INSTITUTO DE PESQUISAS TECNOLÓGICAS

COMUNICAÇÃO TÉCNICA

Nº 178455

DNA data storage transmuting digital information into DNA: an innovative odyssey

Adriano Galindo Leal

Palestra apresentada no Palestra apresentada no InnScidSP, 24/07-04/08/2023. 54 slides.

Instituto de Pesquisas Tecnológicas do Estado de São Paulo S/A - IPT
Av. Prof. Almeida Prado, 532 | Cidade Universitária ou
Caixa Postal 0141 | CEP 01064-970
São Paulo | SP | Brasil | CEP 05508-901
Tel 11 3767 4374/4000 | Fax 11 3767-4099

www.ipt.br

INSTITUTO DE PESQUISAS TECNOLÓGICAS

DNA Data Storage

Transmuting Digital Information into DNA: An Innovative Odyssey

Dr. Adriano Leal

Senior Researcher

at IPT's Digital Technologies Business Unit

Artificial Intelligence and Analytics Section

SÃO PAULO
GOVERNO DO ESTADO

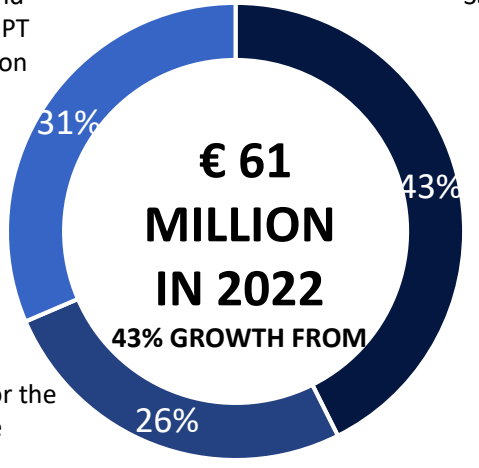
WHO ARE WE ?

IPT PROVIDES TECHNICAL SOLUTIONS FOR INDUSTRY, GOVERNMENTS AND SOCIETY, ENABLING THEM TO OVERCOME THE CHALLENGES OF OUR TIME

INCOMES

Sale of projects and services through IPT Support Foundation (FIPT)

São Paulo State Government basic funding



Sale of projects for the public and private sectors.

OUR NUMBERS (2022)

- **123 YEARS OF CONTRIBUTIONS TO SOCIETY**
- **> 1,000 EMPLOYEES AND PARTNERS**
- **41% REVENUE IN INNOVATION PROJECTS**
- **> 1,830 CUSTOMERS SERVED**
- **SATISFIED CUSTOMERS NPS 84 (LEVEL OF EXCELLENCE)**
- **> 19,900 TECHNICAL DOCUMENTS ISSUED**
- **> 2,000 TESTING AND ANALYSIS PROCEDURES IN THE PORTFOLIO**

WHAT WE DO ?

RESEARCH, DEVELOPMENT AND INNOVATION

- PRODUCTS AND PROCESSES
- SOFTWARES
- FROM THE BENCH TO THE PILOT
- FUNDING
- EMBRAPII

TESTS, TRIALS AND ANALYSIS

- TECHNICAL ANALYSIS OF PRODUCTS AND MATERIALS
- PRODUCT EVALUATION
- PRODUCT CERTIFICATION

INSPECTION AND MONITORING

- CONSTRUCTION AND SCTRUCTURES
- MACHINERY AND EQUIPMENT
- ACCREDITED INSPECTION BODY

METROLOGICAL DEVELOPMENT, MEASUREMENTS AND CALIBRATIONS

- PROFICIENCY PROGRAMS
- STANDARDS DEVELOPMENT
- ADVANCED METROLOGY

CERTIFIED REFERENCE MATERIALS

- METALS
- CERAMIC
- MINERAL
- VISCOSITY
- NORMAL SAND

TECNOLOGICAL EDUCATION

- PROFESSIONAL MASTER
- EXTENSION COURSES
- COURSES ON DEMAND

BUSINESS UNITS

BIONANOMANUFACTURING
Processes, Chemistry, PPEs, Biotech, Nanotech, Microfabrication

CITIES, INFRASTRUCTURE AND ENVIRONMENT
Territorial planning, Sustainability, Risks, Civil works

ENERGY
Generation, Infrastructure, Efficiency, Clean energy

BUILDING AND HOUSING
Confort, Performance, Safety, Materials, Sustainability

ADVANCED MATERIALS
Metallic, Polymeric, Composite, Cellulosic, Corrosion

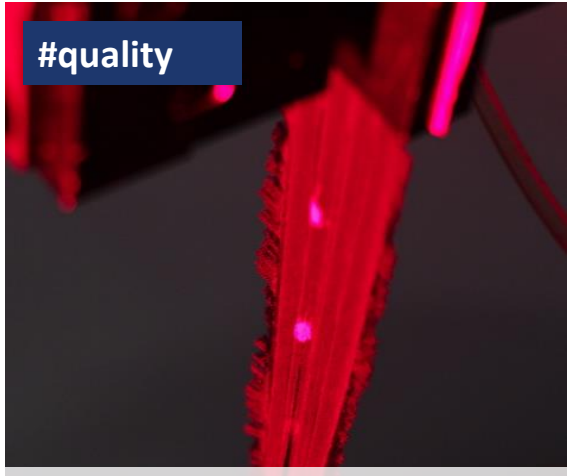
DIGITAL TECHNOLOGIES
IoT, Embedded Systems, Intelligent Transport Systems, AI, Analytics

METROLOGICAL AND REGULATORY TECHNOLOGIES
Mechanics, Electrical, Flow Measurement, Aerodynamics, Chemistry

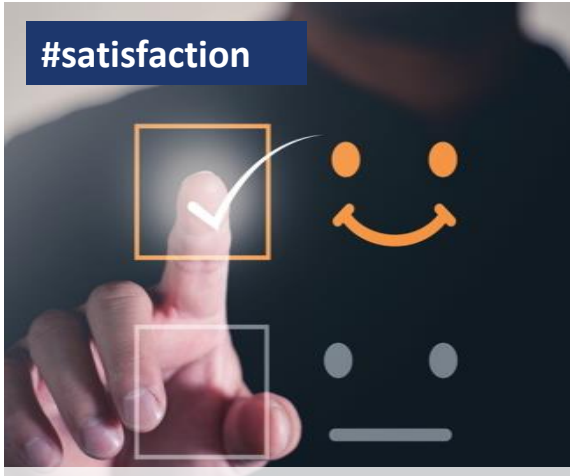
ipt
SÃO PAULO
GOVERNO DO ESTADO

IPT´S HALLMARKS

+ 120,000 square meters of laboratories
+ 1,000 qualified professionals
Countless ways to innovate



+ 2,000 tests and calibrations
+ 20,000 technical documents per year
Reference in quality services



Level of Excellence in NPS
NPS 84
(Net Promoter Score)

Open Experience
IPT

O IPT opens it campus to the largest open innovation action in hardtech in Brazil, connecting distinct sotckholders of this ecosystem.

Cornerstone of the CITI Project – São Paulo State International Technology and Innovation Center.

Mode 1

Innovation hub

Become part of in a unique and transformative ecosystem that brings together companies and starups that undertake together in the creation of technologies that drive new businesses.

Mode 2

Innovation center

Install your company´s Technology Center within the IPT campus and leverage your development capacity.

ipt SÃO PAULO GOVERNO DO ESTADO

PROF. DR. ADRIANO LEAL

Adriano Leal (IEEE Member 2006) received a B.Sc. degree in electrical engineering and an M.Sc. and Ph.D. from Polytechnic School at the University of São Paulo in 1996, 1999, and 2006, respectively.

For 11 years, he worked as an R & D engineer for the GAGTD research group in the Polytechnic School at the University of São Paulo. He was responsible for studying and developing automation and information systems for generation, transmission, and electricity distribution.

From April 2007 until November 2010, he has been a Researcher at Elucid Solutions, a consulting and TI company for several utility companies in Brazil.

Since December 2010, he has been a Senior Research Engineer at IPT – Institute for Technological Research. He was IEEE PES South Brazil Chapter President from 01/2015 until 04/2019.

From 05/2019 until 01/2020, he was Visiting Researcher at Technische Universität Berlin / TUB, at the Network Information Theory Group, under Dr. Renato Luís Garrido Cavalcante. Research and training activities were carried out on machine learning techniques to process data from the Internet of Things that comprise the Cyber-Physical Ecosystem of Smart Cities.

On his return, with the support of IPT's top management, he set up the artificial intelligence and analytics section, training researchers on various AI projects and helping assemble the IASMIN Platform.

Let's take flight on our mental paraglides and soar over the project's landscape.

EMBARKING ON AN AERIAL ODYSSEY: LET'S BEGIN...

Why DNA Data Storage

- DNA can be stable for thousands of years

- Potential to store Petabytes of data into 1 gram of DNA

Article | Open Access | Published: 26 May 2022

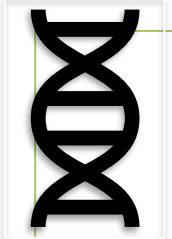
Bioarchaeological and palaeogenomic portrait of two Pompeians that died during the eruption of Vesuvius in 79 AD

Gabriele Scorrano, Serena Viva, Thomaz Pinotti, Pier Francesco Fabbri, Olga Rickards & Fabio Macciardi

Scientific Reports **12**, Article number: 6468 (2022) | Cite this article

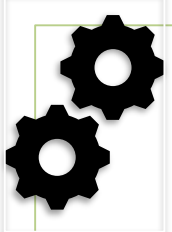
- DNA reading will never be outdated and will always be improved

Scope

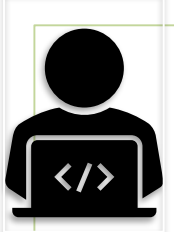


Enzymatic synthesis

Chemical synthesis



Process miniaturization

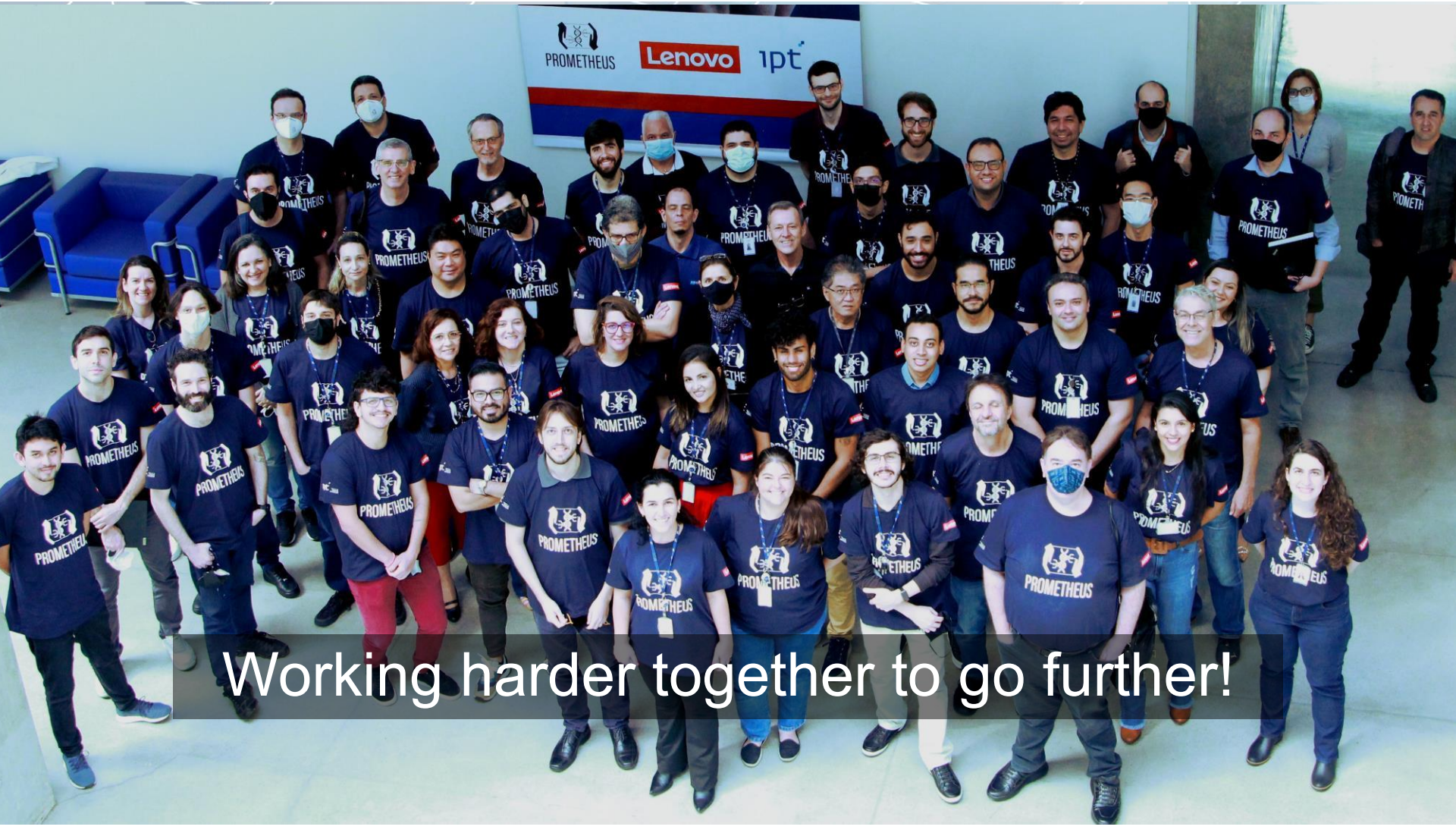


Coding and decoding

Sequencing methods

DNA synthesis

	Enzymatic Synthesis	Chemical Synthesis
Benefits	▪ Fast ▪ Environmentally Friendly	▪ Synthesis Control ▪ Standardized Technique
Challenges	▪ Synthesis Control	▪ Hazardous chemicals

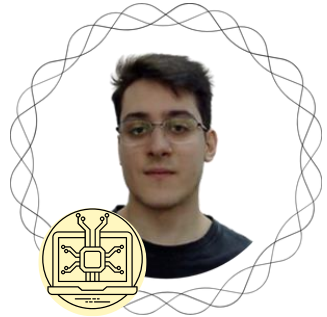

PROMETHEUS
Lenovo
ipt
Working harder together to go further!

The CODEC Team

Diego

Alessandro

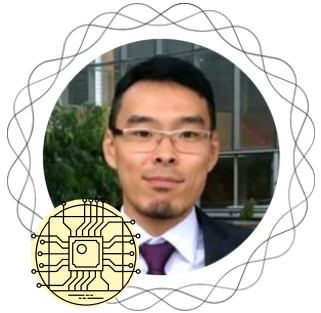
Adriano



Matheus

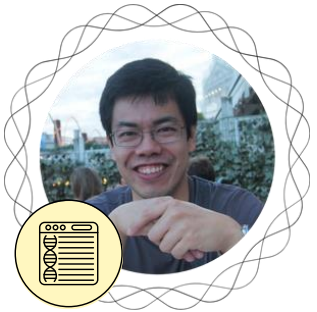
Maria Cristina

André Martins



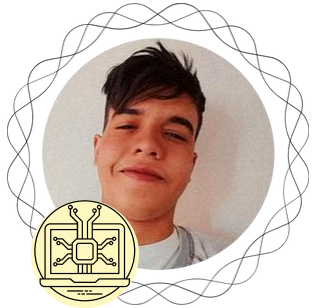
Thiago

Takeo

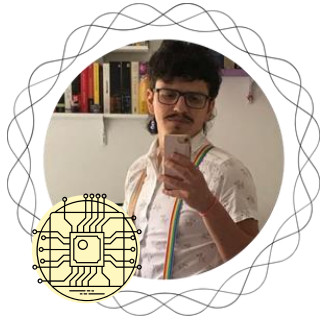


André Fujita

Cristina Maria



Leonardo



Allan

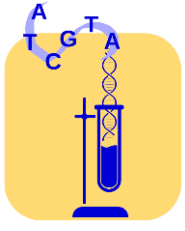
Nyanko

DNA data storage pipeline

1. Converting information to binary code

2. Converting binary code to DNA code

3. DNA Synthesis



4. Storage

5. Recovery

6. DNA Sequencing

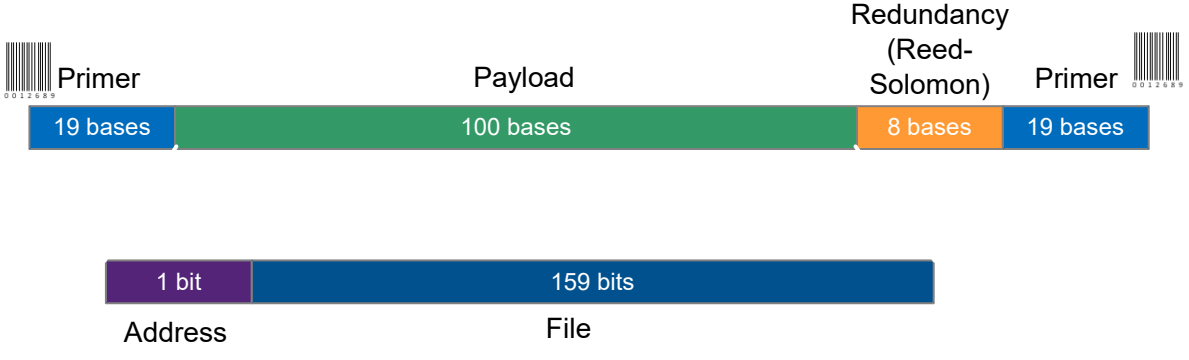
7. Decoding

Our First results

38byte .txt file containing:
IPT and Lenovo, a successful partnership!

encoded to

2 DNA sequences with 146 bases each



Bases = nitrogeneous bases
= A T C G

 |

DNA sequences containing encoded data

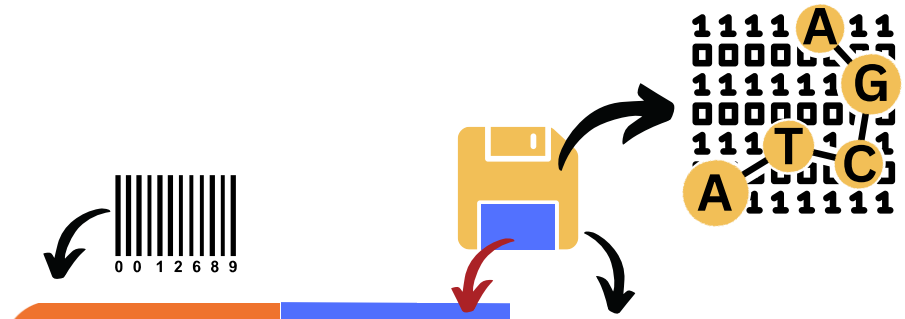
Forward Primer

5'CACGACGTTGTAAAACGACAGACAGGAGAAGCGTACTATATAAGGCCACAGA
CGATAAGGTGCTATCCGGTAGCATGCTGCACGACTATATCGTGTACGGTCACGC
TATATCGCATCACGGGACGCCGGGTCATAGCTGTTTCCTG

5'CACGACGTTGTAAAACGACTCGTGGCAGATCAGTCCATAGCCGTCCAGACAA
GAACAGTACGGCCAAGAACATATCGTCCCAGATCCGACCATATCCTCTCTGATA
CGCCATATAACACAGAGGGATGGGTCATAGCTGTTTCCTG

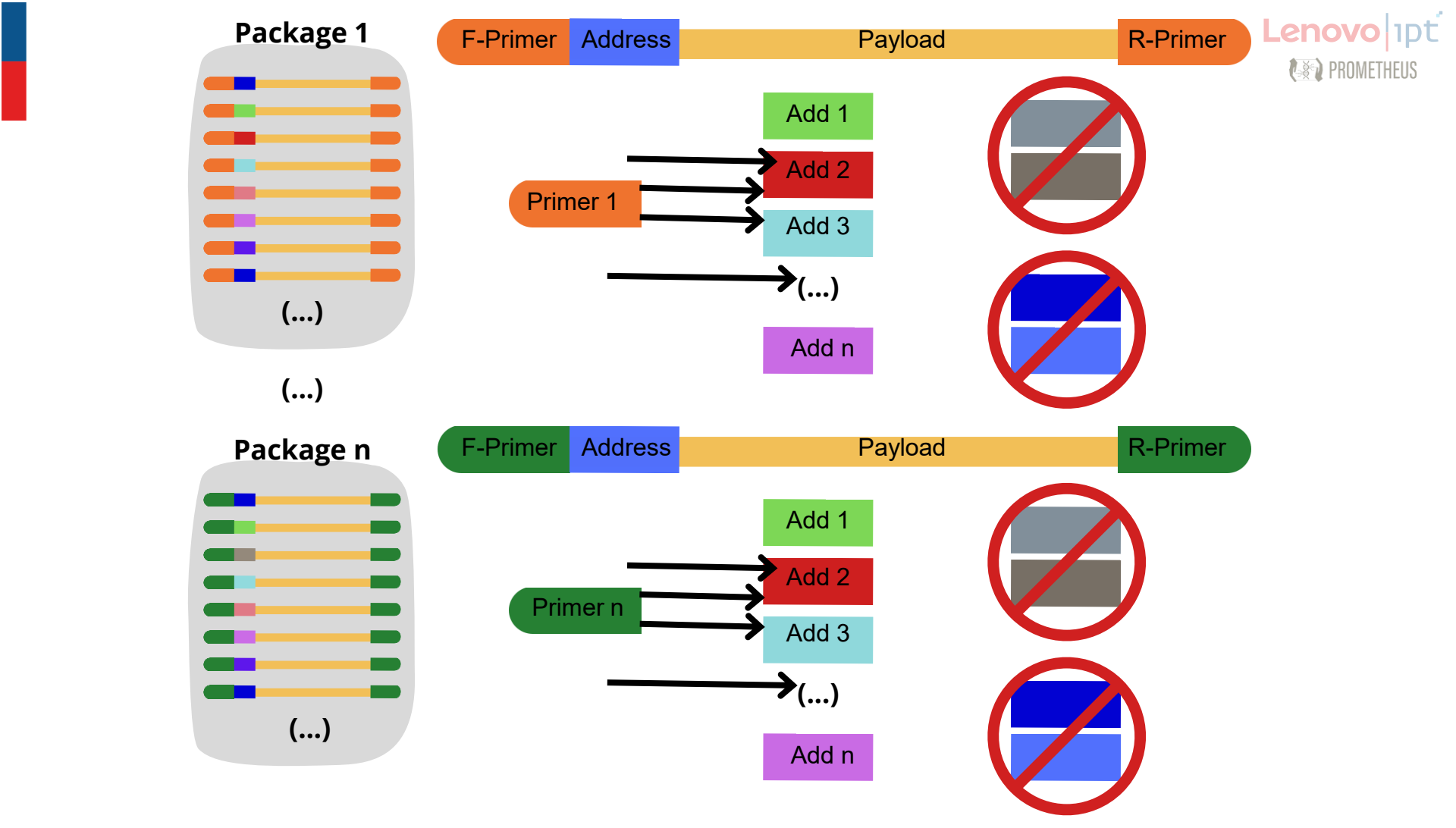
Reverse Primer

Now, more complex..



F-Primer | Address | Payload | R-Primer

5'-(...)TACGCCGCAACCTGCCATCCCAAGCTAAAAGGGGAATCCCAATGGCAATCTGCACCGTCGGACGA(...)-3'


Package 1
(...)
(...)
Package n
(...)
F-Primer
Address
Payload
R-Primer
Add 1
Add 2
Add 3
(...)
Add n
Primer 1
F-Primer
Address
Payload
R-Primer
Add 1
Add 2
Add 3
(...)
Add n
Primer n

DNA DATA STORAGE ALLIANCE

A SNIA TECHNOLOGY AFFILIATE

Lenovo | ipt

PROMETHEUS

Technical Working Group

DNA Archive Rosetta Stone - DARS

"The goal of our group is to define a simple and minimal reserved area in the archive that will function like a “rosetta-stone” and will give the reader/user an idea about how the rest of the archive is encoded."

DNA ZONE

Alliance Codec/Custom

Free text

<Sector 1>
<UID>
<Description>
<Codec Params>
<Compression>
<Encryption>
<Map File Primers>
<Oligo Size>
...

<Timestamp>
<Vendor Details>
<Hash>

Proprietary Codec

<Bucket 1>
<Primers>
<Contents>
<Desc>

<Bucket 2>
<Primers>
<Contents>
...

DNA Map file

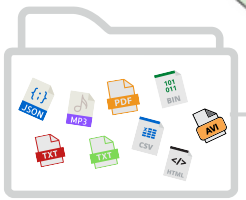
File #1
File #2

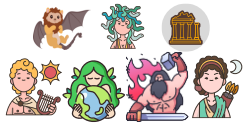
Bucket 1

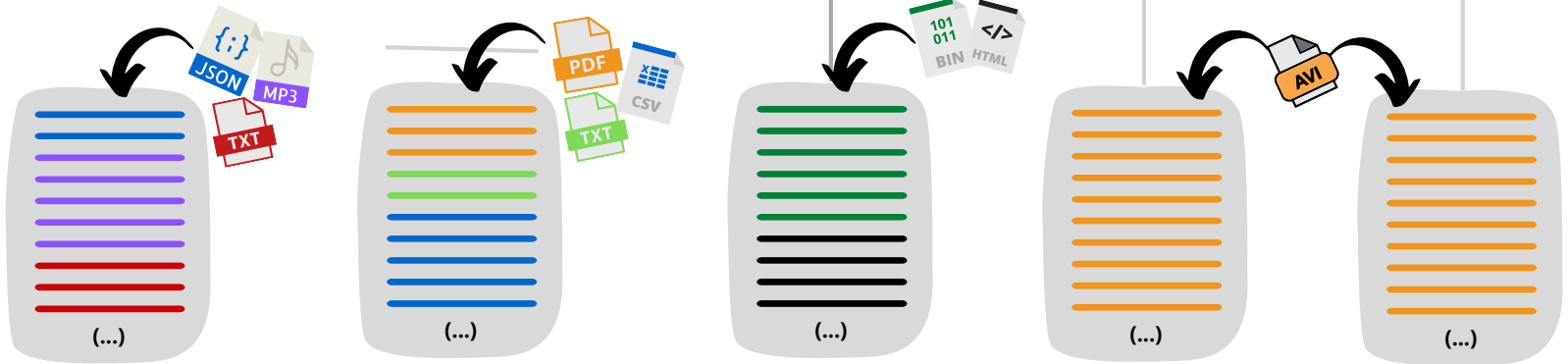
File #3

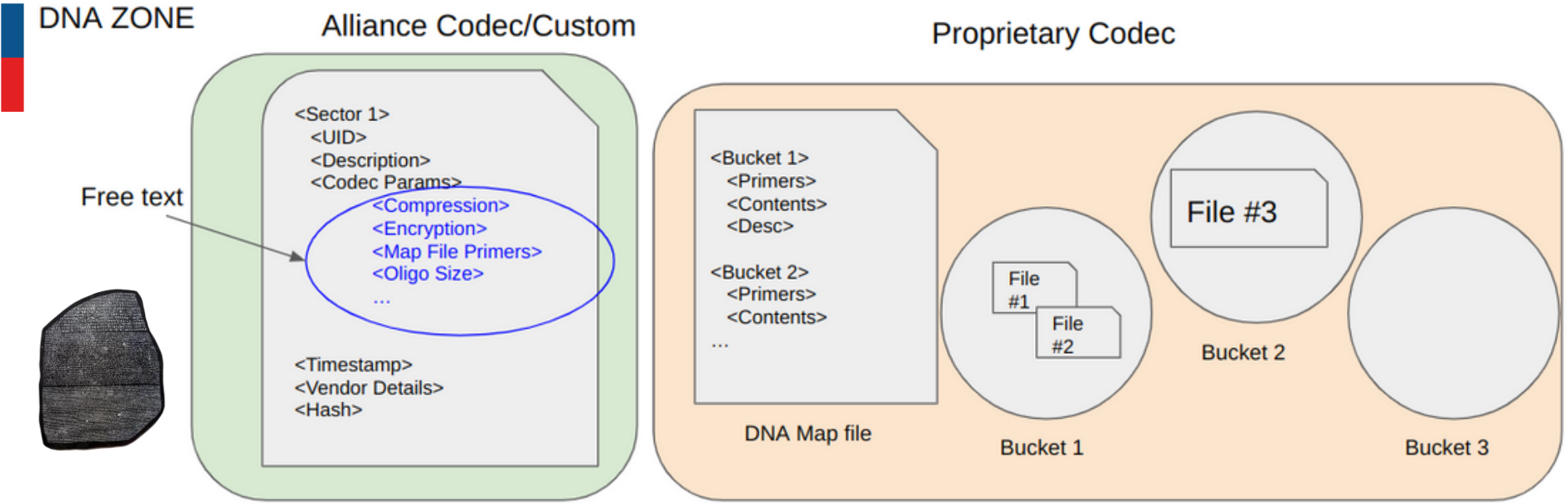
Bucket 2

Bucket 3





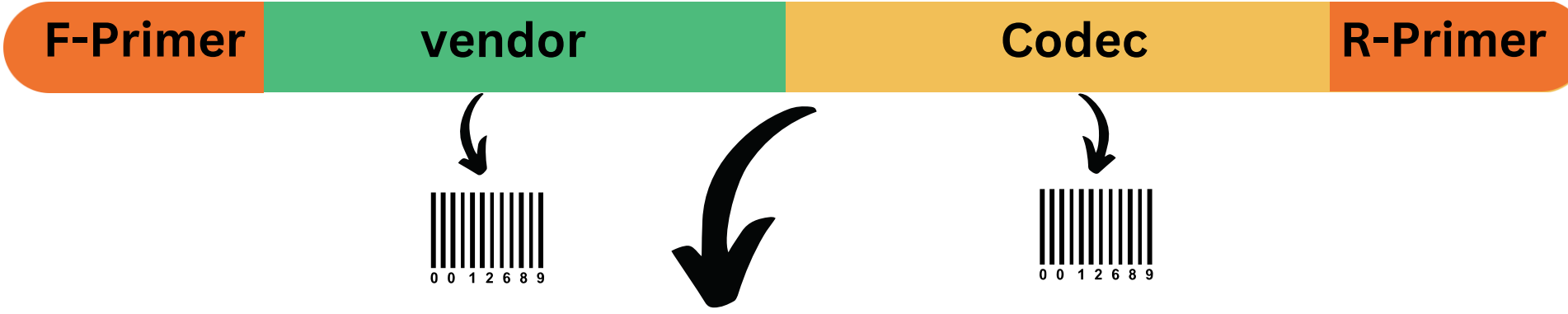




Sector 0: DNA origin - Company name; codec used

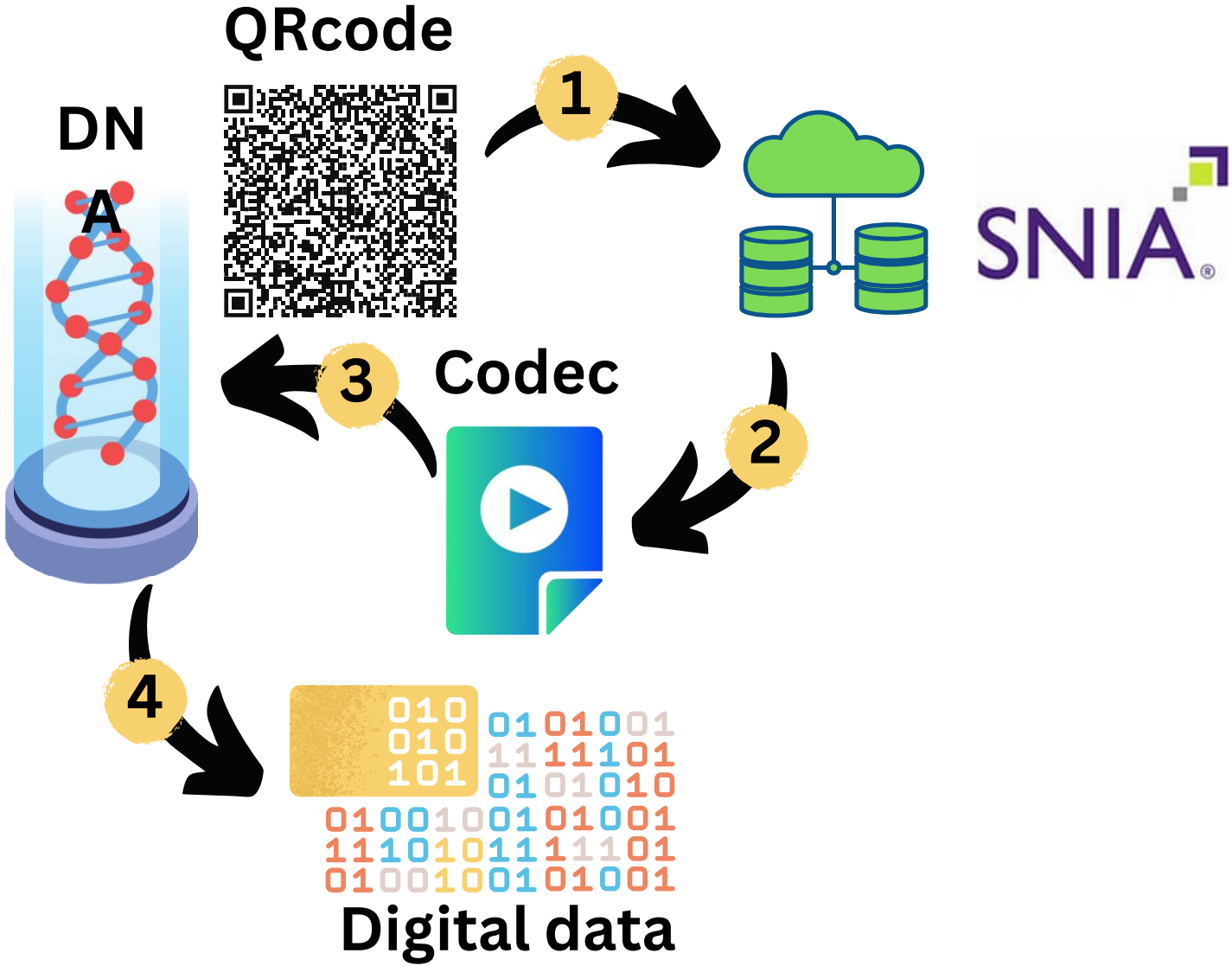
Sector 1 (S1): File Metadata

Sector 0: Company name; codec used



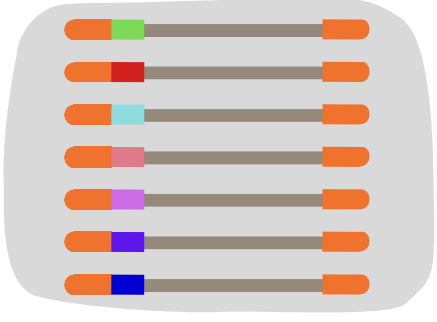
ID	Vendor Name	Contact Details	Notes
ATGCGATGCGATGCGATGC GATGCGATGCGATGCG	Twist Bioscience	twistbioscience.com/get_codec +14155307827	

ID	Vendor Name	Codec Details	Notes
ATGCGATGCGATGCGATGC GATGCGATGCGATGCG		Version 1.4	



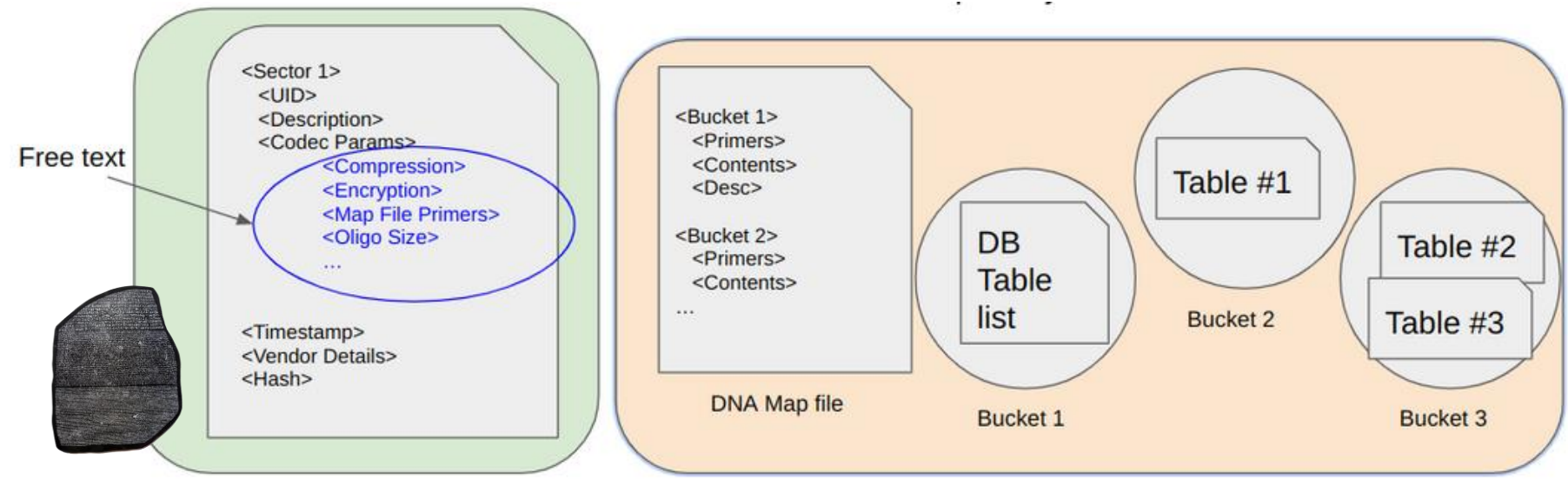
Lenovo | ipt
PROMETHEUS

Sector 1 (S1): File Metadata

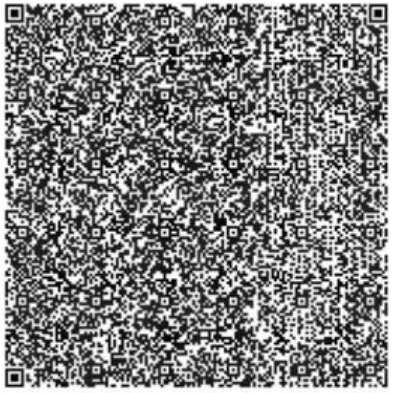


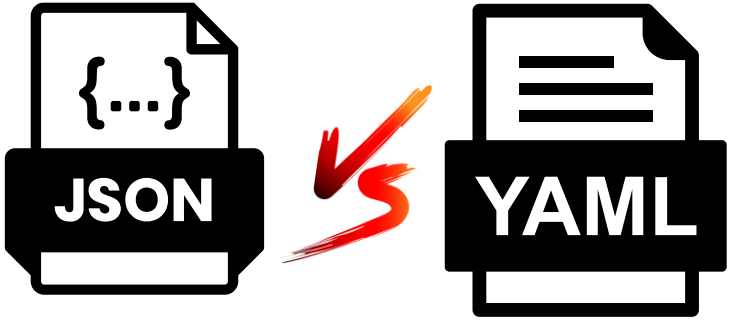
Emergency rescue

- File data
  - List os files
  - file sizes
  - Creation date
  - MD5 Checksum
  - User owner
  - File location (package)
- CODEC
  - Encoding parameters
  - Primer data
  - Number of oligos



Sector 1 (S1): File Metadata





QR codes can hold a maximum of 3KB of data (177x177 nodes) according to https://dhiway.com/a-definitive-guide-to-qr-code/#:~:text=A%20QR%20code%20has%20a,characters%20or%204%2C269%20alphanumeric%20ones.

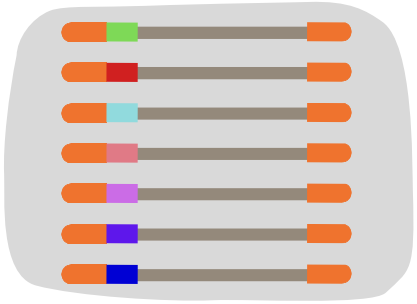
codec.json

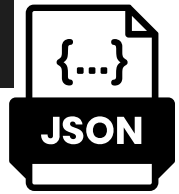
C: > Users > thiagoaoyagi > Desktop > codec.json > ...

```
{
  "address_method": "baselevel_ordered",
  "inner_ecc_method": "ReedSolomon",
  "interleaving": true,
  "mapped_oligo_size": 108,
  "mapping_method": "direct",
  "max_block_size": 5100,
  "outer_ecc_method": "ReedSolomon",
  "primer_cut": 3,
  "primer_list": [
    "CACGACGTTGTAGGACGAC",
    "CAGGGAGCAGCTATGACCC",
    "GGTTGGCCCAGTCACGACG",
    "GGAGGCAGCTATGACCATG",
    "TGTAAAACGACGGCCAGTG",
    "AGCGGATAACAATGGCACC",
    "TCCGCCGACTCACTATAGG",
    "ATGTCGGGCTAACCCCGCC",
    "GCTAGTTATTGCTCAGCGG",
    "TAGTTATTGCTCAGCGGTG"
  ],
  "rand_type": "None",
  "segment_size": 216,
  "xor_kernel": "Xbf1F2"
}
```

Package 0

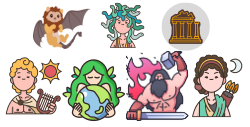
Metadata



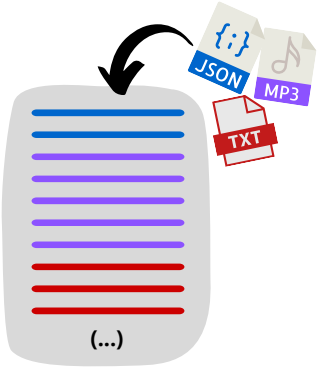


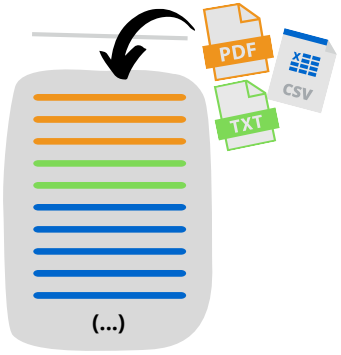
Emergency rescue

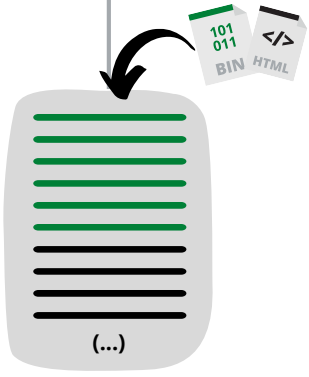
- File data
  - List os files
  - file sizes
  - Creation date
  - MD5 Checksum
  - User owner
  - File location (package)
- CODEC
  - Encoding parameters
  - Primer data
  - Number of oligos

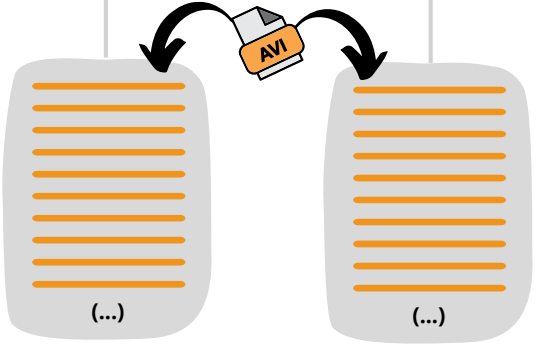


Prometheus' Codec v0.3

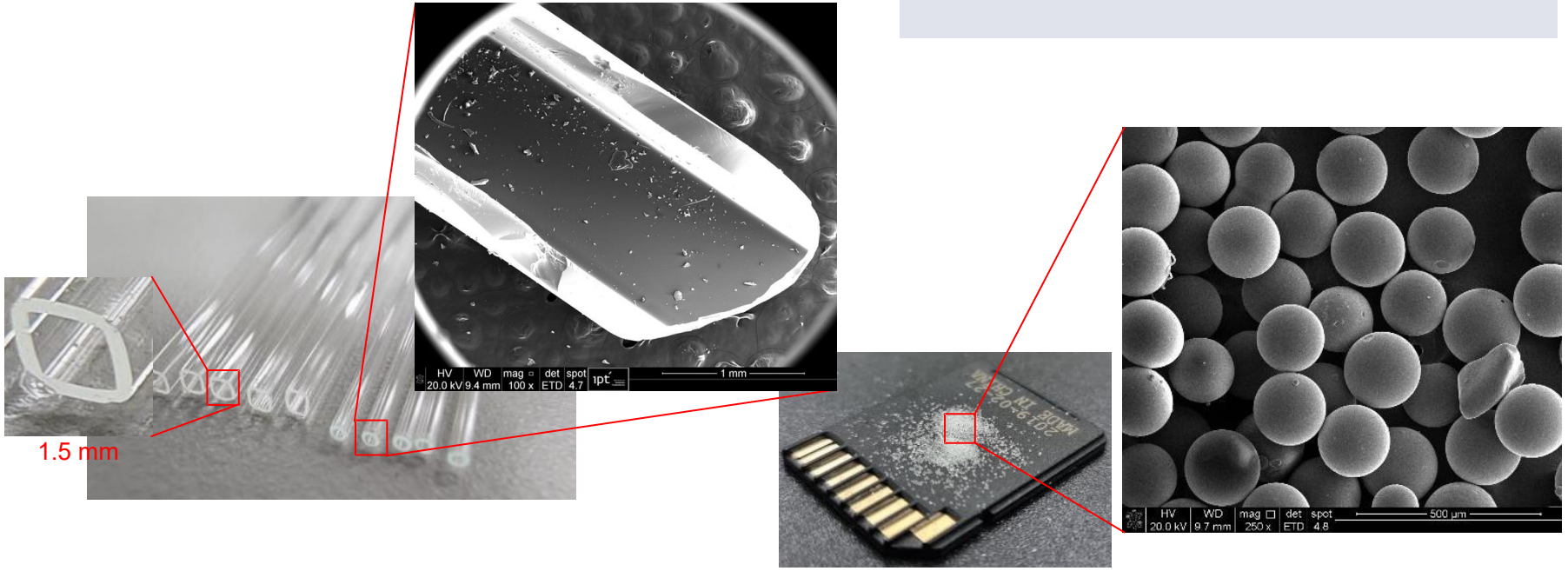




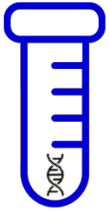




Chemical DNA synthesis on different surfaces



DNA storage



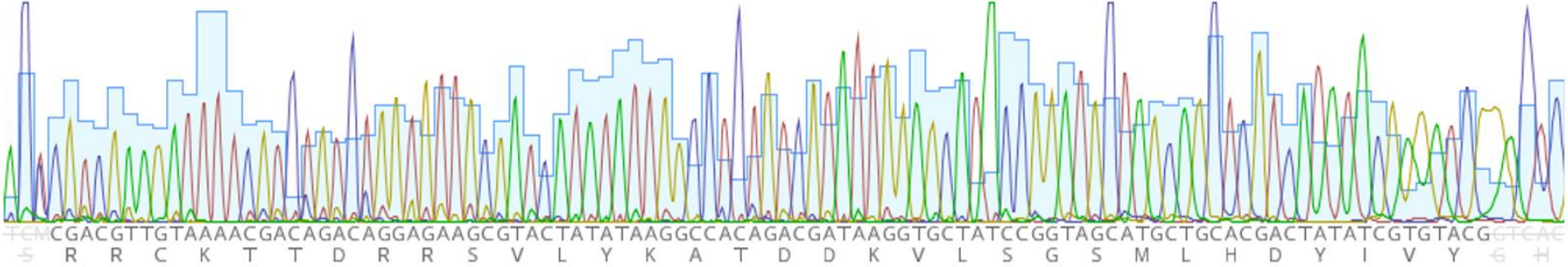
Creating copies by PCR methodology

PCR = Polymerase Chain Reactions

Surface storage with DNA

Sequencing

Sequencing of synthesized DNA



5'CACGACGTTGTAAAACGACAGACAGGAGAAGCGTACTATATAAGGCCACAGACGATAAGGTGCT
ATCCGGTAGCATGCTGCACGACTATATCGTGTACGGTCACGCTATATCGCATCACGGGACGCCGG
GTCATAGCTGTTTCCTG

5'CACGACGTTGTAAAACGACTCGTGGCAGATCAGTCCATAGCCGTCCAGACAAGAACAGTACGG
CCAAGAACATATCGTCCCAGATCCGACCATATCCTCTCTGATACGCCATATAACACAGAGGGATGG
GTCATAGCTGTTTCCTG

Alignment and consensus → Detach of primers → Check of errors via logical redundancy (Reed-Solomon) →

```
primer        |-----------------|                                                                                                                   |-----------------|
read1        TC*CGACGTTGTAAAACGACAGACAGGAGAAGCGTACTATATAAGGCCACAGACGATAAGGTGCTATCCGGTAGCATGCTGCACGACTATATCGTGTACGGTCACGC*--**C*---*ACG---------------------------
read2        ----------------------------G*A--********TAAGGC-*CAGACGATAAGGTGCTATCCGGTAGCATGCTGCACGACTATATCGTGTACGGTCACSCTATATCGCATCACGGGACGCCGGGTCATAGCTGTTTC**GA
consensus     C*CGACGTTGTAAAACGACAGACAGGAGAAGCGTACTATATAAGGCCACAGACGATAAGGTGCTATCCGGTAGCATGCTGCACGACTATATCGTGTACGGTCACGCTATATCGCATCACGGGACGCCGGGTCATAGCTGTTTC**G
original      CACGACGTTGTAAAACGACAGACAGGAGAAGCGTACTATATAAGGCCACAGACGATAAGGTGCTATCCGGTAGCATGCTGCACGACTATATCGTGTACGGTCACGCTATATCGCATCACGGGACGCCGGGTCATAGCTGTTTCCTG
errors         x                                                                                                                                                  xx
```

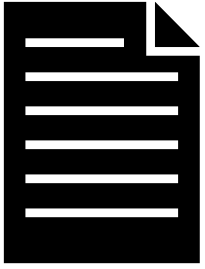
```
primer        |-----------------|                                                                                                                   |-----------------|
read1        TC*CGACGTTGTAAAACGACTCGTGGCAGATCAGTCCATAGCCGTCCAGACAAGAACAGTACGGCCAAGAACATATCGTCCCAGATCCGACCAT*-CCTCTC-*A**-GC******A-----------------------------
read2        -----------------------------****G----T---CGT-CAGAC-AGAACAGTACGGCCAAGAACATATCGTCCCAGATCCGACCATATCCTCTCTGATACGCCATATAACACA*AGGGATGGGTCATAG*TGTTTC**GA
consensus     C*CGACGTTGTAAAACGACTCGTGGCAGATCAGTCCATAGCCGTCCAGACAAGAACAGTACGGCCAAGAACATATCGTCCCAGATCCGACCATATCCTCTCTGATACGCCATATAACACA*AGGGATGGGTCATAG*TGTTTC**G
original      CACGACGTTGTAAAACGACTCGTGGCAGATCAGTCCATAGCCGTCCAGACAAGAACAGTACGGCCAAGAACATATCGTCCCAGATCCGACCATATCCTCTCTGATACGCCATATAACACAGAGGGATGGGTCATAGCTGTTTCCTG
errors         x                                                                                                                  x               x       xx
```

 |

Lenovo | ipt
PROMETHEUS

Fasta

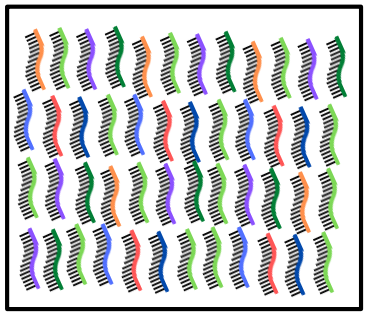
Fastq

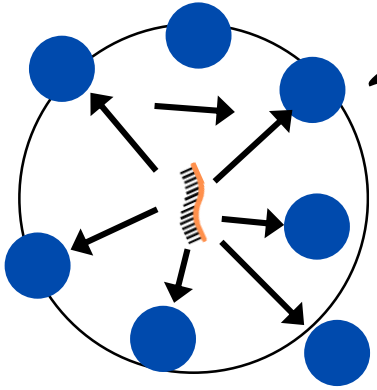


Analogy!

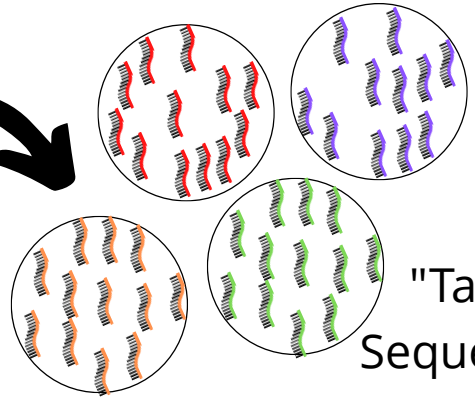
Review of Post-Sequencing Processing

Fastq
(reads)



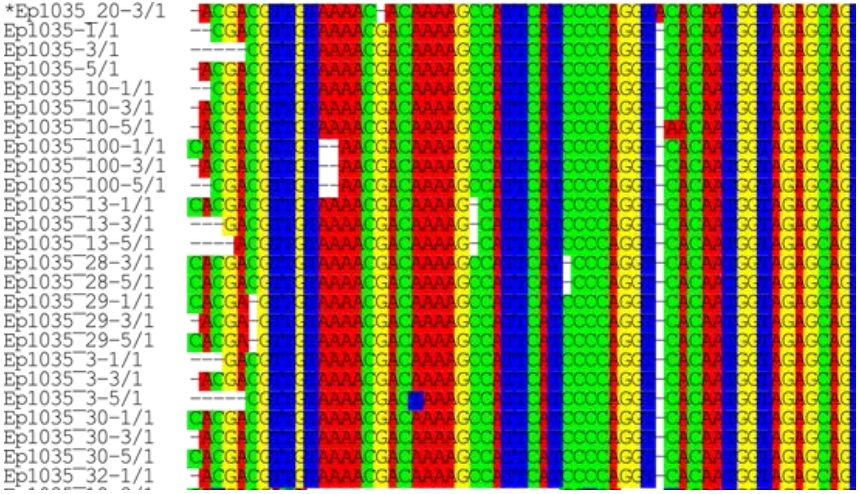


Clustering



"Target
Sequences"

Alignment



Lenovo | ipt
PROMETHEUS

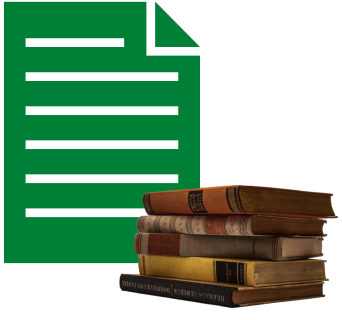
Fastq

>>>

CLUSTERING
ALIGNMENT
CONSENSUS

>>>

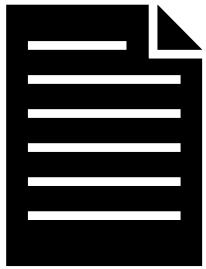
Fasta



Lenovo | ipt
PROMETHEUS

Pre-processing sequenced data

Fastq



Sequence clustering

Fastq

Let's check the pile for the page 15

Analogy!

Pre-processing sequenced data

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam finibus vitae ___ ac blan?it. Cras eget bibendum tortor, scelerisque pharetra leo. Suspendisse vestibulum odio sapien, id ultrices nibh ultricies finibus. Integer fermentum mauris et ultrices commoda.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam finibus vitae elit ac biandit. Cras eget bibendum rortor, scelerisque p_aretra leo. Suspendisse vestibulum odio sapien, id ultrices nibh ultricies finibus. Integer fermentum mauris ut ultrices commodo.

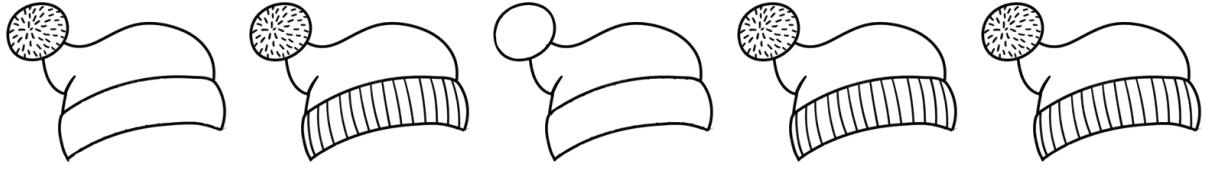
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam finibus vitae elit ac blandit. Cras eget bibendum tortor, scelerisque pharetra l??. Suspendisse vestibulum odio sapien, id ultrices nibh ultricies finibus. Integer fermentum mauris at ultrices commodu.

- Identify similar pages
- group them
- make a consensus
- repeat n times

Analogy!

Lenovo | ipt
PROMETHEUS

Which one is the correct version of Santa Claus?



Alignment and consensus

For DNA storage, the challenge is similar, find and group the right sequence and then, get consensus, but repeating these millions of times

If any error remains in the consensus, Artemis will attempt to fix it using error correction methods such as Reed-Solomon or Hamming code.

Fasta

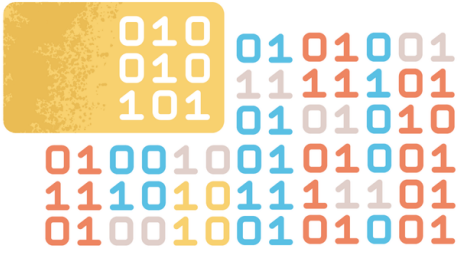
>>> DECODING >>>

Binary

Inverse mapping back to bit sequence

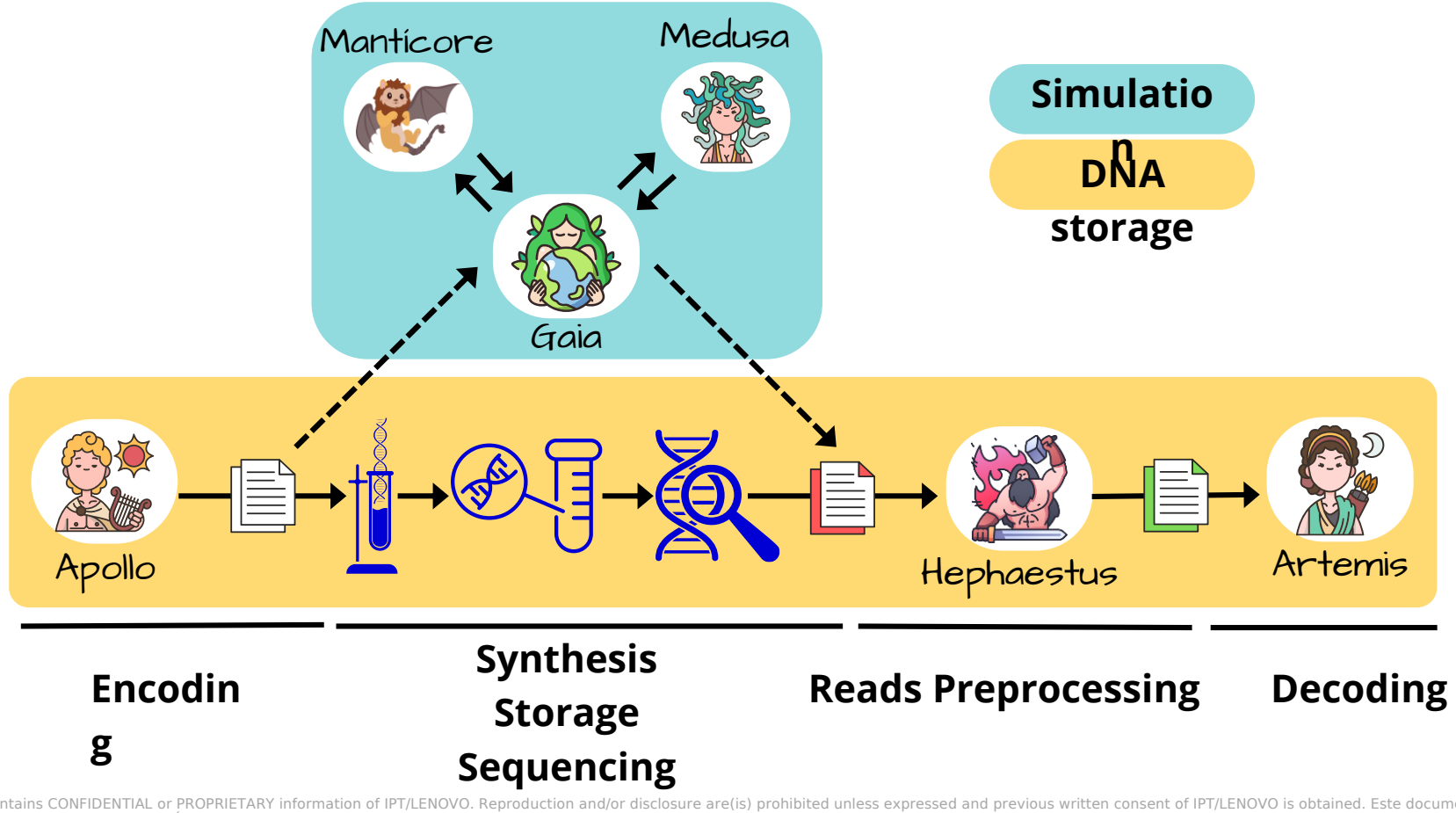
File reassembling according to segment addresses

IPT and Lenovo, a successful partnership!

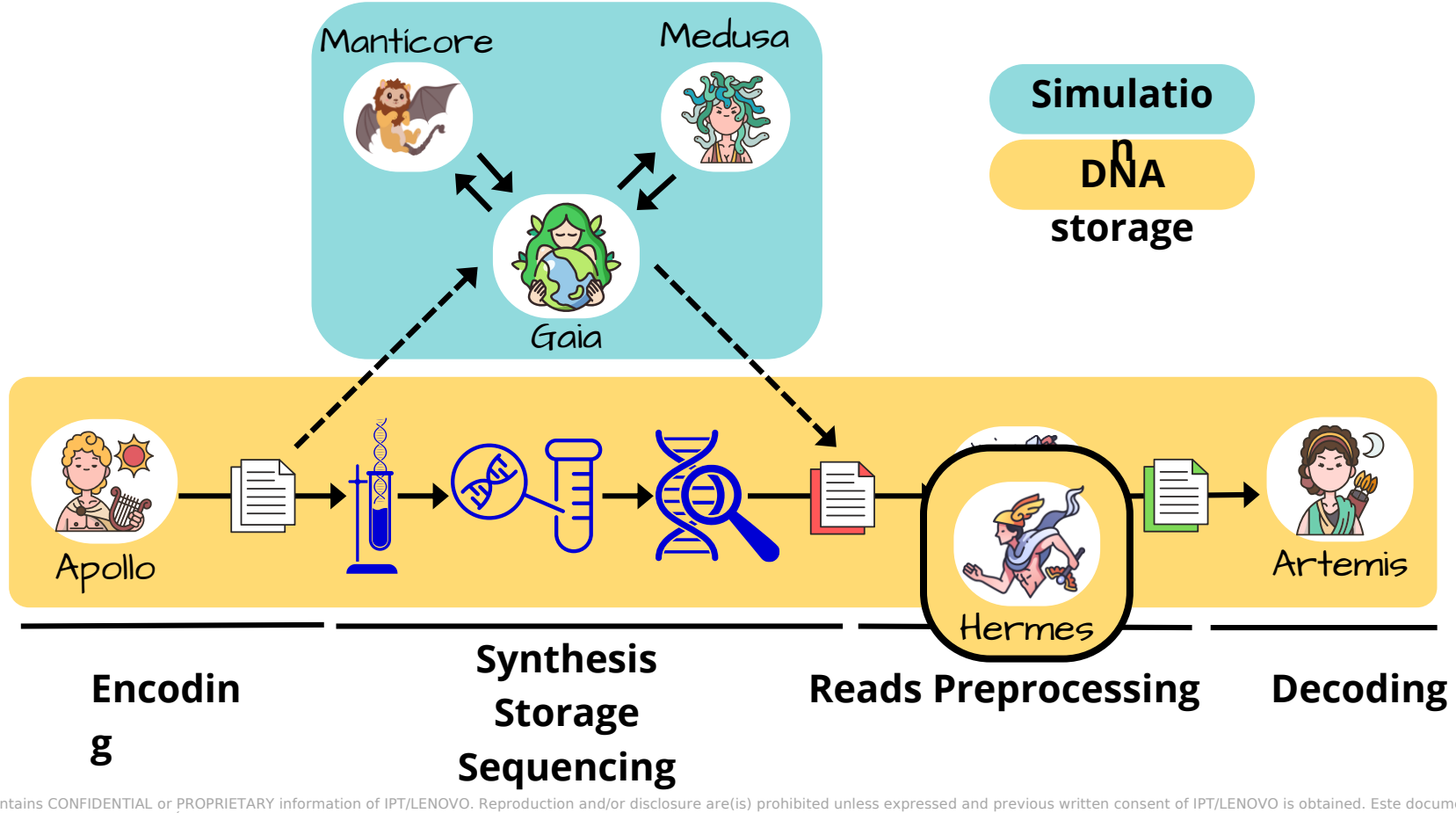


But the Process was Slow and we Needed to Sophisticate...

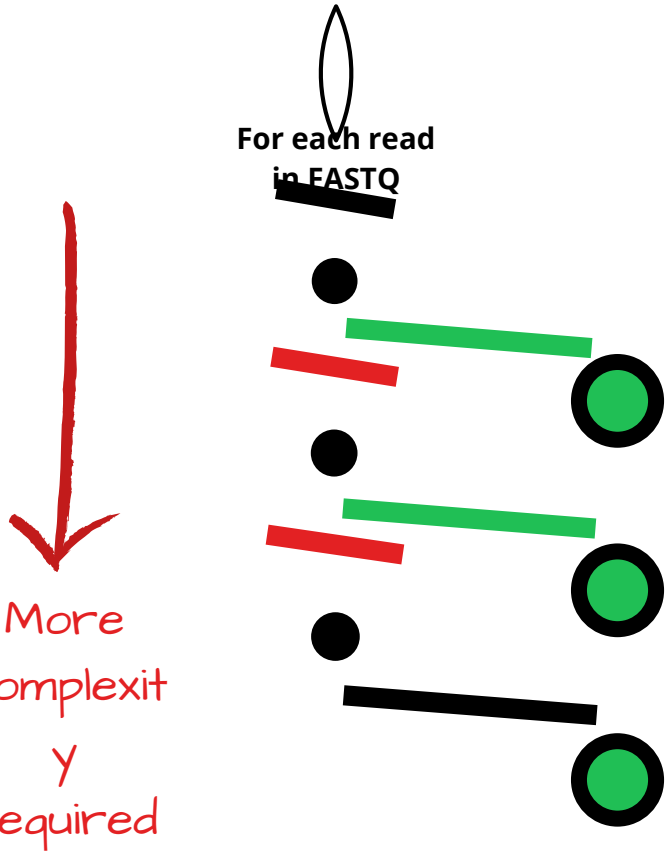
CODEC v0.4 Update



CODEC v0.4 Update



Post-sequencing processing pipeline (under discussions)



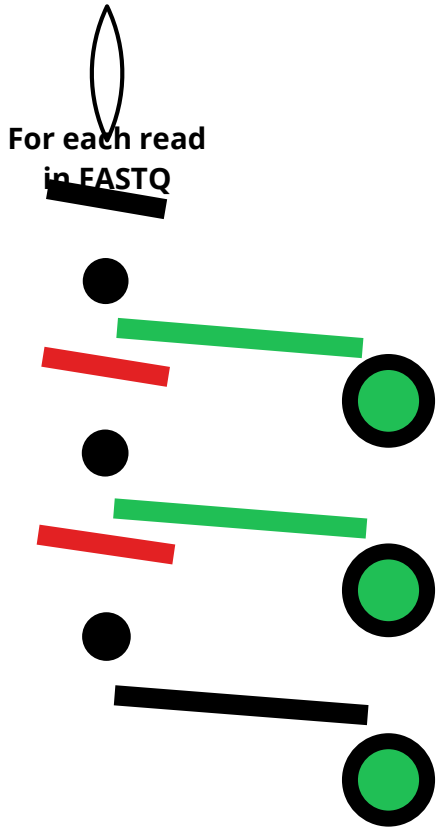
Sort level

Level 1:
Parity-checked oligo

Level 2:
Exact address match

Level 3:
Imperfect address match or Hephaestus (?)

Post-sequencing processing pipeline (under discussions)



More complexity required

Sort level

Level 1:
Parity-checked oligo

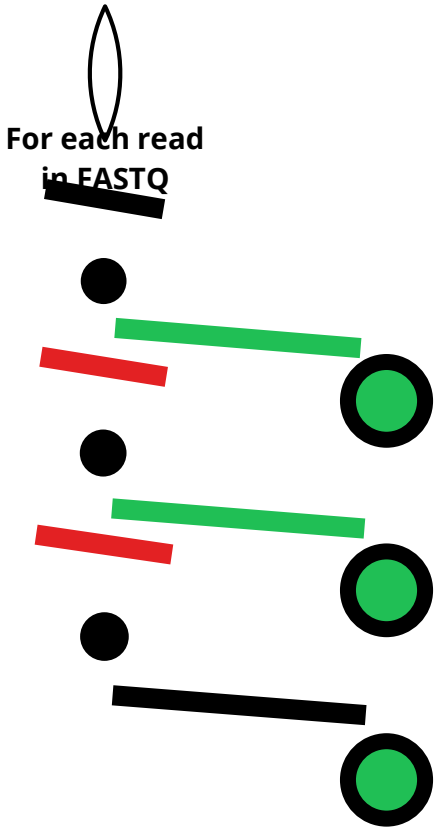
Level 2:
Exact address match

Level 3:
Imperfect address match or Hephaestus (?)

Data sorted by address

Address 1	
Address 2	Data ✓
Address 3	
. . .	. . .
Address N	Data ✓

Post-sequencing processing pipeline (under discussions)



More complexity required

Sort level

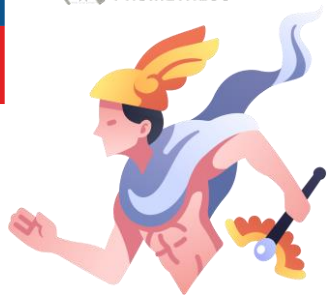
Level 1:
Parity-checked oligo

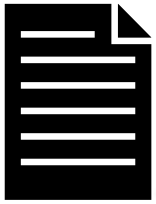
Level 2:
Exact address match

Level 3:
Imperfect address match or Hephaestus (?)

Data sorted by address

Address 1	
Address 2	Data
Address 3	Data
. . .	. . .
Address N	Data





Fastq

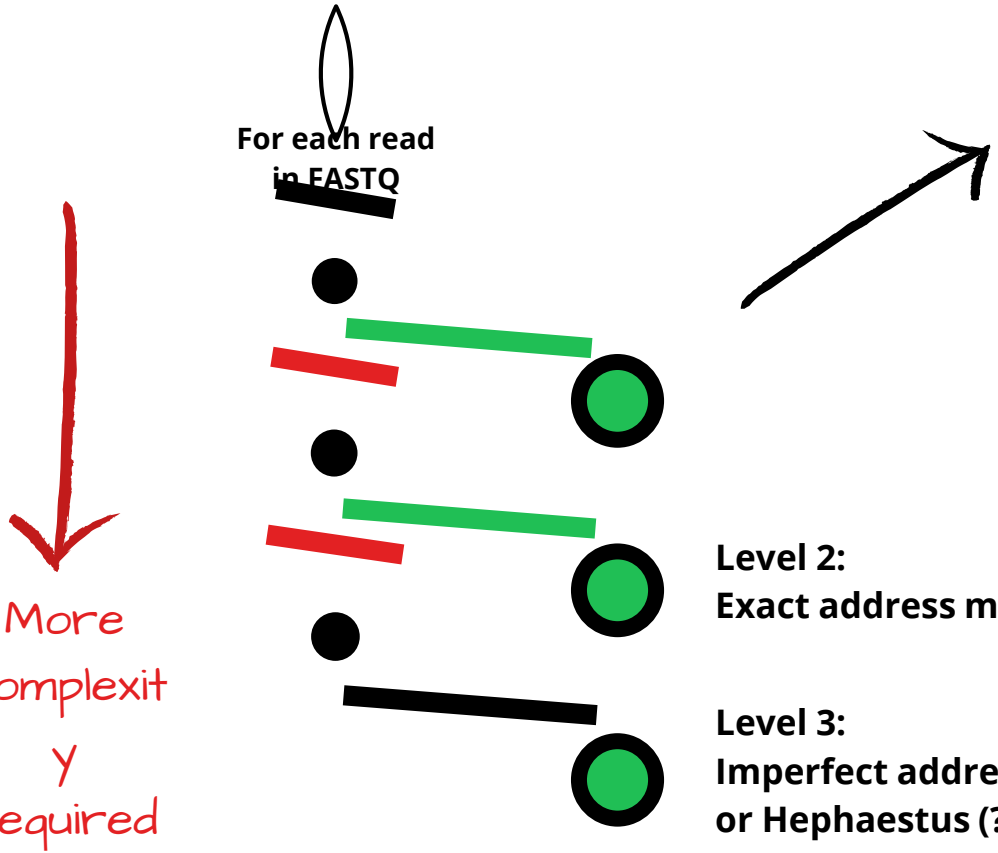
Address

ECC sequence

5'-(...)TACGCCGCAACCTGCCATCCCAAGCTAAAAGGGGAATCCCAATGGCAATCTGCACCGTCGGACGA(...)-3'

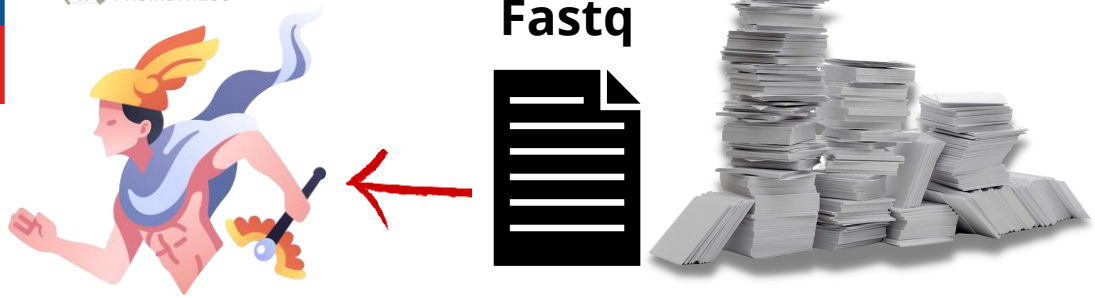
- Integrity check (sequence length and ECC)
- Try to identify the address
- Pass integrity and identify the address → This is a perfect sequence, hold to send to Artemis for decoding
- Fails integrity, but identify the address
- Fails integrity, and fails to get address

Post-sequencing processing pipeline (under discussions)



Sorting Level 1

- Entire sequence free of errors (verified by ECC);
- High-fidelity
- No consensus required



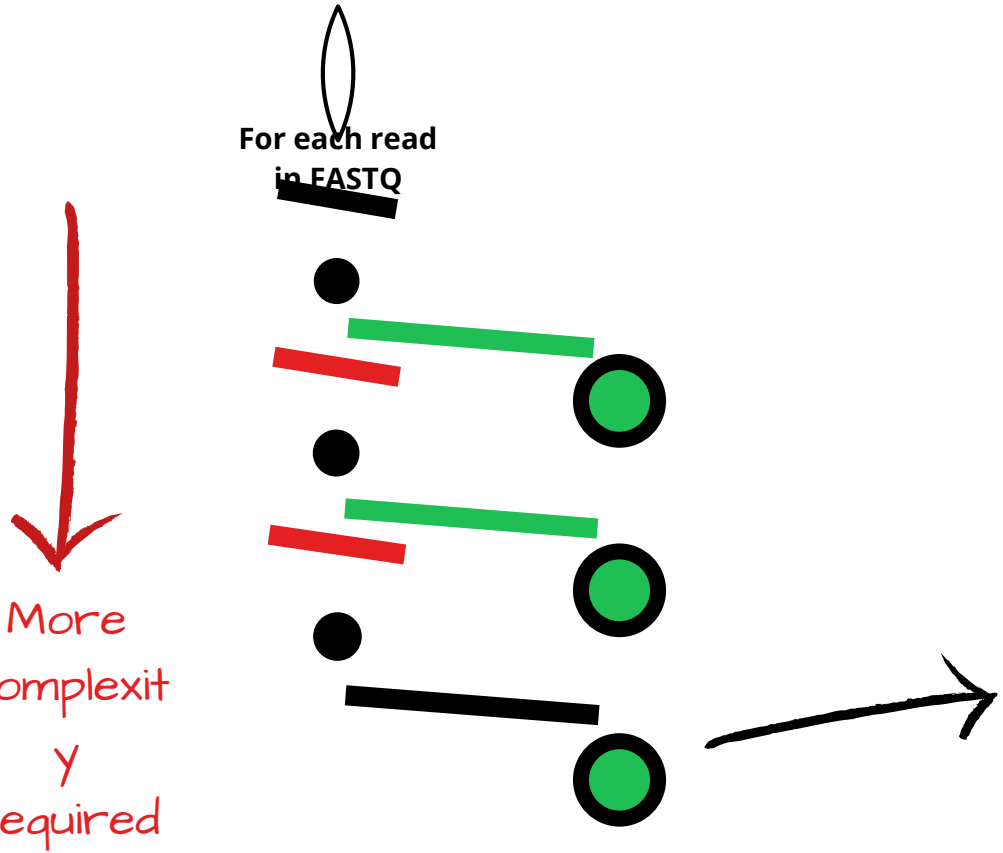
Address

ECC sequence

5'-(...)TACGCCGCAACCTGCCATCCCAAGCTAAAAGGGGAATCCCAATGGCAATCTGCACCGTCGGACGA(...)-3'

- Integrity check (sequence length and ECC)
- Try to identify the address
- Pass integrity and identify the address
- Fails integrity, but identify the address → Put this sequence in the pile for that address. If we don't find a perfect one, do MSA and consensus using the pile.
- Fails integrity, and fails to get address

Post-sequencing processing pipeline (under discussions)



Sorting Level 1

- Entire sequence free of errors (verified by ECC);
- High-fidelity
- No consensus required

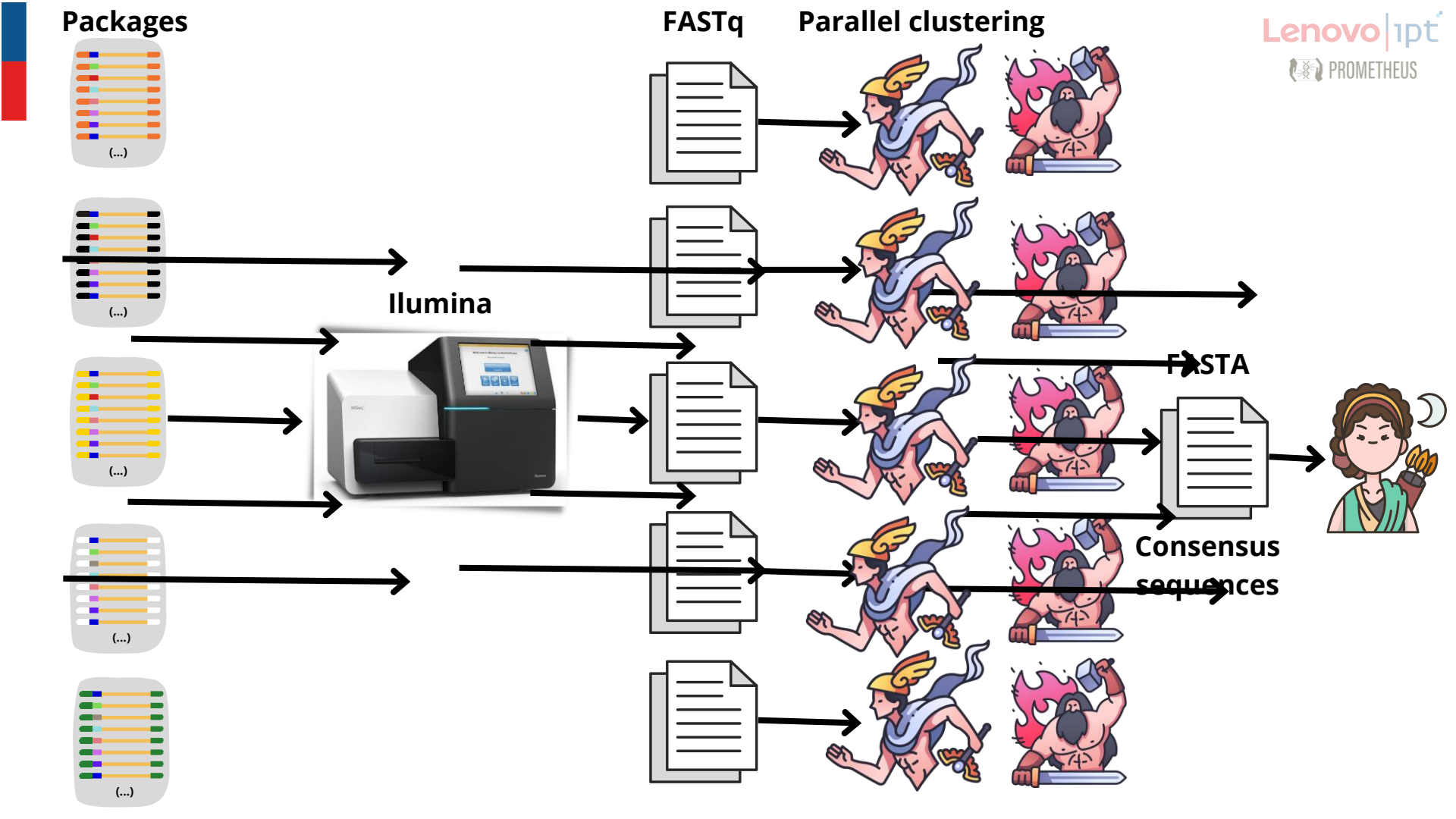
Sorting Level 2

- Address is free of errors, but payload is not
- Consensus required
- Works better with distance-designed addresses

Sorting Level 3

Imperfect match: find address closest to the tag found in oligo

- Hephaestus: cluster all reads that have not been categorized.


Packages
(...)
(...)
(...)
(...)
(...)
Ilumina
FASTq
Parallel clustering
FASTA
Consensus
sequences
Lenovo | ipt
PROMETHEUS

Finally, a small Sample

PROMETHEUS

Lenovo | ipt

Input file: photo from workshop
(1 data block)

- 5863 original sequences (targets)
- 64493 simulated sequences (reads)
- 100% of sequences have 2 substitution errors

Thank you!

Dr. Adriano Leal

leal@ipt.br

linkedin.com/school/iptsp/

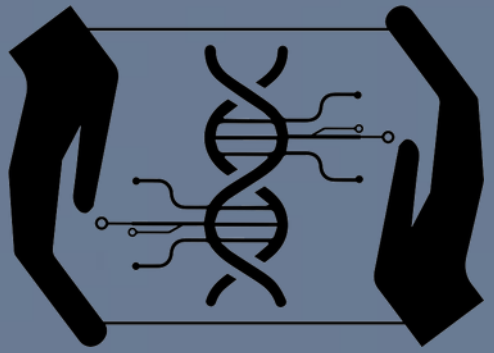
instagram.com/ipt_oficial/

youtube.com/@IPTbr/

www.ipt.br

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

SÃO PAULO GOVERNO DO ESTADO



Lenovo™

ipt