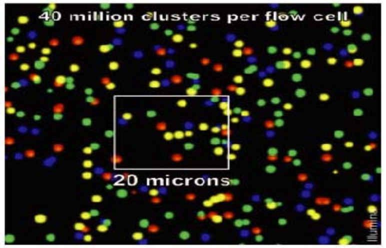
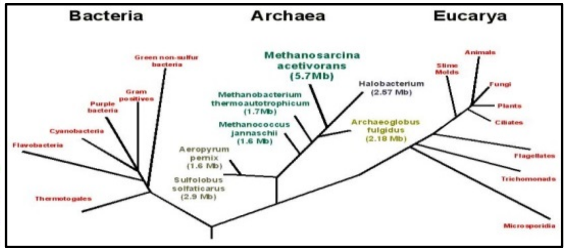
生物信息学：导论与方法

Bioinformatics: Introduction and Methods

https://www.coursera.org/course/pkubioinfo

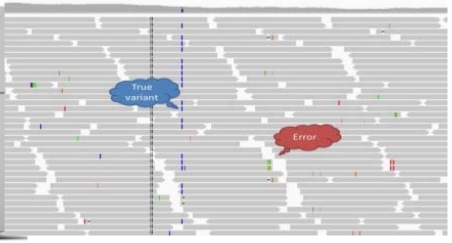
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA

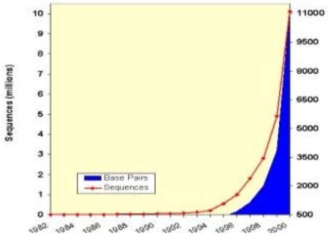


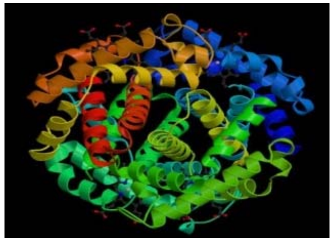

Bacteria
Archaea
Eucarya
Green non-sulfur bacteria
Gram positives
Purple bacteria
Cyanobacteria
Flavobacteria
Thermotogales
Methanosarcina acetivorans (5.7Mb)
Methanobacterium thermoautotrophicum (1.7Mb)
Methanococcus jannaschii (1.6 Mb)
Halobacterium (2.57 Mb)
Archaeoglobus fulgidus (2.18 Mb)
Aeropyrum pernix (1.6 Mb)
Sulfolobus solfataricus (2.9 Mb)
Slime Molds
Animals
Fungi
Plants
Ciliates
Flagellates
Trichomonads
Microsporidia

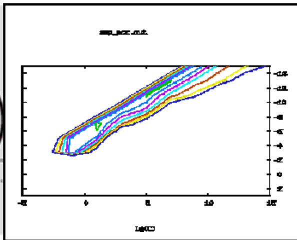
Ontology, and Identification of Molecular Pathways

Supplementary Learning Materials

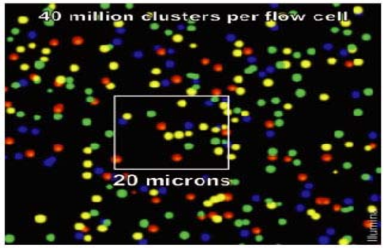


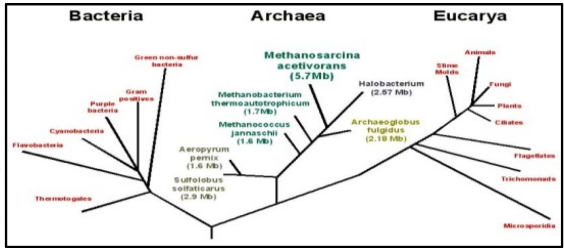






```
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA
```



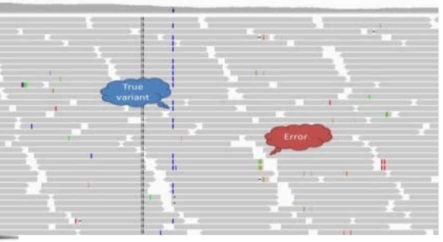

Bacteria
Archaea
Eucarya
Green non-sulfur bacteria
Gram positives
Purple bacteria
Cyanobacteria
Flavobacteria
Thermotogales
Methanosarcina acetivorans (5.7Mb)
Halobacterium (2.57 Mb)
Methanobacterium thermoautotrophicum (1.7Mb)
Methanococcus jannaschii (1.6 Mb)
Archaeoglobus fulgidus (2.18 Mb)
Aeropyrum pernix (1.6 Mb)
Sulfolobus solfaticarus (2.9 Mb)
Animals
Slime Molds
Fungi
Plants
Ciliates
Flagellates
Trichomonads
Microsporidia

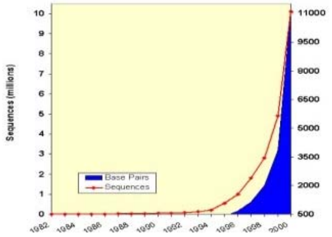
Brief Introduction to Database

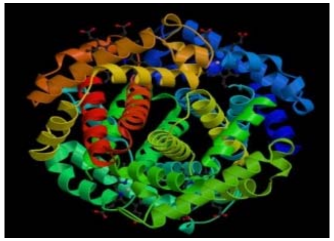
北京生命科学研究所 谢忱

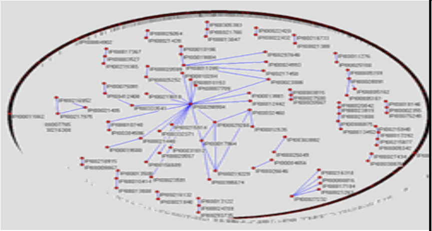
Chen Xie, Ph.D.

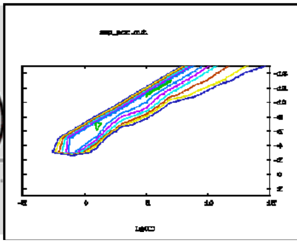
National Institute of Biological Sciences, Beijing











What is database?

Database is the collection of data

Database management system (DBMS) is the collection of interrelated data and a set of programs to access those data

DBMS provides a efficient, reliable, convenient and safe multi-user storage of and access to massive amounts of persistent data

Why do people use DBMS?

Major disadvantages of file-processing system

- Data redundancy and inconsistency
- Difficulty in accessing data
- Data isolation
- Integrity problems
- Atomicity problems
- Concurrent-access anomalies
- Security problems

Data models

Relational model

Entity-relationship model

Object-based data model

Semistructured data model

Relational model

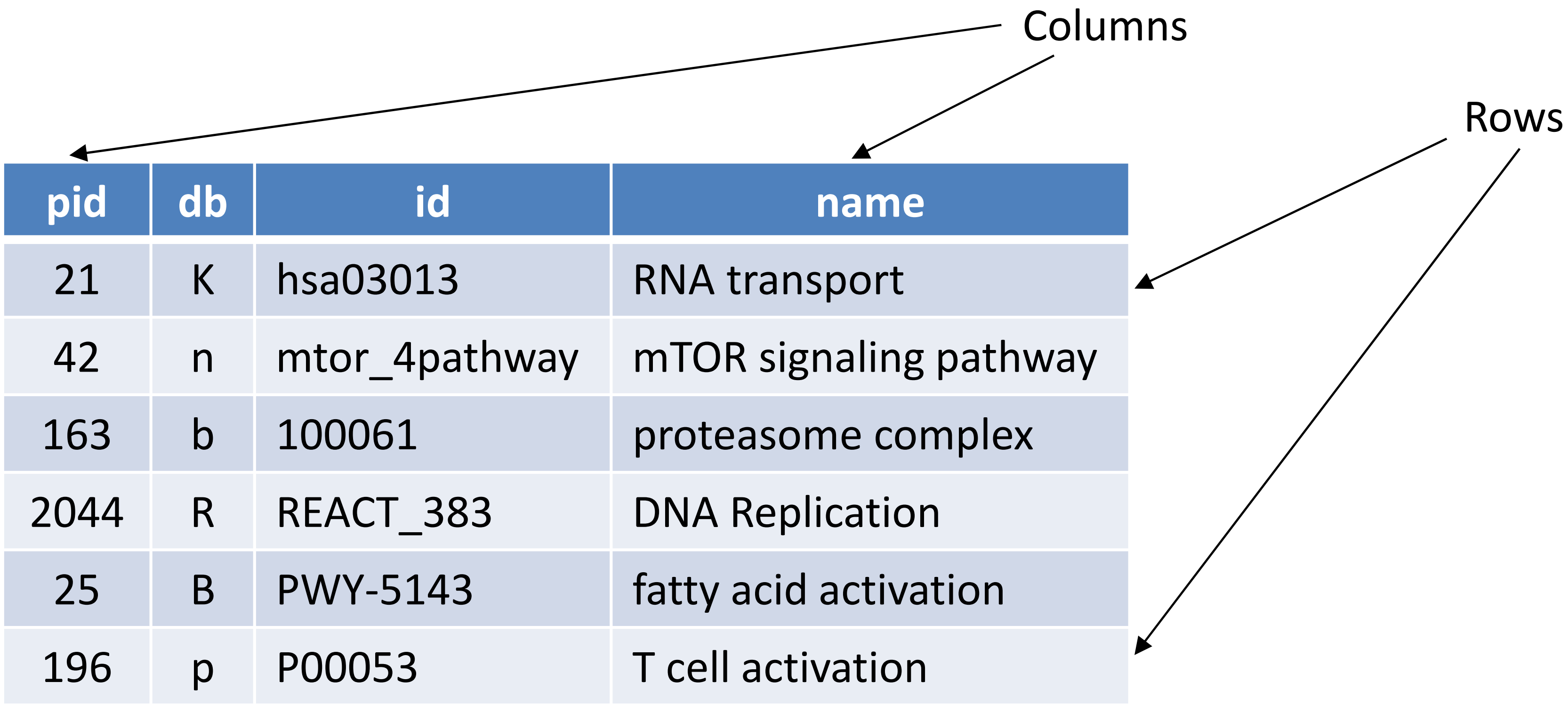
Database is a set of named relations (or tables)

Each relation has a set of named attributes (or columns)

Each tuple (or row) has a value for each attribute

Each attribute has a type (or domain)

An example for relational model



pid	db	id	name
21	K	hsa03013	RNA transport
42	n	mtor_4pathway	mTOR signaling pathway
163	b	100061	proteasome complex
2044	R	REACT_383	DNA Replication
25	B	PWY-5143	fatty acid activation
196	p	P00053	T cell activation

Table Pathways

Key

Column whose value is unique in each row

pid in Table Pathways

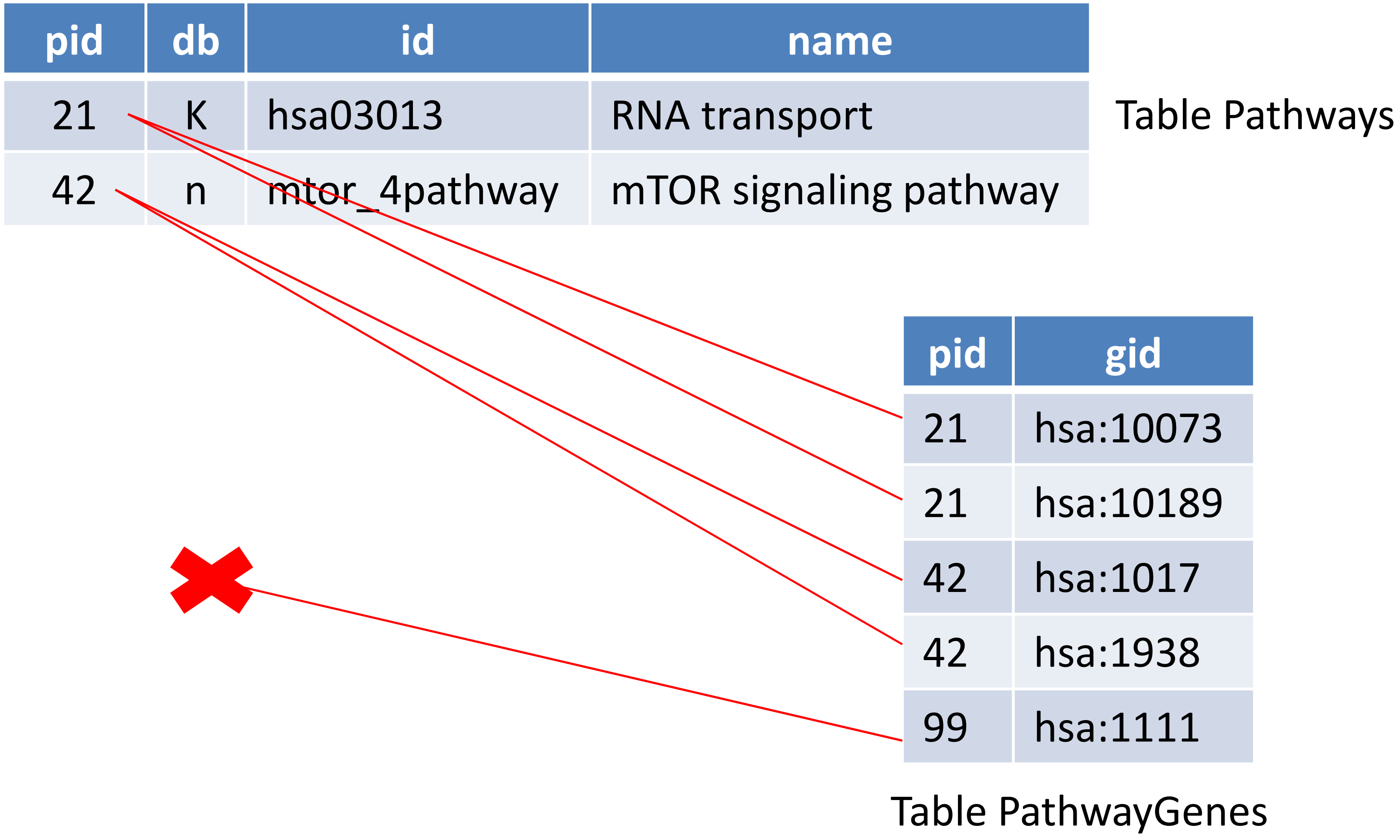
Set of columns whose combined values are unique

(pid, gid) in Table PathwayGenes

pid	gid
21	hsa:10073
21	hsa:10189
42	hsa:1017
42	hsa:1938
99	hsa:1111

Table PathwayGenes

Referential integrity



pid	db	id	name
21	K	hsa03013	RNA transport
42	n	mtor_4pathway	mTOR signaling pathway

Table Pathways

pid	gid
21	hsa:10073
21	hsa:10189
42	hsa:1017
42	hsa:1938
99	hsa:1111

Table PathwayGenes

Database Languages

Data-Definition Language (DDL)

Data-Manipulation Language (DML)

SQL for DDL

```
CREATE TABLE Pathways
(
    pid     INTEGER    PRIMARY KEY,
    db      TEXT,
    id      TEXT,
    name    TEXT
);
```

SQL for DML

Find the name of the pathway with pid 21

```
SELECT name
FROM Pathways
WHERE pid = 21;
```

Find the name of all pathways having the gene with gid hsa:1017

```
SELECT Pathways.name
FROM Pathways, PathwayGenes
WHERE Pathways.pid = PathwayGenes.pid AND PathwayGenes.gid =
‘hsa:1017’;
```

Open source database softwares

MySQL

SQLite

PostgreSQL

References and further reading

A. Silberschatz, H. Korth, S. Sudarshan. Database System Concepts, 6th edition. New York. McGraw-Hill. 2011.

J. Widom. Introduction to Databases.

https://www.coursera.org/course/db

生物信息学：导论与方法

Bioinformatics: Introduction and Methods

https://www.coursera.org/course/pkubioinfo