data1\\GCA_000006945.2_ASM694v2_genomic.fna')
readBCTGCA20 = BCTGCA20.read()
```



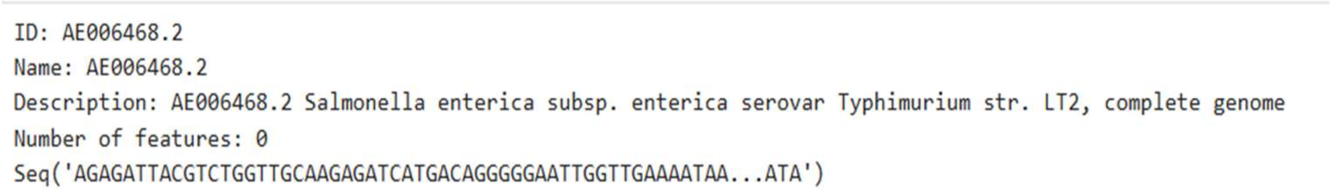
```
## Reading Sequences using BioPython
## Using BioPython we can extract ID, Name, Description, Number of Features, Seq of the Genes
## Seq stores the sequence

fBactDataGCA="C:\\Kalyan\\GeneSequence\\dataset\\data1\\GCA_000006945.2_ASM694v2_genomic.fna"
sequences=[i for i in SeqIO.parse(fBactDataGCA, 'fasta')]
```



```
SeqIO.write(gene_of_interest,'C:\\Kalyan\\GeneSequence\\dataset\\data1\\GCA_AE006468_2.fna', 'fasta')
```



```
SeqIO.write(gene_of_interest,'C:\\Kalyan\\GeneSequence\\dataset\\data1\\GCA_AE006471_2.fna', 'fasta')
```



```
ID: AE006468.2
Name: AE006468.2
Description: AE006468.2 Salmonella enterica subsp. enterica serovar Typhimurium str. LT2, complete genome
Number of features: 0
Seq('AGAGATTACGTCTGGTTGCAAGAGATCATGACAGGGGGAATTGGTTGAAAATAA...ATA')
```

Install BioPython, PyTorch, GFFUtils

Import BioPython Libraries

Read FASTA File

Parse FASTA File

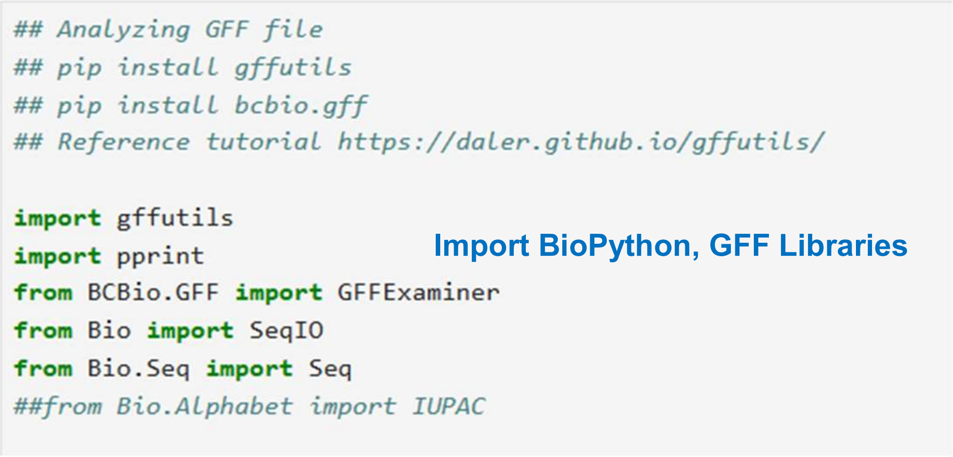
Identify Genes of Interest

Store them in Separate Fasta Files

Understand Sequence Data

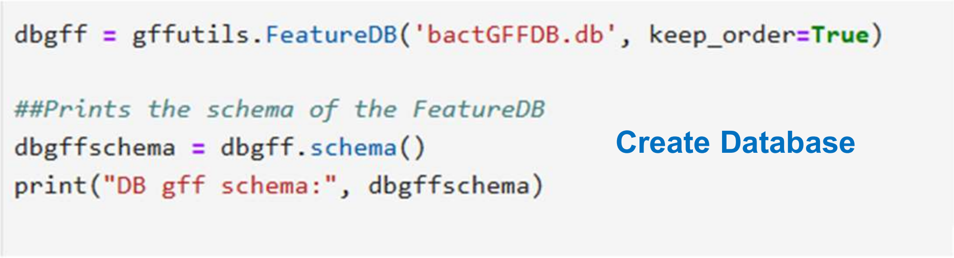
SeqID, No. of Genes, Length, Extra Characters, Spaces

Processing Features GFF – Python Code



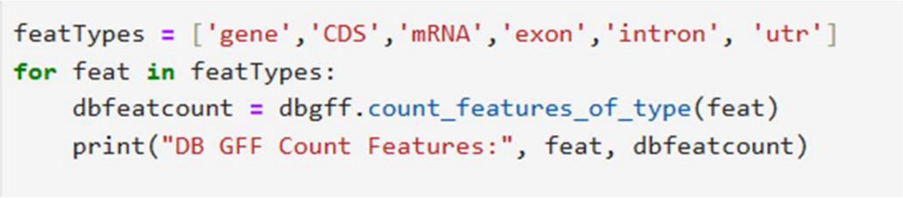
```
## Analyzing GFF file
## pip install gffutils
## pip install bcbio.gff
## Reference tutorial https://daler.github.io/gffutils/

import gffutils
import pprint
from BCBio.GFF import GFFExaminer
from Bio import SeqIO
from Bio.Seq import Seq
##from Bio.Alphabet import IUPAC
```

Import BioPython, GFF Libraries

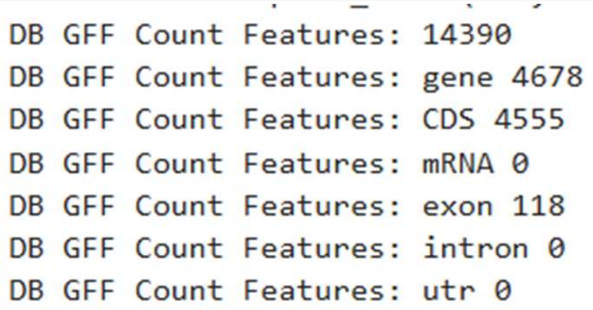


```
dbgff = gffutils.FeatureDB('bactGFFDB.db', keep_order=True)

##Prints the schema of the FeatureDB
dbgffschema = dbgff.schema()
print("DB gff schema:", dbgffschema)
```

Create Database



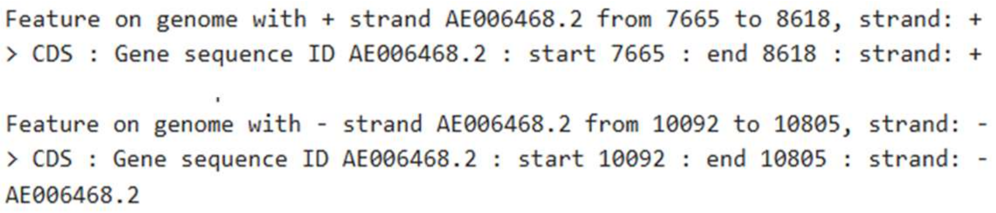
```
featTypes = ['gene','CDS','mRNA','exon','intron', 'utr']
for feat in featTypes:
    dbfeatcount = dbgff.count_features_of_type(feat)
    print("DB GFF Count Features:", feat, dbfeatcount)
```



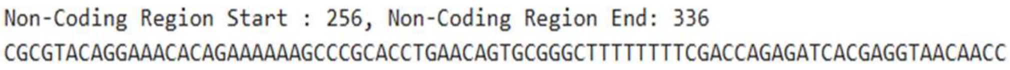
```
DB GFF Count Features: 14390
DB GFF Count Features: gene 4678
DB GFF Count Features: CDS 4555
DB GFF Count Features: mRNA 0
DB GFF Count Features: exon 118
DB GFF Count Features: intron 0
DB GFF Count Features: utr 0
```

Understand Features
CDS
Gene
Exon
Intron
mRNA

Process CDS and Non-CDS + and - strands



```
Feature on genome with + strand AE006468.2 from 7665 to 8618, strand: +
> CDS : Gene sequence ID AE006468.2 : start 7665 : end 8618 : strand: +

Feature on genome with - strand AE006468.2 from 10092 to 10805, strand: -
> CDS : Gene sequence ID AE006468.2 : start 10092 : end 10805 : strand: -
AE006468.2
```



```
Non-Coding Region Start : 256, Non-Coding Region End: 336
CGCGTACAGGAAACACAGAAAAAAGCCCGCACCTGAACAGTGCGGGCTTTTTTTTCGACCAGAGATCACGAGGTAACAACC
```

Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

htw

Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

www.htw-berlin.de