



UNIVERSITAS
GADJAH MADA



Machine Learning in Bioinformatics

Afiahayati, Ph.D.
Lab. of Intelligent System
Dept. of Computer Science and Electronics
Universitas Gadjah Mada
afia@ugm.ac.id



Afiahayati, S.Kom., M.Cs., Ph.D

2004 – 2008 **S.Kom**, Universitas Gadjah Mada
2008 – 2010 **M.Cs.**, Universitas Gadjah Mada
Oct 2010 Short Course of Computational Logic, TU Dresden
2011 – 2015 **Ph.D**, Keio University, Japan

Thesis :

2008 Multiple Sequence Alignment Using Hidden Markov Model (S.Kom)

2010 Multiple Sequence Alignment Using Hidden Markov Model With Augmented Set

And Its Influence On Phylogenetic Tree Accuracy (M.Cs)

2015 Development of *de novo* assemblers for metagenomic sequencing data (Ph.D)

Selected Awards :

1. Mahasiswa berprestasi UGM 2007
2. Accenture High Performer Scholarship 2007
3. Asea Uninet Scholarship 2009, On Place Master Program, funded by Austrian Gov.
4. Asia Development Scholarship 2011
5. Google Anita Borg Memorial Scholarship 2012
6. Schlumberger Faculty for The Future Fellowship Award 2013



UNIVERSITAS
GADJAH MADA



Collaborations

- Covid19 Genome Analysis : dr. Gunadi, Ph.D. (FKKMK UGM) , drh. Hendra Wibawa, DVM, MSc, PhD (Disease Investigation Center Wates)
- Cancer Analysis : Prof. Sofia Mubarika, Ph.D. (FKKMK UGM)
- GamaComet : drg. Ryna Y., Ph.D. (FKG UGM)
- Metagenomic Analysis : Prof. Sakakibara (Keio Univ, Japan)
- NGS analysis for Banana : Prof. Siti Subandiyah (Fac. of Agriculture, UGM)



Outlines

- Bioinformatics discipline
- Genetics Background
- Datasets
- Overview Machine Learning
- Genome Assembly
- Protein Secondary Structure Prediction
- GAMAComet
- Several potential research topics in Indonesia



UNIVERSITAS
GADJAH MADA

Bioinformatics Discipline

Bioinformatics?

- You do not need A-level biology for the course
 - We don't have any background requirements
- AI might help, databases might help...



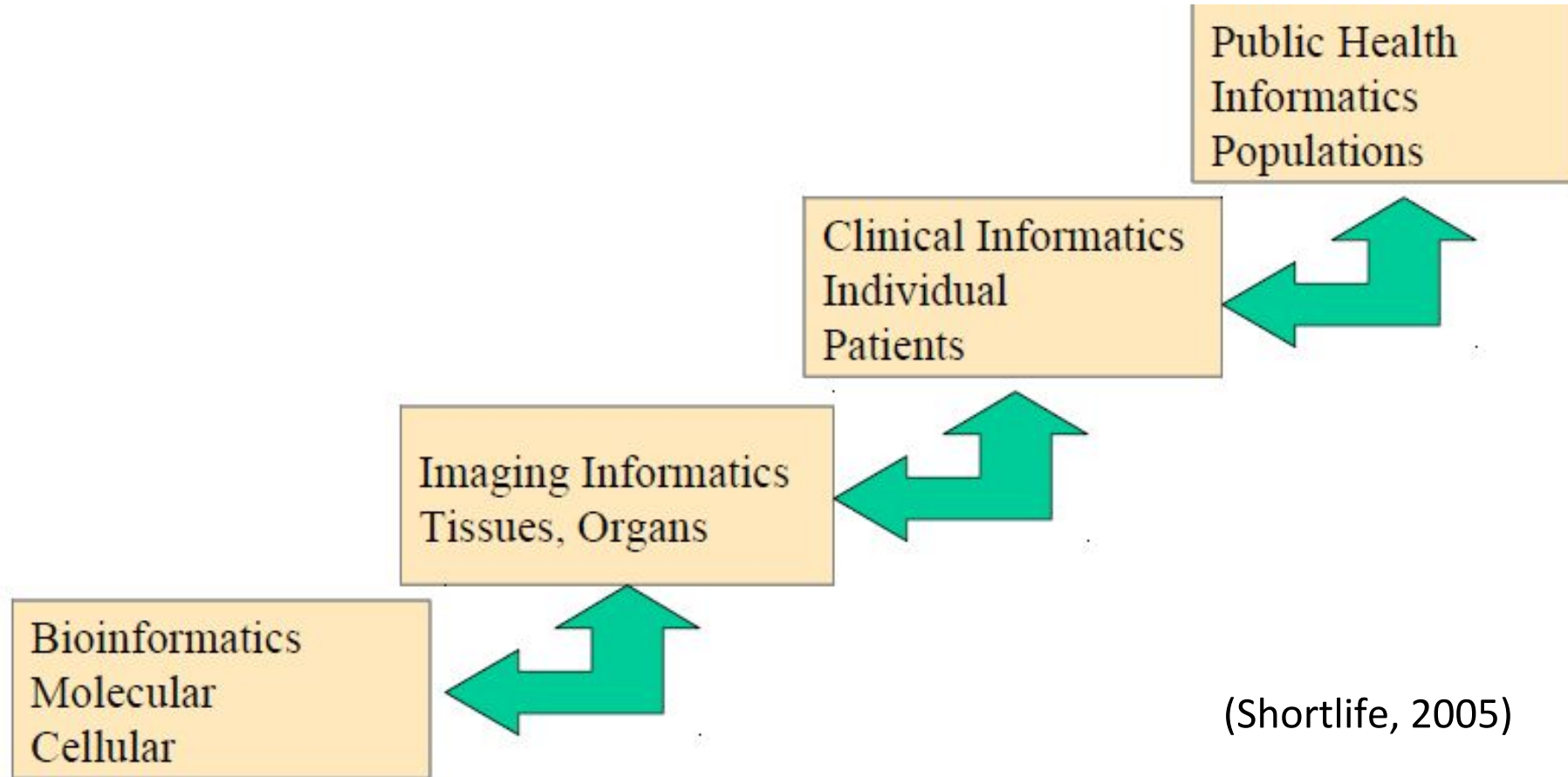
Bioinformatics?

- An intersection of AI and genetics
 - Two very popular (most wanted) sciences
- An opportunity to:
 - Use some of the most interesting computational techniques to solve some of the most important and rewarding questions

Health Information



UNIVERSITAS
GADJAH MADA



(Shortlife, 2005)



UNIVERSITAS
GADJAH MADA

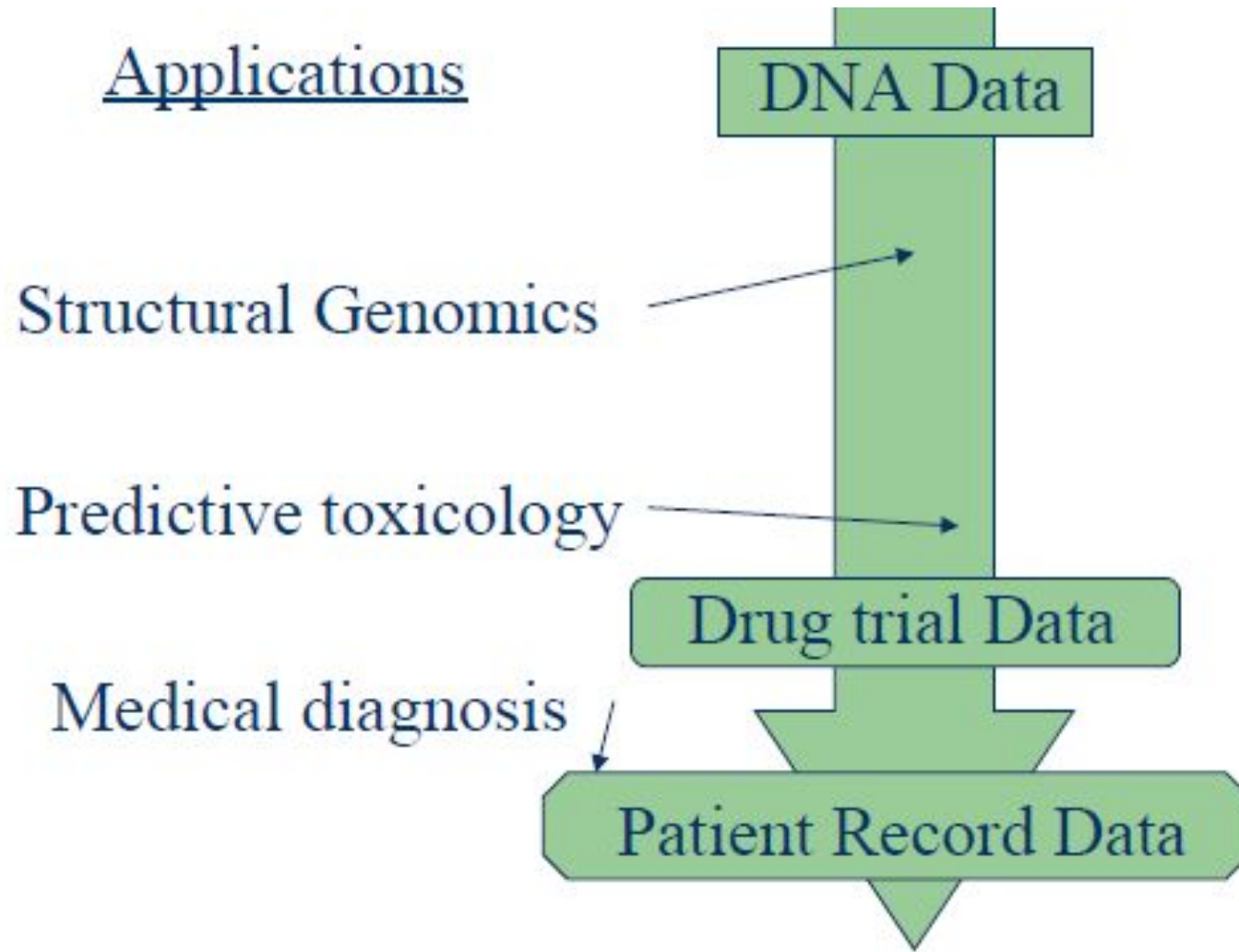
Genetic Backgrounds

Bioinformatics Data



UNIVERSITAS
GADJAH MADA

Applications

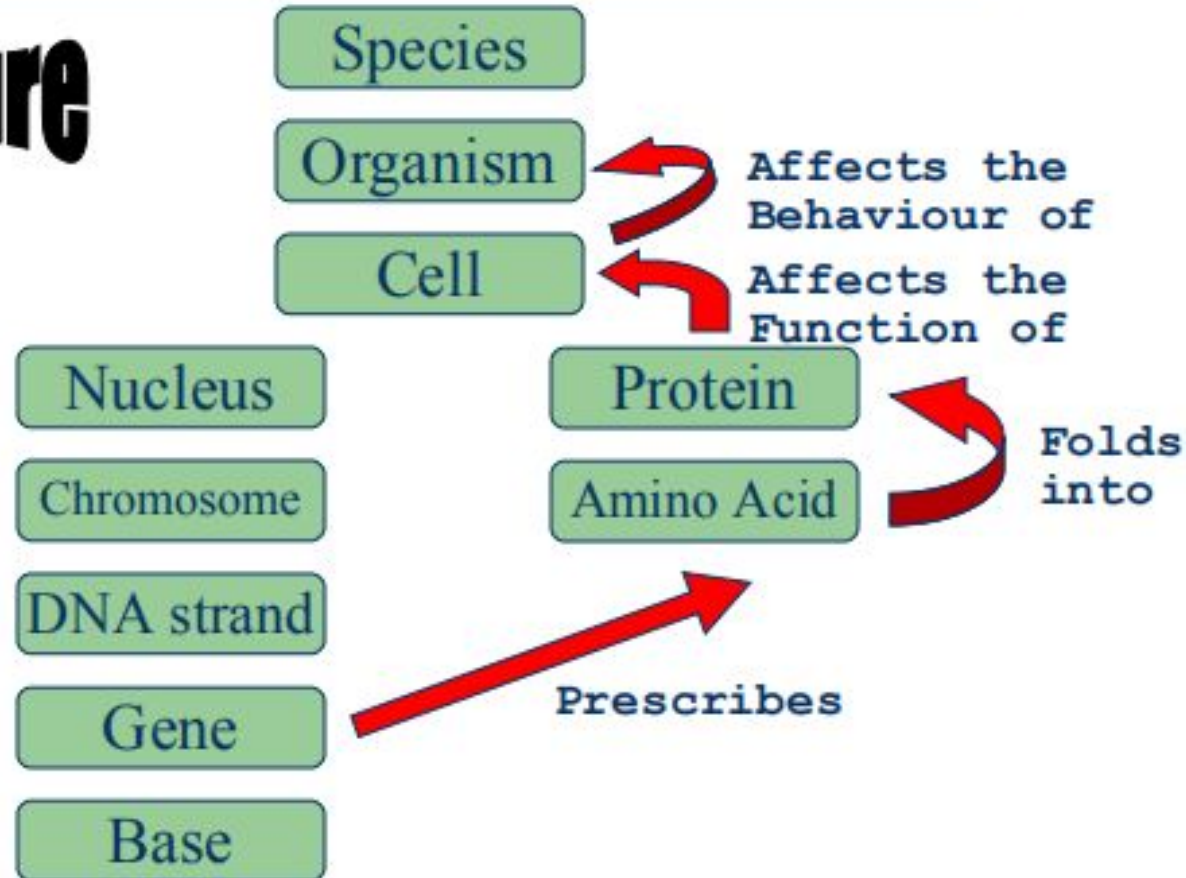


Substructure and Effect (Top Down/Bottom Up)



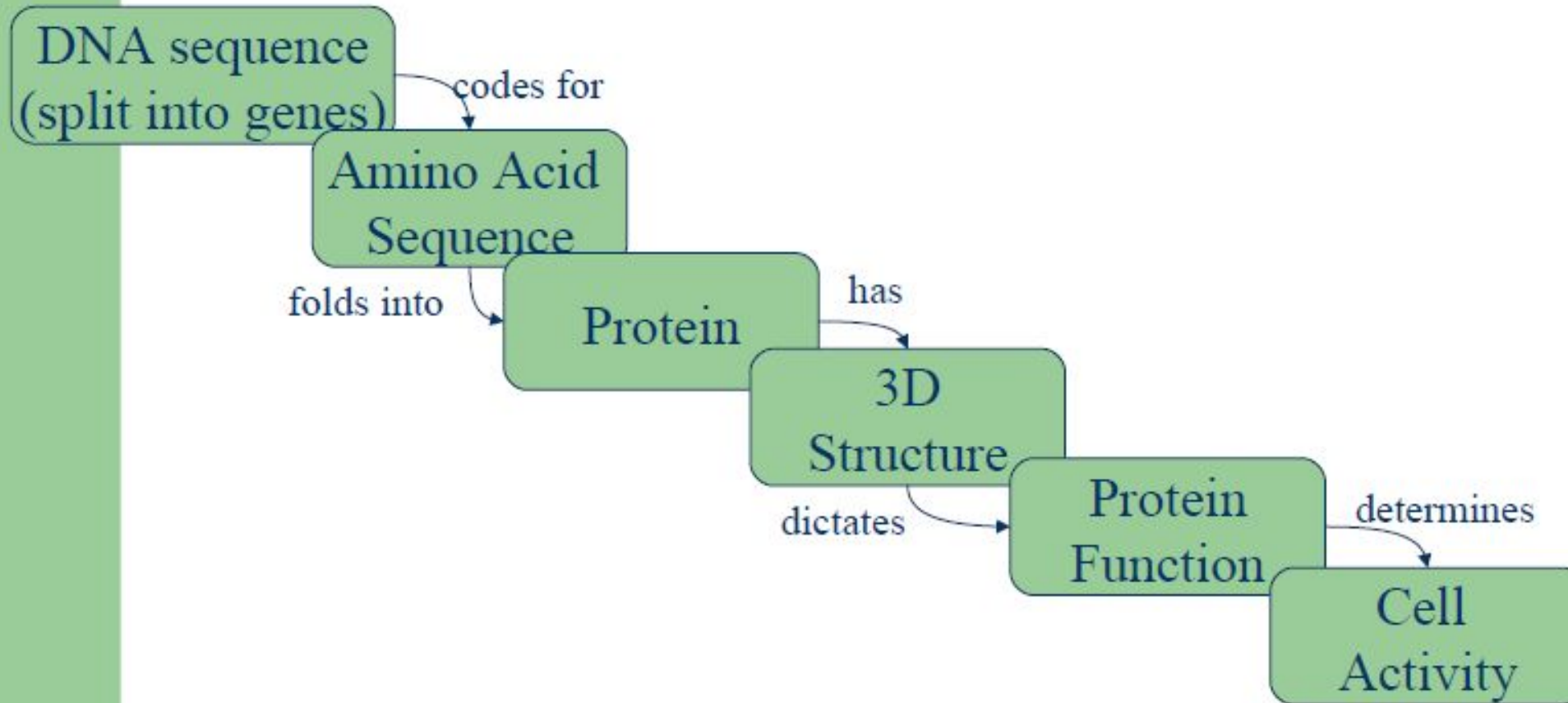
UNIVERSITAS
GADJAH MADA

Substructure





From DNA to Cell Function



DNA Sequence



Analyze Data Workflow Shared Data Visualization Help Login or Register Using 79%

This dataset is large and only the first megabyte is shown below.
[Show all](#) | [Save](#)

```
@1/1
CCCTGGGCAACTTCCTGTTCTTCTTTGTTCTATTCCCCTTACCTAATTAATAAAGTTTTAACTAATAGCCAACCTGGGTAAAGTGAAA
+
@@CFF3DDDAHHFIIII>HIIIIICHIIIIIIGHII>EHEB@DGEDGGEHFINGIHGEHIGHGGEGHIGHI:FHHHEIDENHEBDFBFEF
@1/2
ACCCTACTGCCGTGTCCAGTTTCCATTGGCTGGAATAAGACCTCACATTTTACACTTTACCCAGTTGGCTATTAGTTTAAAACTTTATTA
+
@@@DDFFFHFHAHINGHAGFBHII>HGIJJGII;FFGH<@BCDDGA<BFFFE@?B@GGGGIIGG;@DECAEHHFE=>CBECFD;@@;@##
@2/1
CCCGATTGTACACCTGTTCAATTCTGAGATAGGAGGAAAACCACCCTATGGTGGGAGGTGAGACATGTTGGCAGCAATGCTGTCTAGTTA
+
@CCFFDFHHHGHIIJHGIHGHIIJJJJCHGHIEIEIFCHIGIJJJJEIJ?FGDGH;DAEEH>EHBDFFCBAAEDDDDDDDDED>CE
@2/2
TGTTACTGTCCACCCAACATTTTCAGTGGAGTAAAGAATAACTAGACAGCATTGCTGCCAACATGTCTCACCTCCCACCATAGGGTGGTT
+
CC@FFFDHHHHHJJIIIGGGIIJJJCHIEGFFHHIJJIIJJIIJJIIJJIDGEGHIIJJIIIBHFGFDEGHJFFHHFF<B;CEC6;88?
@3/1
GCAATCTGGGTGGAAGTTCTTTAATATGAACATTTCAACCACCTTCATTCTACCATGTCCACTATCAGCACATTCAAACCTGATCCAGCCA
+
@@@FFDEFHFFHHJJJJJHIIIGIJJJFIGJJJJIIJJJEGHCGHGHIGIGGGHGGIIIHGHIIJJGGGEGIEGCDHHBCEHFFFFFFFCE
```

History

search datasets

Unnamed history
2 shown, 1 [hidden](#)

4 GB

2: Trimmomatic on SRR1220154 (fastq-dump)

This is a new dataset and not all of its data are available yet

1: SRR1220154 (fastq-dump)

Amino Acids

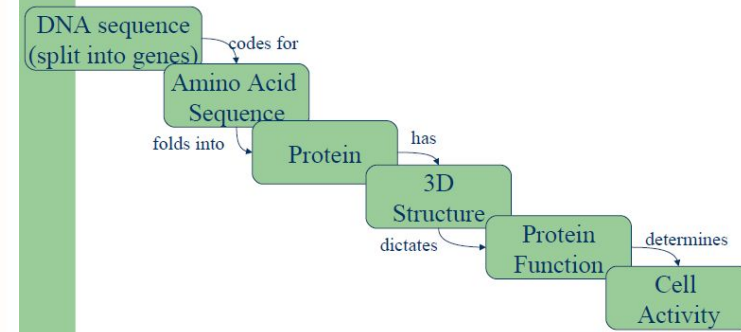


UNIVERSITAS
GADJAH MADA

Genetic Code

		Second letter				
		U	C	A	G	
First letter U	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp	U C A G
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C A G
	A	AUU } AUC } Ile AUA } AUG Met	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	U C A G

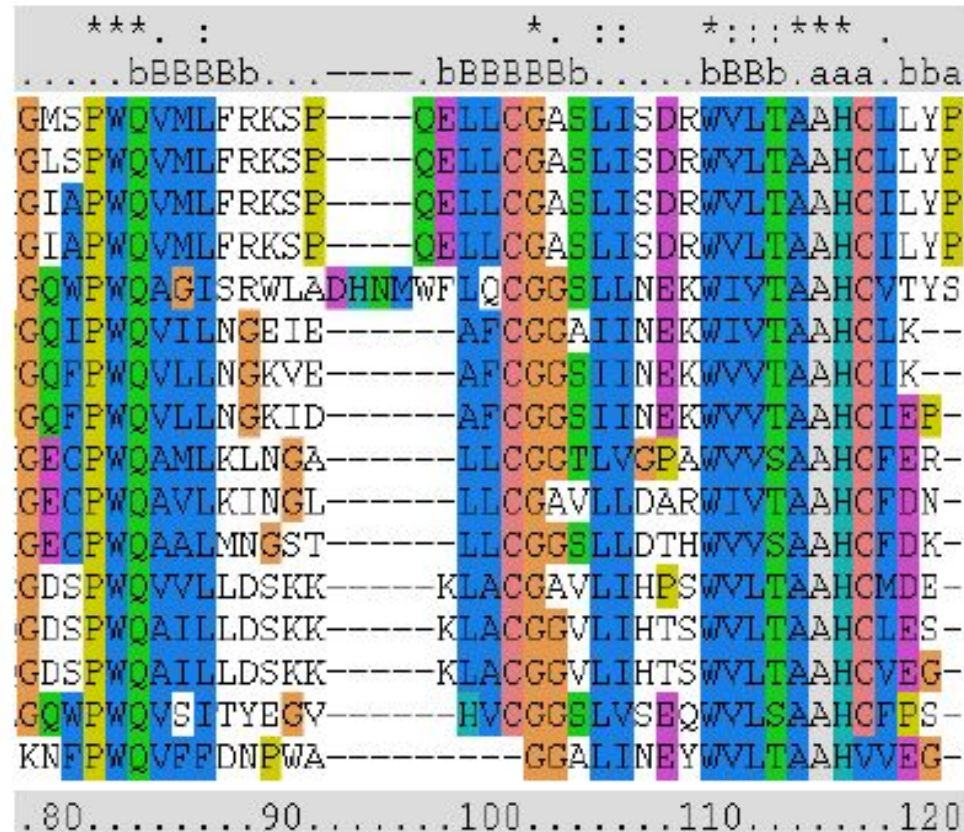
A=Ala=Alanine
 C=Cys=Cysteine
 D=Asp=Aspartic acid
 E=Glu=Glutamic acid
 F=Phe=Phenylalanine
 G=Gly=Glycine
 H=His=Histidine
 I=Ile=Isoleucine
 K=Lys=Lysine
 L=Leu=Leucine
 M=Met=Methionine
 N=Asn=Asparagine
 P=Pro=Proline
 Q=Gln=Glutamine
 R=Arg=Arginine
 S=Ser=Serine
 T=Thr=Threonine
 V=Val=Valine
 W=Trp=Tryptophan
 Y=Tyr=Tyrosine



Multiple Sequence Alignment

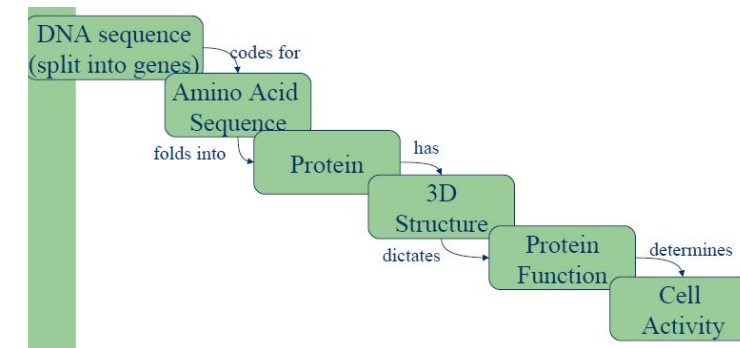
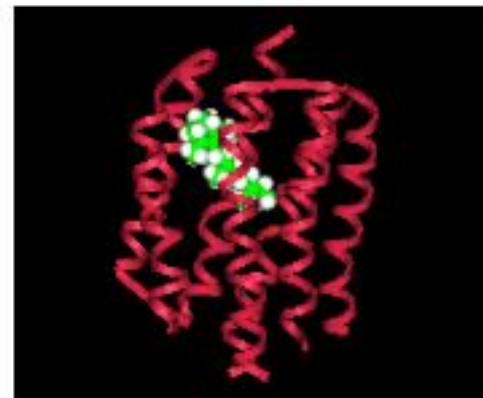
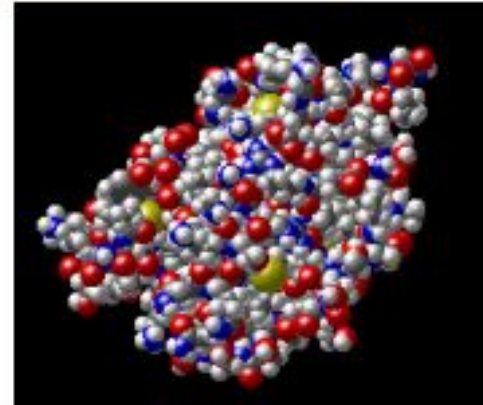


- Protein sequences form families
 - Learn much more about a gene by looking at its family
- Multiple sequence alignment algorithms
 - Profiles
 - PSI-BLAST



Proteins

- DNA codes for
 - strings of amino acids
- Amino acids strings
 - Fold up into complex 3d molecule
 - 3d structures: conformations
 - Between 200 & 400 “residues”
 - Folds are proteins
- Residue sequences
 - Always fold to same conformation
- Proteins play a part
 - In almost every biological process





UNIVERSITAS
GADJAH MADA

Datasets

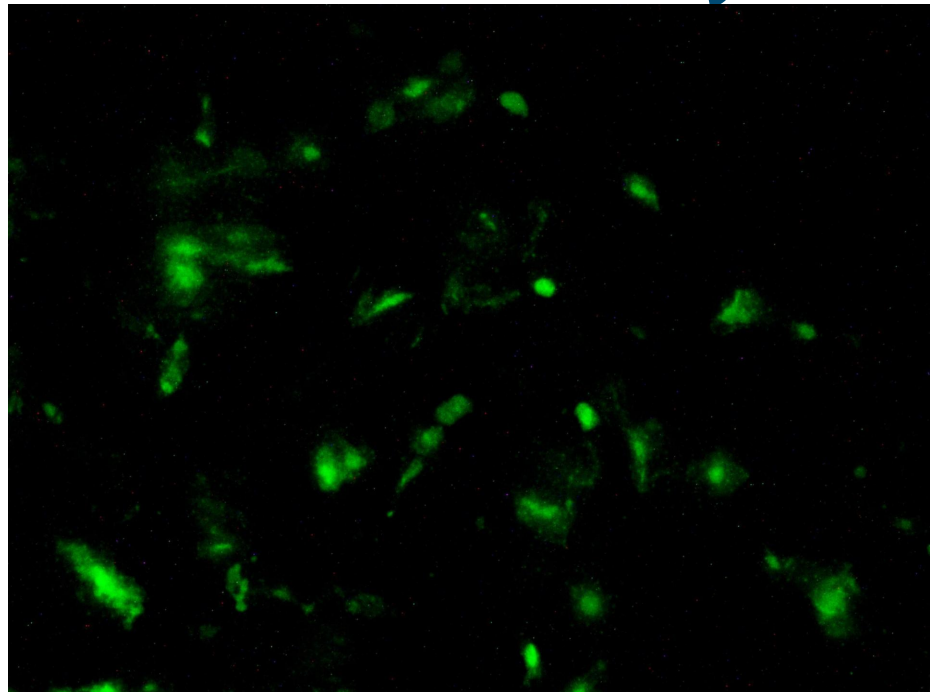


Datasets

- Primary Datasets
 - Data acquisition : DNA sequence, Comet Assay
- Secondary Datasets, benchmarking datasets
 - NCBI (DNA sequence, protein, ...)
 - RS126 (SSP)
 - CullPDB (SSP)
 - Cb513 (SSP)

Primary dataset : DNA Seq

Primary dataset : Comet Assay



Data Article

RNA-seq data of banana bunchy top virus (BBTV) viruliferous and non-viruliferous banana aphid (*Pentalonia nigronervosa*)



Siti Subandiyah ^{a, b}, Ruth Feti Rahayuniati ^{a, d}, Sedyo Hartono ^a, Susanto Somowiyarjo ^a, Afiahayati ^c, Alan Soffan ^{a, b, *}

^a Department of Plant Protection, Faculty of Agriculture, Universitas Gadjah Mada, Yogyakarta, Indonesia

^b Research Center for Biotechnology, Universitas Gadjah Mada, Yogyakarta, Indonesia

^c Department of Computer Science and Electronics Faculty of Mathematics and Natural Sciences Universitas Gadjah Mada, Yogyakarta, Indonesia

^d Department of Agrotechnology, Faculty of Agriculture, Universitas Jenderal Soedirman, Purwokerto, Indonesia

ARTICLE INFO

Article history:

Received 29 September 2019

Received in revised form 12 November 2019

Accepted 13 November 2019

Available online 21 November 2019

Keywords:

Banana aphid

BBTV

RNA-Seq

ABSTRACT

Banana bunchy top disease (BBT) is one of the most economically serious viral diseases of banana caused by banana bunchy top virus (BBTV: Nanoviridae: Babuvirus). BBTV is a circular, ssDNA virus which is suitable in the phloem tissue and currently only being transmitted by the banana aphid (*Pentalonia nigronervosa*) in a persistent, non-propagative, circulative manner. Interaction of BBTV and banana aphid had been studied in several ways, such as transmission and translocation of BBTV inside the banana aphid body at cellular level. However, the molecular mechanism underlying the interaction between BBTV and banana aphid have been poorly understood.

Secondary Datasets

Benchmarking datasets :CullPDB, CB513



← → ↻ 🔒 princeton.edu/~jzthree/datasets/ICML2014/

Index of /~jzthree/datasets/ICML2014

<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
 Parent Directory		-	
 cb513+profile_split1.npy.gz	2014-01-21 00:49	9.3M	
 cullpdb+profile_5926.npy.gz	2018-10-28 23:53	103M	
 cullpdb+profile_5926_filtered.npy.gz	2018-10-28 23:53	95M	
 cullpdb+profile_6133.npy.gz	2014-01-21 00:37	108M	
 cullpdb+profile_6133_filtered.npy.gz	2014-01-21 00:37	100M	
 cullpdb.tar.gz	2017-02-25 23:02	1.1G	
 dataset_readme.txt	2018-10-28 23:53	2.7K	
 slides.pdf	2015-10-01 21:16	822K	

Secondary Datasets

NCBI



UNIVERSITAS
GADJAH MADA

ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

All Databases Search

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: <https://www.coronavirus.gov>.
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy

Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code

Analyze
Identify an NCBI tool for your

Research
Explore NCBI research and

Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI News & Blog
The BLAST Docker and databases are now ready to use on Google and Amazon clouds.

Secondary Datasets

GISAID



UNIVERSITAS
GADJAH MADA

gisaid.org


GISAID Login

About us Database Features Events Collaborations References Registration Help

In Focus

GISAID Comments on the WHO Report on the Public Health Implications of Implementation of the Nagoya Protocol

Member States representatives met at the 72nd annual World Health Assembly in Geneva, Switzerland to discuss and debate a report prepared by the WHO on the public health implications of implementation the Nagoya Protocol.




GISAID comments on that paper, highlighting the current issues around sharing of seasonal influenza viruses, the consequences of delays in virus sharing, and the connection to the discussion at the Convention on Biological Diversity.

> [read GISAID's Comments on the Report](#)


● ● ● ● ●

Genomic epidemiology of hCoV-19

Showing 1,171 of 1,171 genomes sampled between Dec 2019 and Feb 2020.



COVID-19 Global Cases



EpiCoV Data Curation Team

The Francis Crick Institute, London

Gabriel Lihue Rojo
Hospital de Niños Dr. Ricardo Gutierrez, Buenos Aires

Recent hCoV-19 data submissions

- [hCoV-19/Belgium/ULG-10204/2020](#)
- [hCoV-19/South Korea/A6/2020](#)
- [hCoV-19/USA/CA-UW-606/2020](#)



UNIVERSITAS
GADJAH MADA

Overview Machine Learning

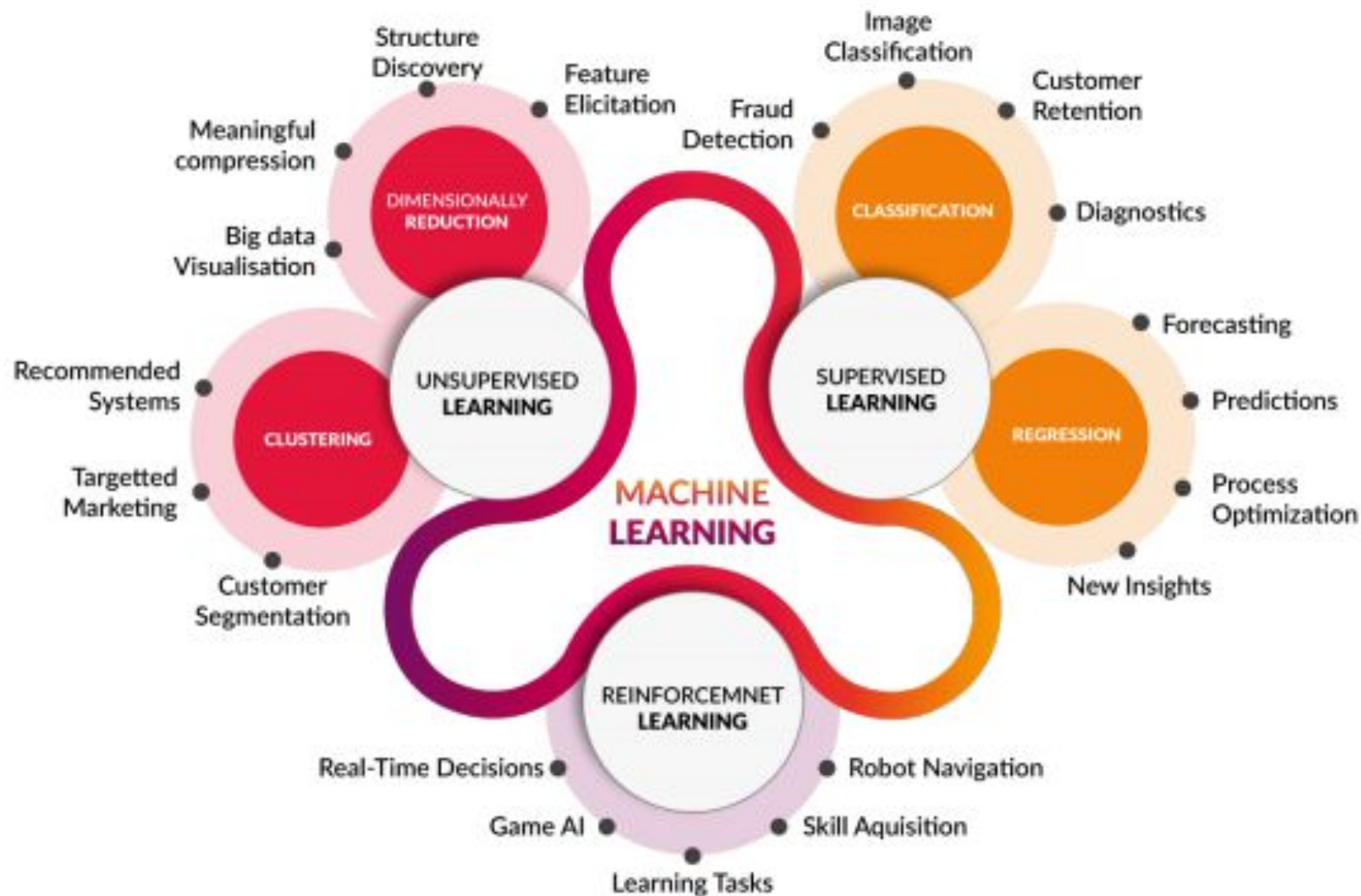
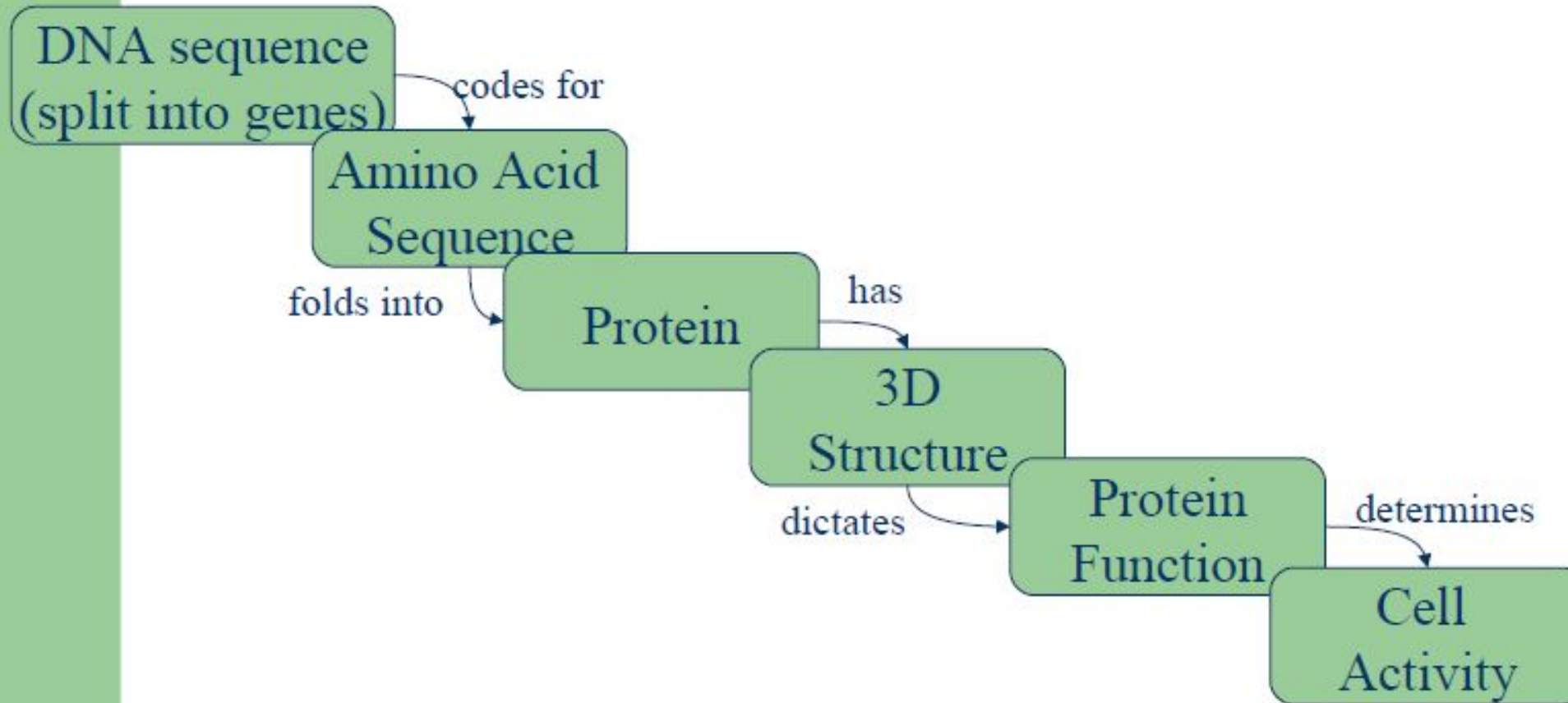


Figure: ML Map (Image source: <http://www.cognub.com/index.php/cognitive-platform/>)

Machine Learning in Bioinformatics



UNIVERSITAS
GADJAH MADA





UNIVERSITAS
GADJAH MADA

Genome Assembly

DNA Data => Next Generation Sequencing

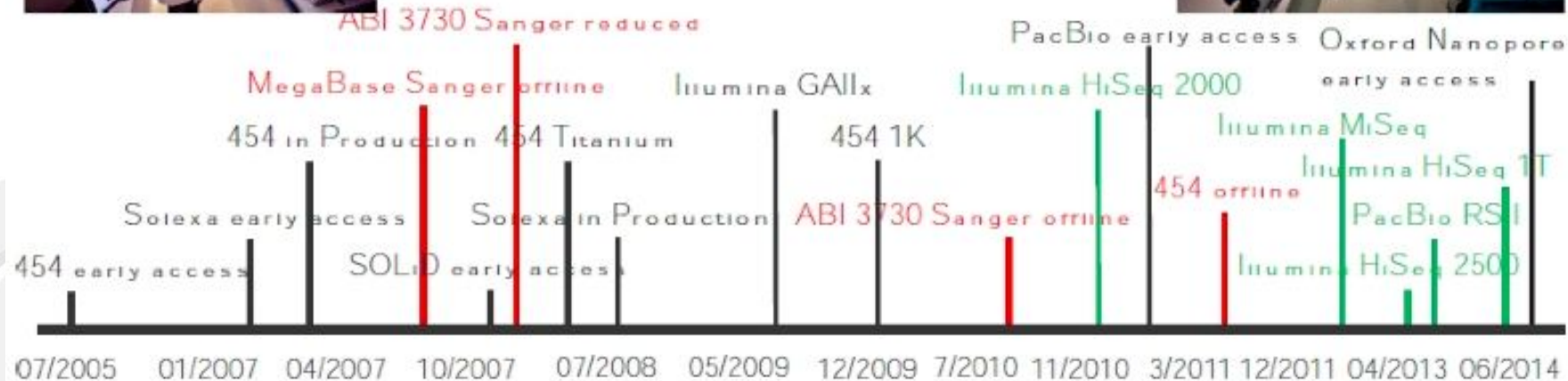


UNIVERSITAS
GADJAH MADA







Staying state of the art



Sanger Sequencing to Next-Gen Sequencing by Synthesis



Sequencing Platform

						
	Illumina HiSeq 1T	Illumina HiSeq 2500	Illumina HiSeq 2000	Illumina NextSeq 500	Illumina MiSeq	Pacific Biosciences RSII
Reads (Single-Read/Cluster)	>1,500 Million per Flowcell	200 Million per Flowcell	>1,000 Million per Flowcell	400 Million per Flowcell	>10 Million per Flowcell	0.06 Million per SMRT Cell
Readlength	2 X 150bp Max*	2 X 250bp Max	2 X 150bp Max*	2 X 150 Max	2 X 300bp Max	11,000bp Avg; 50,000bp Max
Total Bases	500 Gb per Flowcell	130 Gb per Flowcell	350 Gb per Flowcell	>100 Gb per Flowcell	5-20 Gb per Flowcell	>0.4 Gb per SMRT Cell
Run Time	7 Days for 2 X 150	4.5 Days for 2 X 250	16 Days for 2 X 150	1 Days for 2 X 150	2 Days for 2 X 300	0.08-0.12 Days (2-4 hours)
Applications	Primary Sequence Generator at JGI	Rapid output HiSeq	Supplement / Backup Platform	Development / Supplement	16x iTags, Library QC, R&D	Assembly improvement, de novo, SynBio validation, methylation/epigenome

NGS Data



genome.gov/human-genome-project

COVID-19 is an emerging, rapidly evolving situation. [CDC health information](#) [NIH research information](#)

NIH National Human Genome Research Institute

Begin your search here



[About Genomics](#)

[Research Funding](#)

[Research at NHGRI](#)

[Health](#)

[Careers & Training](#)

[News & Events](#)

[About NHGRI](#)



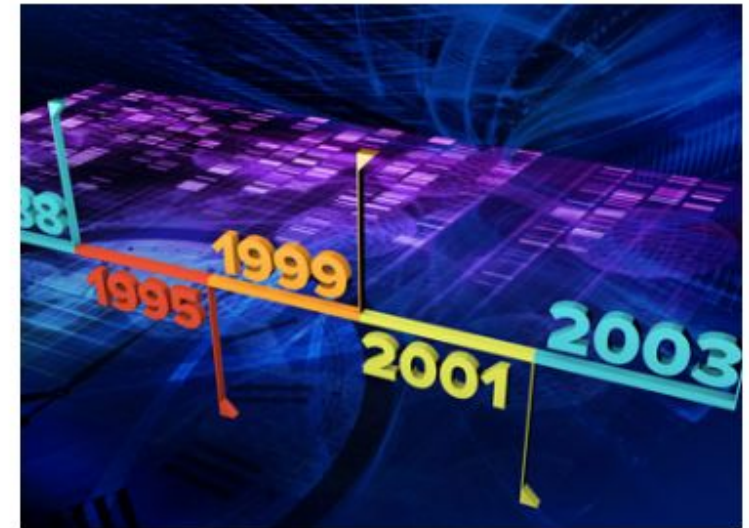
What is the Human Genome Project?>

The Human Genome Project was the international research effort to determine the DNA sequence of the entire human genome.



Human Genome Project Results>

In 2003, an accurate and complete human genome sequence was finished two years ahead of schedule and at a cost less than the original estimated budget.



Human Genome Project Timeline of Events>

Key moments and press releases from the history of the Human Genome Project.

DNA Sequences



UNIVERSITAS
GADJAH MADA

Analyze Data Workflow Shared Data Visualization Help Login or Register Using 79%

This dataset is large and only the first megabyte is shown below.
[Show all](#) | [Save](#)

```
@1/1
CCCTGGGCAACTTCCTGTTCTTCTTTGTTCTATTCCCCTTACCTAATTAATAAAGTTTTAACTAATAGCCAACCTGGGTAAAGTGAAA
+
@@CFF3DDDAHFFIIII>HIIIIICHIIIIIIGHII>EHEB@DGEDGGEHFINGIHGEHIGHGGEGHIGHI:FHHHEIDEHHEBDFBFEF
@1/2
ACCCTACTGCCGTGTCCAGTTTCCATTGGCTGGAATAAGACCTCACATTTTACACTTTACCCAGTTGGCTATTAGTTTAAAACTTTATTA
+
@@@DDFFFHFHAHINGHAGFBHII>HGIJJGII;FFGH<@BCDDGA<BFFFE@?B@GGGGIIGG;@DECAEHHFE=>CBECFD;@@;@##
@2/1
CCCGATTGTACACCTGTTCAATTCTGAGATAGGAGGAAAACCACCCTATGGTGGGAGGTGAGACATGTTGGCAGCAATGCTGTCTAGTTA
+
@CCFFDFHHHGHIIJHGIHGHIIJJJJCHGHIEIEIFCHIGIJJJJEIJ?FGDGH;DAEEH>EHBDFFCBAAEDDDDDDDDED>CE
@2/2
TGTTACTGTCCACCCAACATTTTCAGTGGAGTAAAGAATAACTAGACAGCATTGCTGCCAACATGTCTCACCTCCCACCATAGGGTGGTT
+
CC@FFFDHHHHHJJIIIGGGIIJJJCHIEGFFHHIJJIIJJIIJJIDGEGHIIJJIIIBHFGFDEGHJFFHHFF<B;CEC6;88?
@3/1
GCAATCTGGGTGGAAGTTCTTTAATATGAACATTTCAACCACCTTCATTCTACCATGTCCACTATCAGCACATTCAAACCTGATCCAGCCA
+
@@@FFDEFHFFHHJJJJJHIIIGIJJJFIGJJJJJJJEGHCGHGHIGIGGGHGGIIIHGHIIJJGGGEGIEGCDHHBCEHFFFFFFFCE
```

History

search datasets

Unnamed history
2 shown, 1 [hidden](#)

4 GB

2: Trimmomatic on SRR1220154 (fastq-dump)

[1: SRR1220154 \(fastq-dump\)](#)

This is a new dataset and not all of its data are available yet



DNA data Sequence



<http://metagenomicsrevealed.yolasite.com/studies.php>
Environmental samples



<http://www.well.ox.ac.uk/ogc/hiseq-2500-upgrade>
<http://www.seqwright.com/researchservices/nextgen454intro.html>

**Sequenced by
NGS technologies**

```
TCACTGATGCTA
ATCTCTCTCCAA
GTCGCTTAACCA
CTTCCGAAGCTC
TCCGATCCAATC
...
...
...
ATCCGTAATTAG
```

DNA reads



Advantages :

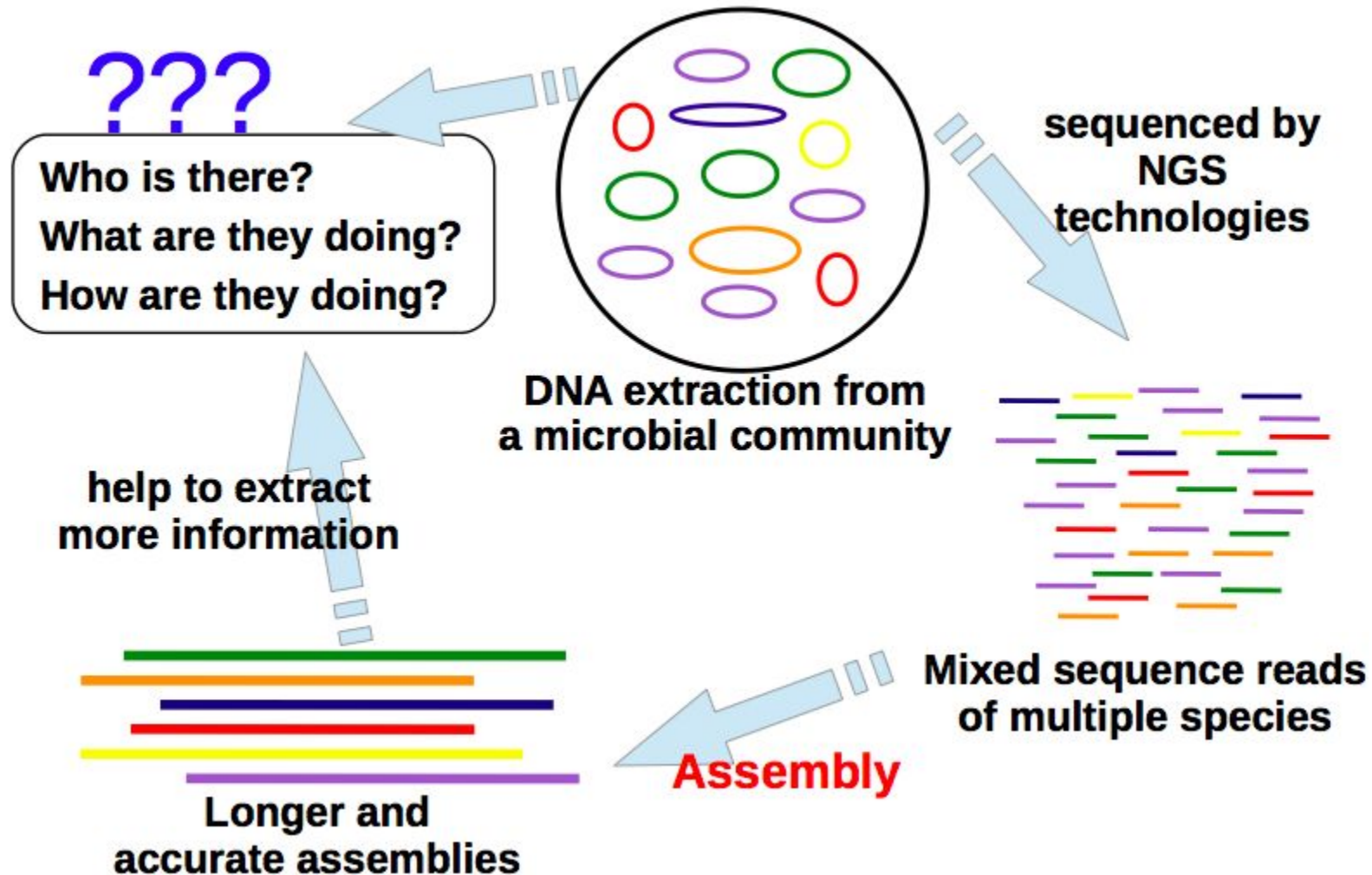
- > Lower cost
- > High throughput sequencing
- > Deep sequencing

(Afiahayati et al., 2015)

Metagenomic Analysis

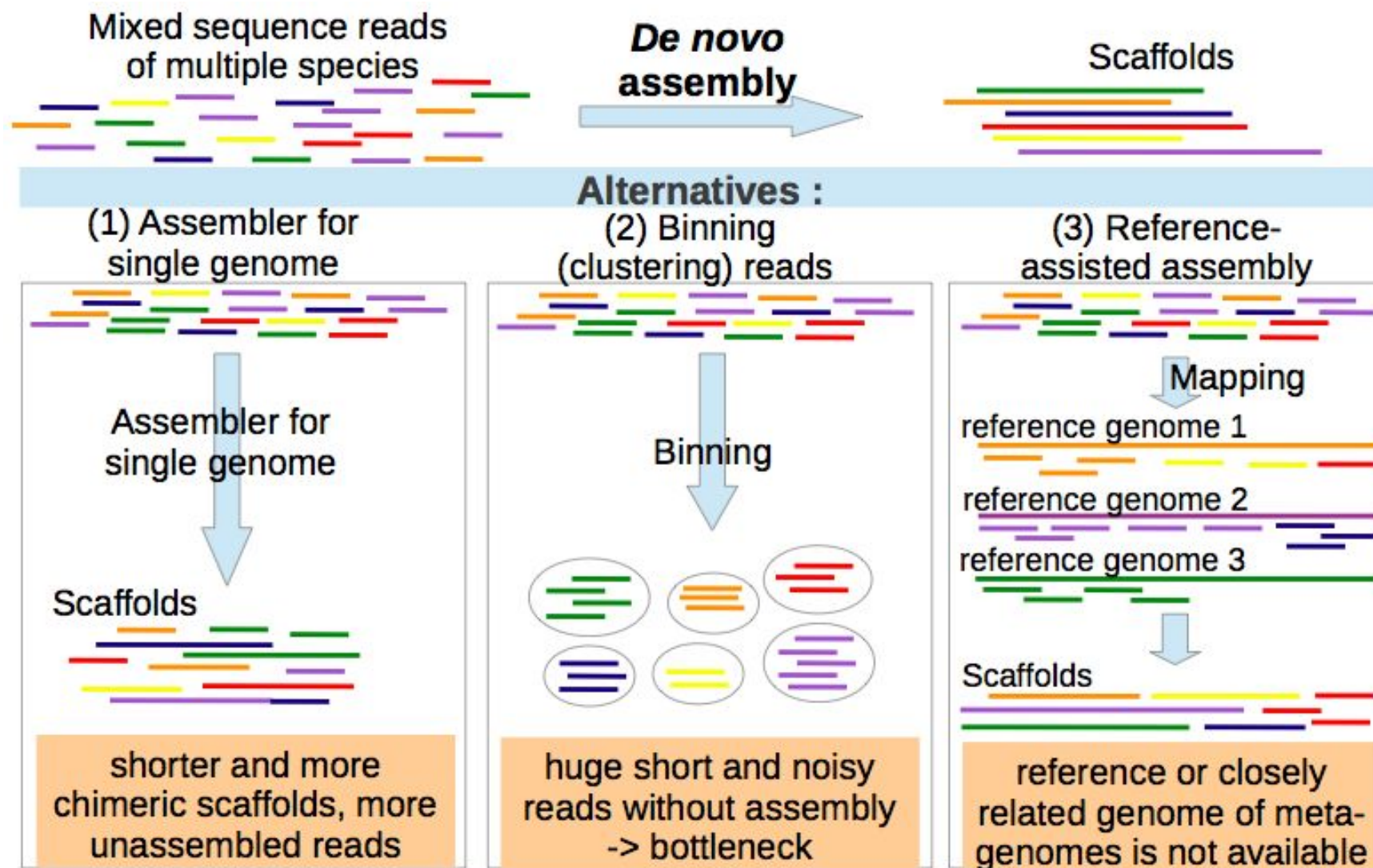


UNIVERSITAS
GADJAH MADA





Metagenomic Analysis - Metagenomic Assembly



De novo metagenomic assemblers are dispensable

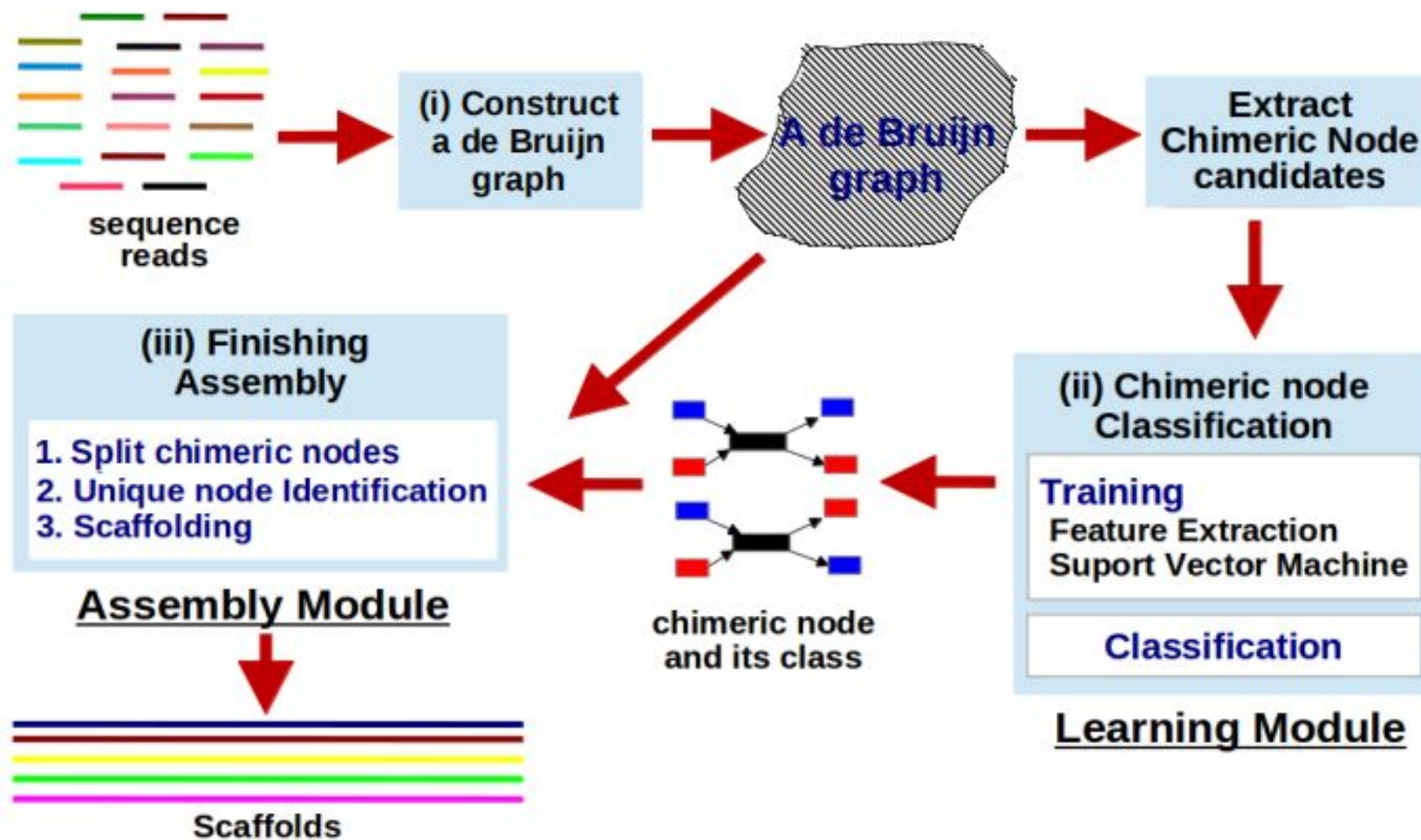


MetaVelvet-SL

<http://metavelvet.dna.bio.keio.ac.jp/MSL.html>

(Afiahayati et al., 2015)

System of MetaVelvet-SL



Computational Infrastructure for Bioinformatics Analysis



UNIVERSITAS
GADJAH MADA

- NGS => Super Big data => super performance computing
- High performance computing infrastructure :
 - Computing cluster :
 - Multiple nodes(servers) with multiple cores
 - High computer memory (64 GB)
 - High performance storage(TB, PB level)
 - Fast networks(10Gb ethernet, infiniband)
 - Operating System : Unix based (Linux)
 - Server room
 - System admin



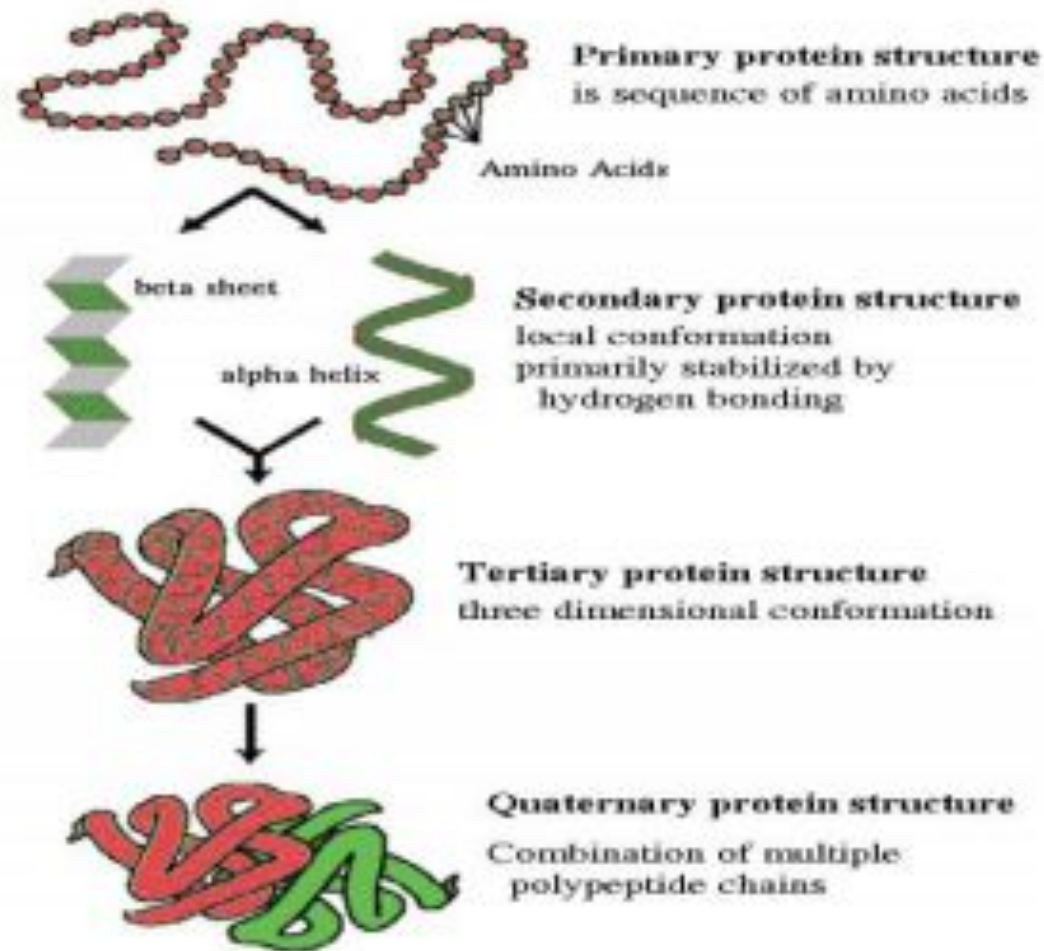
UNIVERSITAS
GADJAH MADA

Protein Secondary Structure Prediction

Secondary Structure Prediction



UNIVERSITAS
GADJAH MADA



Kelas DSSP	Simbol 8 Kelas	Simbol 3 Kelas	Nama Kelas
α -helix 3 ₁₀ -helix	H G	H	<i>Helix</i>
B-Strand B-bridge	E B	E	<i>Sheet</i>
Loop/Irregular B-Turn Bend π -helix	L T S I	C	<i>Coil / Loop</i>

Secondary Structure Prediction



Sekuens 1: Panjang sekuens = 67

SPP	VPSLATISLENSWSGLSKQIQLAQGNNGIFRTPIVLVDNKGNRVQITNV TSKVVTSTNIQLLLNRNI
SSP 3	CCCHHHHHHHHHHHHHHHHHHHHHHCCCCCEEEEEEECCCCCCEEE EECCCHHHHHCECCECCHHHC
SSP 8	LLLHHHHHHHHHHHHHHHHHHHHHTTTTTTEEEEEEEELLSSSSLEEEEET TSHHHHTBLLBLLGGGL
PSSP 3	CC CCCCCCCCCCCCCCCCCCCC
PSSP 8	LL LLLLL

Secondary Structure Prediction



UNIVERSITAS
GADJAH MADA

- Machine learning methods :
 - Multi Layer Perceptron
 - Support Vector Machine
 - Convolutional Neural Network
 - Recurrent Neural Network (as a sequence prediction)

Secondary Structure Prediction menggunakan CNN

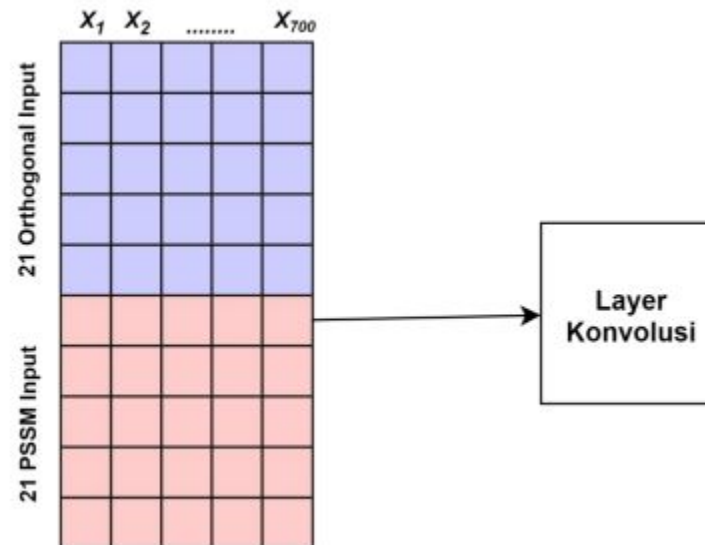


UNIVERSITAS
GADJAH MADA

Orthogonal
Encoding

	A	C	E
A	1	0	0
C	0	1	0
E	0	0	1
D	0	0	0
...
...
Y	0	0	0

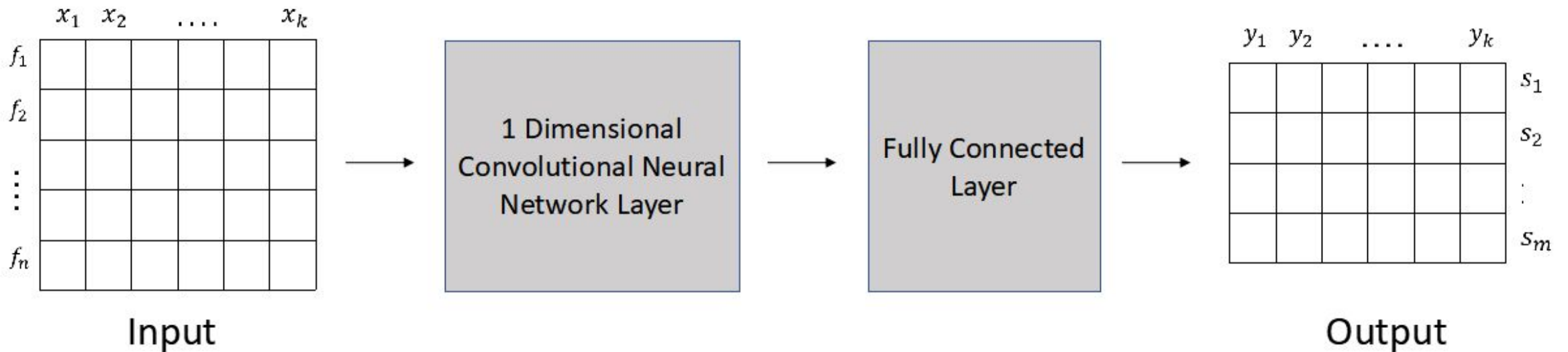
Input Layer -
Convolutional Layer



Secondary Structure Prediction menggunakan CNN



UNIVERSITAS
GADJAH MADA





UNIVERSITAS
GADJAH MADA

GAMAComet

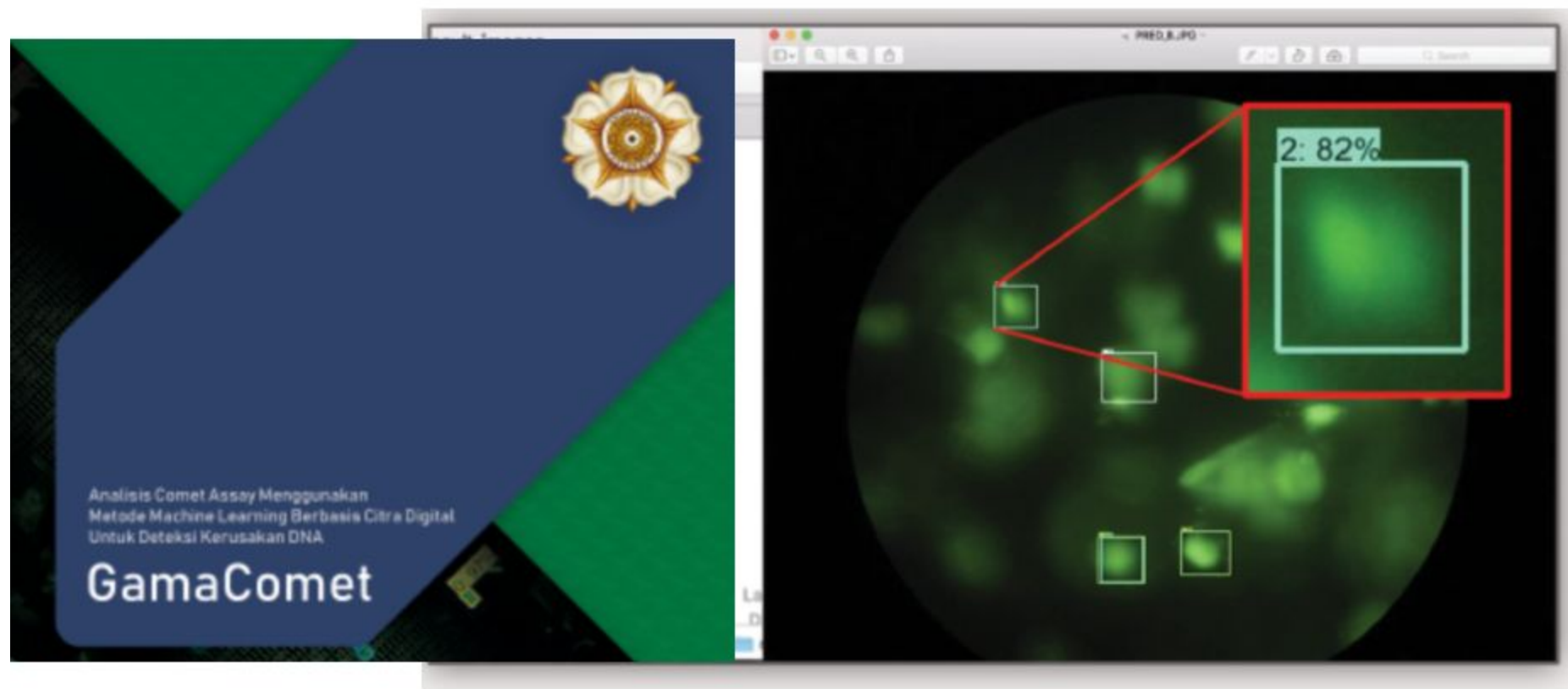


UNIVERSITAS
GADJAH MADA

GAMAComet

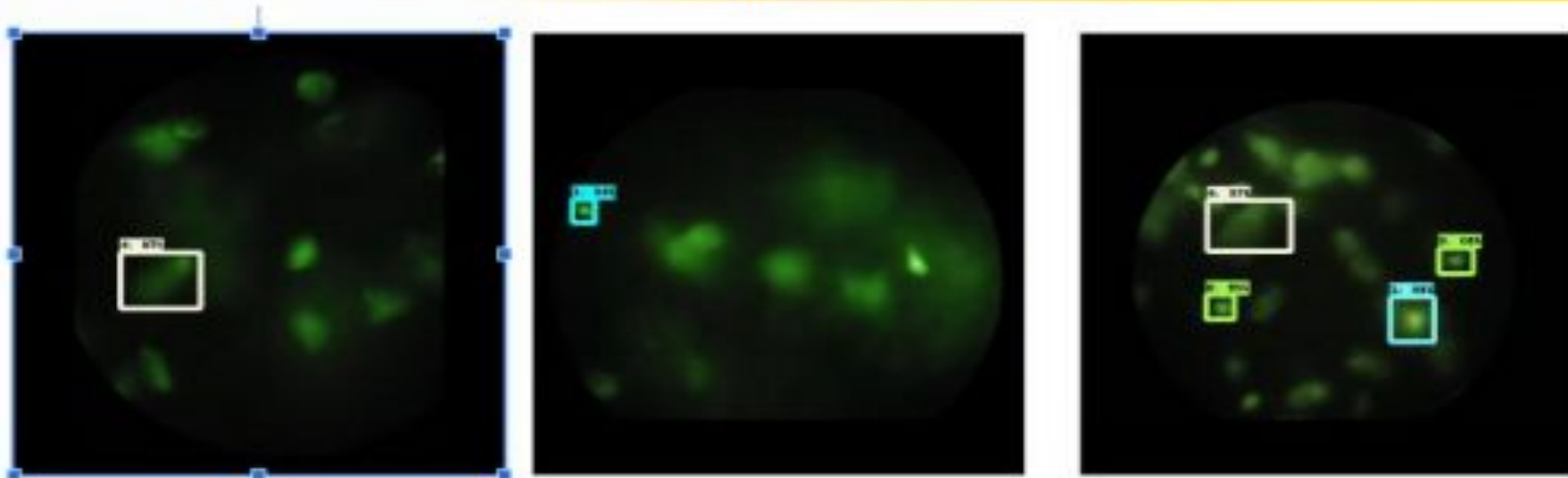
(kerjasama dengan FKG UGM)

Hak Cipta : 000143480

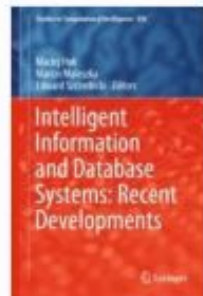


GamaComet :

Analisis Comet Assay Menggunakan Machine Learning Untuk Deteksi Kerusakan DNA



- Done :
 - ✓ Classification using super tiny dataset and transfer learning
 - ✓ Automatic segmentation classification - FasterRCNN
- OnGoing :
 - ✓ Classification using ELM
 - ✓ Automatic segmentation - 2D Otsu Methods
- Next :
 - ✓ Data oversampling
 - ✓ Segmentation - using pixel classification
- GRANT :
 - ✓ PTUPT




[Asian Conference on Intelligent Information and Database Systems](#)

..... ACIIDS 2019: [Intelligent Information and Database Systems: Recent Developments](#) pp 279-289 | [Cite as](#)

Comet Assay Classification for Buccal Mucosa's DNA Damage Measurement with Super Tiny Dataset Using Transfer Learning

Authors

[Authors and affiliations](#)

Afiahayati , Edgar Anarossi, Ryna Dwi Yanuarieska, Fajar Ulin Nuha, Sri Mulyana

Chapter

First Online: 06 March 2019

248

Downloads

Part of the [Studies in Computational Intelligence](#) book series (SCI, volume 830)

Abstract

Paper GamaComet (1)

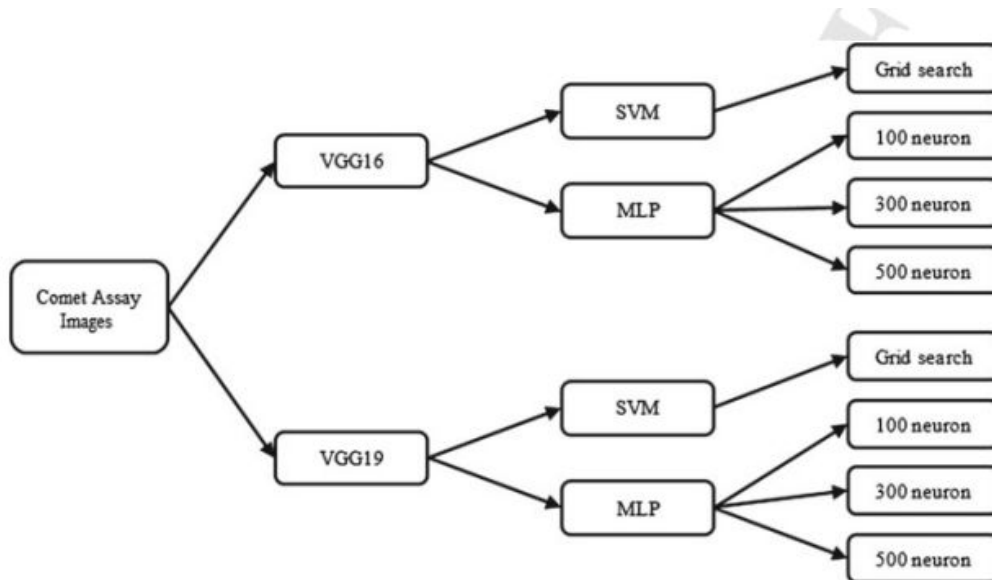


Fig. 5 Classification process using transfer learning

Table 6 Results of transfer learning using MLP and SVM as classification methods with VGG16 as a feature extractor

Data/architecture	SVM	MLP 100	MLP 300	MLP 500
No augmentation	0.6202	0.7049	0.7049	0.639
Augmented	0.6272	0.664	0.68	0.604

Table 7 Results of MLP and SVM classification with VGG19 as a feature extractor

Data/architecture	SVM	MLP 100	MLP 300	MLP 500
No augmentation	0.6238	0.6557	0.623	0.623
Augmented	0.6356	0.692	0.668	0.56

Table 8 The accuracy comparison of the overall classification model and OpenComet software [15]

Data/architecture	CNN	CNN-SVM	Transfer learning 1 (VGG16-MLP)	Transfer learning 2 (VGG19-MLP)	OpenComet
No augmentation	0.635	0.623	0.705	0.656	0.115
Augmented	0.519	0.568	0.68	0.692	–

Several potential research topics in Indonesia



UNIVERSITAS
GADJAH MADA

- Bioinformatics basic tasks
- Indonesia's Cancer analysis
- Indonesia's NGS analysis
- COVID19 Genome Analysis
- Machine learning in bioinformatics : Protein Secondary Structure Prediction, ...

Youtube Channel : AfiaKenkyu Please subscribe! :)



UNIVERSITAS
GADJAH MADA

The screenshot shows the YouTube channel page for 'AfiaKenkyu', which has 143 subscribers. The channel is categorized under 'VIDEOS'. The 'Uploads' section displays a grid of 10 video thumbnails with their respective titles and view counts:

- Fuzzy C Means Clustering (Part 2)**: 10 views • 18 hours ago. Video duration: 14:02.
- Fuzzy C-Means Clustering (Part 1)**: 14 views • 19 hours ago. Video duration: 14:41.
- Contoh Kasus Neural Network - Multi Layer...**: 34 views • 3 days ago. Video duration: 13:18.
- Contoh Kasus Linear Classifier - Neural Network - ...**: 47 views • 3 days ago. Video duration: 13:17.
- Neural Network - MLP - Part 2**: 174 views • 1 week ago. Video duration: 13:37.
- Neural Network - Multi Layer Perceptron - Part 1**: Video duration: 12:39.
- K-means Clustering Algorithm - Part 3**: Video duration: 13:08.
- K-means Clustering Algorithm - Part 2**: Video duration: 13:28.
- K-means Clustering Algorithm - Part 1**: Video duration: 14:33.
- Robot Pembersih Daun, Miraikan, Odaiba, Tokyo**: Video duration: 0:31.



UNIVERSITAS
GADJAH MADA

Thank you for your attention!
Maturnuwun!

Assembly with de Bruijn graph



Original sequence

ATGCGAGGCGTGG

ATGCG

GCGAG

AGGCG

CGTGG

reads



k-mers (*k*=3)

ATG TGC GCG CGA

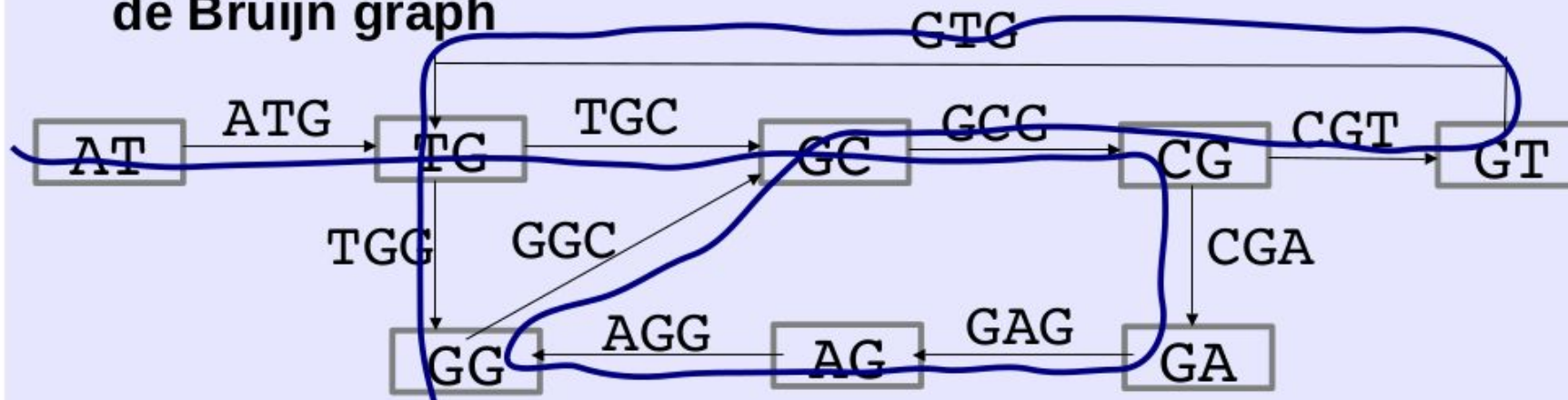
GAG AGG GGC CGT

GTG TGG

ATG → AT, TG
k-mer (k-1)-mer x 2



de Bruijn graph



ATGCGAGGCGTGG