Hello students unit IV bioinformatics

databases and their sequencing

module name biological databases and

their classification format module.

20 myself Ms. Shreeveni Tari outline includes

definition of biological databases.

Features of biological databases,

classification format

of biological databases.

Learning outcomes defines the

term biological database,

explains features of biological databases.

Differentiates the classification

format for biological databases.

CITES examples of biological databases.

Now when we're dealing with

the molecular biology studies,

we are studying three molecules.

We're studying DNA,

RNA and proteins.

And we are studying the sequence

structure and function and

further these are utilized for

the studies related to evolution,

mutation,

metabolic pathways and so now the

data that is generated is very huge,

so we are utilizing bioinformatics

as affairs that will be helping

in storing and retriving of

this biological information.

Now that is stored in the form of

biological databases so we have.

A biological database as a large

organized body of persistent data

associated with computerized

software designed to update,

query and retrieve components of

the data stored within the system.

The chief objective of this biological

databases is to organize data in a

set of structured records to enable

easy retrieval of information.

We need these biological databases

for storing and communicating

large data sets to make biological

data available to the scientists

and also to make biological data

available in computer readable.

Form now we are using this

biological databases for storing,

maintaining, entering data,

searching, sorting,

retriving or presenting or

displaying the biological data.

Now the properties of the biological.

Databases include easy search,

easy to understand, cross referenced,

and connected to the other databases.

And easy retrieval of data.

There are various systems for

classifying the biological databases.

We can utilize the various

parameters for their classification.

The first parameter includes the data type,

the type of data that a biological

database contains, like for example,

whether it contains the structure,

whether it contains the sequences.

Or whether it contains the

functional information or so.

Then based on the data content,

whether it is referring to the DNA molecule,

whether it is referring to the

protein molecule.

OK,

so the type of data content

that is present for a given

biological database based on that.

Also you can classify the third

is the data source where exactly

the data is coming from,

whether it is coming directly from the lab,

whether it is coming from

the other databases,

whether it is coming from literature.

And so. so based on that.

Also you can classify the

biological databases.

Then some of the biological databases are

only related to some of the organisms.

So based on these organisms you can

also classify this biological databases.

Then some of the biological databases

are maintained by some of the curators.

So based on these maintainers also

you can classify the databases

like some of the main maintainers.

Include NCBI EBI, OK, so based on that.

Also you can classify them based

on the data access,

not all the biological data will

be accessible to the public or

to the scientific community.

Some may need some permissions.

OK, so based on the data access

capacity also we can classify

the biological databases.

The last system includes the database design.

Now there are two main approaches

when it comes to the database.

Designing first is the object related.

Where in the information or the data

collected will be stored in the form

of objects and in turn this object

will be linked to one another.

The second approach is the relational

database approach wherein based on

the links or based on the relation,

the datasets will be stored.

Now the popular classification

that is accepted and followed

includes based on the data source.

So where exactly the data is coming from?

That is mainly deciding the classification.

So based on that we have the three

types we have the primary database we

have the secondary database and we

have the third one that is referred to

as integrated or composite database.

Primary database is also called as

archival database here directly.

Whatever experiments you are doing in lab,

the information or the data is directly

given to the primary database so.

Here they will be having the archives

of raw sequence or structural

data submitted by the scientific

community directly.

So it will be populated

with experimentally derived data

such as nucleotide sequence,

protein sequence,

or the macromolecular structure.

So no more information is added.

Whatever you are getting the results

out of your experiment will be given.

The examples of the primary databases

include GenBank DNA data Bank of

Japan and the European Molecular

Biology Laboratory that are having the

storage of the nucleotide sequences.

Second is the protein databank that

is having the three dimensional

structures of biological macromolecules.

The third is the Array Express

archive at EMBL-EBI and GEO.

At NCBI.

That is having the functional genomics data.

So these are few of the examples

of the primary databases.

Then we have the secondary databases

which are comprising computationally

expressed sequence information

from the primary database.

So here the source of information

is from the primary database.

The amount of computational

processing work will vary.

So some will only have the translated

sequence data that are identified

from the open reading frame in DNA,

and some may have additional

annotations and information apart from

the sequences which may be related

to the higher levels of information

regarding structure and information.

This secondary databases often draw

information from the primary databases

as well as controlled vocabularies

and the scientific literature.

They are highly curated.

And often using a complex combination

of computational algorithms,

and manual analysis and interpretations

to derive new knowledge from

the public record of science,

that is from the raw data examples of

secondary databases include SWISS-PROT,

that is having the sequence annotation,

including structure,

function and protein family assignments.

Then we have UniProt

knowledgebase that is dealing

with sequence and functional

information on proteins.

Then we have SCOP that is

dealing with the classification

of protein structural domains and

we have CATH that is having the

classification of protein structure.

Third is the integrated

or composite database.

Now here the information source

is from both the databases that

is the primary databases as

well as the secondary databases.

So the data will be having the

characteristics of both primary

and the secondary.

So integrated databases offers one stop.

Center for the knowledge extraction.

These databases are more like consortiums,

managing an integrating sources of

information to provide a unified

access to the user's example of

composite database is InterPro,

which is an integrated documentation

resource for protein families,

domains and functional sites which

amalgamates the efforts of PROSITE PRINT,

Pfam and ProDom database projects.

Each of the InterPro entry includes

a functional description.

Annotation, literature references and links.

Back to the relevant member databases.

Now these are few of the databases

which are in various categories.

We have the first category that

is of literature wherein we have

a database name Pub Med that is

dealing with the scientific and

medical abstract and citation.

So the literature related files you

will get in the Pub Med database next

is referring to the health wherein we

have the example as OMIM that is

Online Mendelian Inheritance in Men,

that is,

having the information about the

genes and the genetic disorders

that are occurring in human beings.

Next we have the nucleotide sequences.

The databases include the nucleotide that

will be having the sequences for DNA and RNA.

Next,

we have.

The genomes where in Genome and dbSNP.

These databases are dealing

with the genome related information,

then we have the genes wherein

we have the examples as protein

that will be having the protein

sequences and UniProt that will

be having the protein sequences

and the related information.

Then we have chemical related

database named as PubChem compound,

wherein the chemical information with

reference to structure and the other

informational links will be provided.

Then we have pathway related databases,

for example,

Biosystems that is dealing with the

molecular pathways with links to

genes and proteins and we have KEGG

pathway that is dealing with the

information on biological pathways.

We also have Organism related database.

Like for example we have the flybase.

That is,

referring to the genes and genomic

data of the Drosophila species.

We also have Saccharomyces genome

database that is dealing with the

information of the genes and genome

of the Saccharomyces species.

Then we also have the taxonomy related

database that will be maintained for

each of the biological taxa. Now.

These are few of the popular biological

databases which you can utilize,

or which you can study and they are used.

In the application of Bioinfomatics and

various other fields also you can study.

These are few of the references in

the form of books and the web sources.

Thank you.