cDNA libraries and Bioinformatics- tools of bioinformatics, data bases and data base management,

Lanchana H.A
BOT-CPT –4.2
Guest Faculty
DOSR in Botany
Tumkur University

DNA (Gene) Libraries:

- A DNA library is a set of cloned fragments that collectively represent the genes of a particular organism. Particular genes can be isolated from DNA libraries, much as books can be obtained from conventional libraries.
- There are two general types of gene library: a genomic library, which consists of the total chromosomal DNA of an organism; and a cDNA library, which represents the mRNA from a cell or tissue at a specific point of time.
- The choice of the particular type of gene library depends on a number of factors, the most important being the final application of any DNA fragment derived from the library. If the ultimate aim understands the control of protein production for a particular gene or its architecture, then genomic libraries must be used.

- However, if the goal is the production of new or modified proteins, or the determination of tissue-specific expression of timing patterns, cDNA libraries are more appropriate.
- In contrast, however, cDNA libraries represent only mRNA being produced from a specific cell type at a particular time in the cell's development. Thus, it is important to consider carefully the cell or tissue type from which the mRNA is to be deriver in the construction of cDNA libraries.
- There are a variety of cloning vectors available, many based on naturally occurring molecules such as bacterial plasmids or bacteria-infecting viruses. The choice of vector also depends on whether a genomic library or cDNA library is constructed.

Constructing Gene Libraries:

- **Digesting Genomic DNA Molecules:** After genomic DNA has been isolated and purified, it is digested with restriction endonucleases. Important to note that every copy of a given DNA molecule from a specific organism will give the same set of fragments when digested with a particular enzyme.
- DNA from different organisms will, in general, give different sets of fragments when treated with the same enzyme. By digesting complex genomic DNA from an organism it is possible to reproducibly divide its genome into a large number of small fragments, each approximately the size of a single gene. Some enzymes cut straight across the DNA to give flush or blunt ends.

Ligating DNA Molecules:

- The DNA products resulting from restriction digestion to form sticky ends may be joined to any other DNA fragments treated with the same restriction enzyme. Thus, when the two sets of fragments are mixed; base-pairing between sticky ends will result in the annealing of fragments that were derived from different starting DNA.
- Each DNA fragment is inserted by ligation into vector DNA molecule, which allows the whole recombinant DNA to then be replicated indefinitely within microbial cells
- Thus, all of the DNA extracted from an organism and digested with a restriction enzyme will result in a collection of clones. This collection of clones is known as a gene library.

cDNA Libraries:

- cDNAs are DNA molecules copied from mRNA templates. cDNA libraries are constructed by synthesizing cDNA from purified cellular mRNA.
- These libraries present an alternative strategy for gene isolation, especially eukaryotic genes. Because most eukaryotic mRNAs carry 3′-poly(A) tails, mRNA can be selectively isolated from preparations of total cellular RNA by oligo(dT)-cellulose chromatography

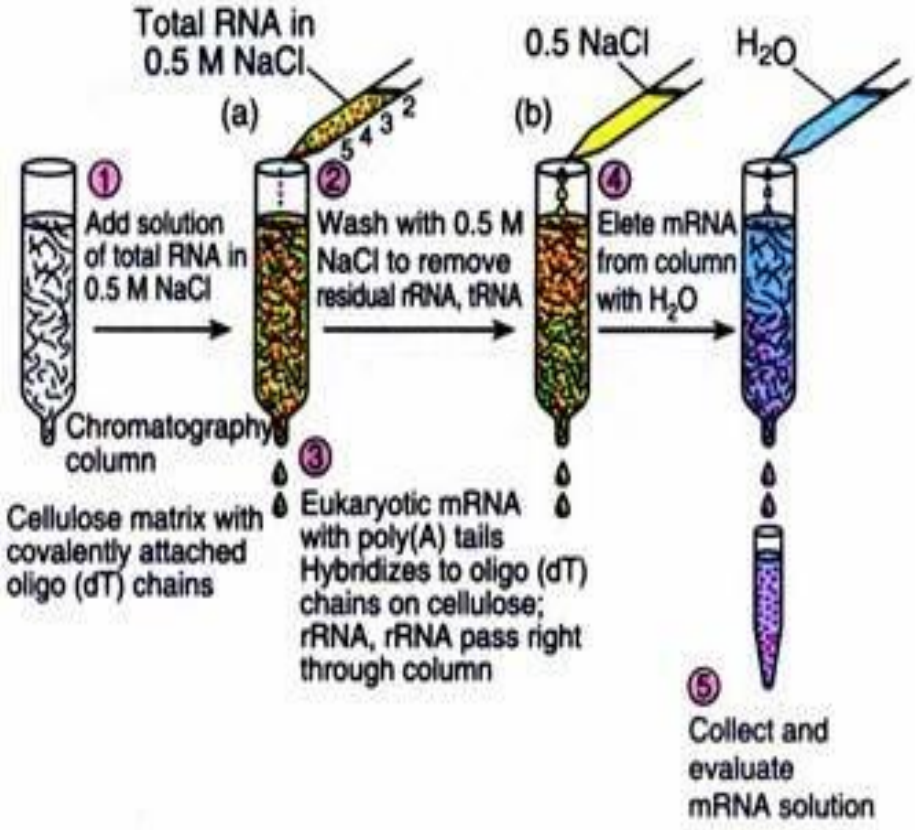


Fig. 4.12: Isolation of eukaryotic mRNA via oligo(dT)-cellulose chromatography. (a) In the presence of 0.5 M NaCl, the poly(A) tails of eukaryotic mRNA anneal with short oligo(dT) chains covalently attached to an insoluble chromatographic matrix such as cellulose. Other RNAs, such as rRNA (green), pass right through the chromatography column. (b) The column is washed with more 0.5 M NaCl to remove residual contaminants. (c) Then the poly(A) mRNA is recovered by washing the column with water because the base pairs formed between the poly(A) tails of the mRNA and the oligo(dT) chains are unstable in solutions of low ionic strength.

- DNA copies of the purified mRNAs are synthesized by first annealing short oligo (dT) chains to the poly(A) tails.
- These oligo(dT) chains serve as primers for reverse transcriptase-driven synthesis of DNA .
- (Random oligonucleotides can also be used as primers, with the advantages being less dependency on poly(A) tracts and increased likelihood of creating clones representing the 5′-ends of mRNAs.) Reverse transcriptase is an enzyme that synthesizes a DNA strand, copying RNA as the template. DNA polymerase is then used to copy the DNA strand and form a double-stranded (duplex DNA) molecule.

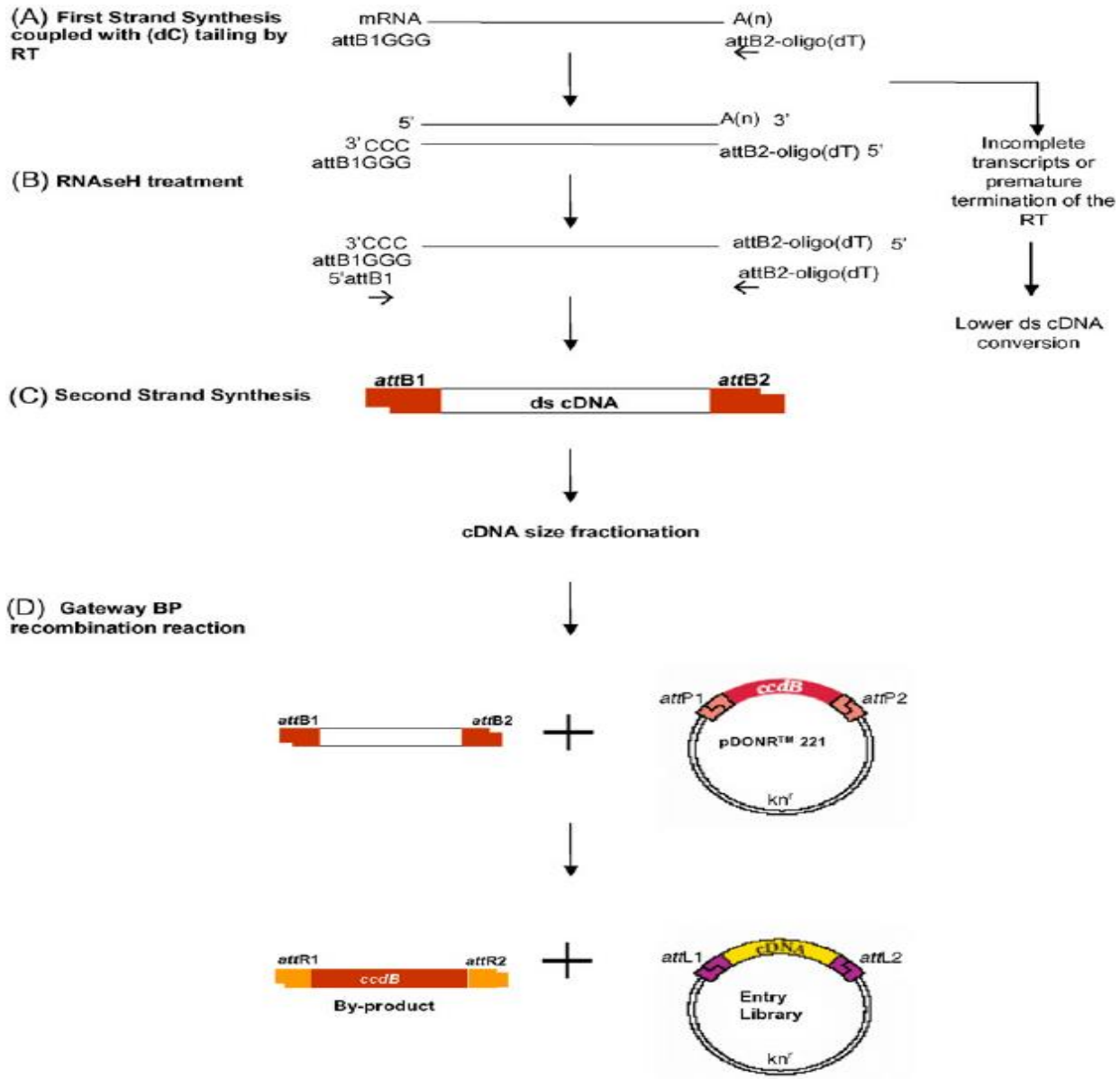
- Ligation of blunt-ended DNA fragments is not as efficient as ligation of sticky ends; therefore, with cDNA molecules additional procedures are undertaken before ligation with cloning vectors.
- One approach is to add cDNA small, double stranded molecules with one internal site for a restriction endonuclease; these are termed nucleic acid linkers. Numerous linkers are commercially available with internal restriction for many of the most commonly used restriction enzymes.

- Linkers are blunt end ligated to cDNA but since they are added much in excess of the cDNA, the ligation process is reasonably successful. Subsequently the linkers are digested with the appropriate restriction enzyme, which provides the sticky ends for efficient ligation to a vector digested with the same enzyme.
- This process may be made easier by the addition of adaptors rather than linkers, which are identical except that the sticky ends are performed and so there is no need of restriction digestion following ligation.

- Therefore, lastly Linkers are added to the DNA duplexes rendered from the mRNA templates, and the cDNA is cloned into a suitable vector. Once a cDNA derived from a particular gene has been identified, the cDNA becomes an effective probe for screening genomic libraries for isolation of the gene itself.

- Because different cell types in eukaryotic organisms express selected subsets of genes, RNA preparations from cells or tissues in which genes of interest are selectively transcribed are enriched for the desired mRNAs. cDNA libraries prepared from such mRNA are representative of the pattern and extent of gene expression that uniquely define particular kinds of differentiated cells.

- cDNA libraries of many normal and diseased human cell types are commercially available, including cDNA libraries of many tumour cells. Comparison of normal and abnormal cDNA libraries, in conjunction with two dimensional gel electrophoretic analysis of the proteins produced in normal and abnormal cells is a promising new strategy in clinical medicine to understand disease mechanisms.


(A) First Strand Synthesis coupled with (dC) tailing by RT
mRNA
attB1GGG
A(n)
attB2-oligo(dT)
5'
A(n) 3'
3' CCC
attB1GGG
attB2-oligo(dT) 5'
Incomplete transcripts or premature termination of the RT
(B) RNAseH treatment
3'CCC
attB1GGG
5'attB1
attB2-oligo(dT) 5'
attB2-oligo(dT)
Lower ds cDNA conversion
(C) Second Strand Synthesis
attB1
ds cDNA
attB2
cDNA size fractionation
(D) Gateway BP recombination reaction
attB1
attB2
attP1
ccdB
attP2
pDONR™ 221
kn^r
attR1
attR2
ccdB
By-product
attL1
cDNA
attL2
Entry Library
kn^r

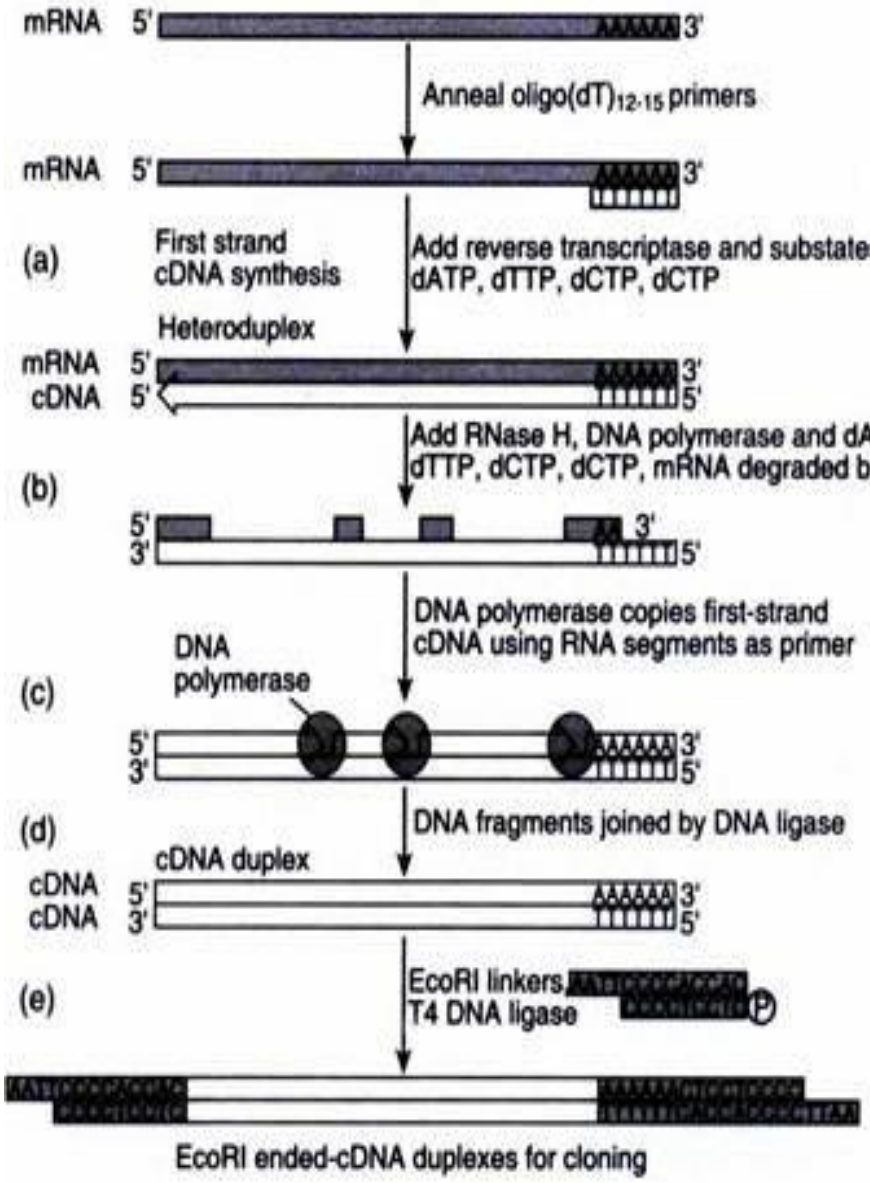


Fig. 4.13: Reverse transcriptase–driven synthesis of cDNA from oligo(dT) primers annealed to the poly(A) tails of purified eukaryotic mRNA. (a) Oligo(dT) chains serve as primers for synthesis of a DNA copy of the mRNA by reverse transcriptase. Following completion of first-strand cDNA synthesis by reverse transcriptase, RNase H and DNA polymerase are added (b). RNase H specifically digests RNA strands in DNA:RNA hybrid duplexes. DNA polymerase copies the first-strand cDNA, using as primers the residual RNA segments after RNase H has created nicks and gaps (c). DNA polymerase has a 5′→3′ exonuclease activity that removes the residual RNA as it fills in with DNA. The nicks remaining in the second-strand DNA are sealed by DNA ligase (d), yielding duplex cDNA. *Eco*RI adapters with 5′-overhangs are then ligated onto the cDNA duplexes (e) using phage T4 DNA ligase to create *Eco*RI ended cDNA for insertion into a cloning vector.

Bioinformatics

- Bioinformatics: The sum of the computational approaches to analyze, manage, and store biological data. Bioinformatics involves the analysis of biological information using computers and statistical techniques, the science of developing and utilizing computer databases and algorithms to accelerate and enhance biological research.
- Bioinformatics is used in analyzing genomes, proteomes (protein sequences), three-dimensional modeling of biomolecules and biologic systems, etc.
- There are data-mining software that retrieve data from genomic sequence databases and also visualization tools to analyze and retrieve information from proteomic databases. These can be classified as homology and similarity tools, protein functional analysis tools, sequence analysis tools and miscellaneous tools.

Databases and Database Management

- A **database** is a computerized archive used to store and organize data in such a way that information can be retrieved easily via a variety of search criteria.
- Databases are composed of computer hardware and software for data management.
- The chief objective of the development of a database is to organize data in a set of structured records to enable easy retrieval of information.

- Each record, also called an **entry**, should contain a number of fields that hold the actual data items, for example, fields for names, phone numbers, addresses, dates.
- To retrieve a particular record from the database, a user can specify a particular piece of information, called **value**, to be found in a particular field and expect the computer to retrieve the whole data record.
- This process is called **making a query.**

TYPES OF DATABASES

- Originally, all databases used a **flat file format**, which is a long text file that contains many entries separated by a delimiter, a special character such as a vertical bar (|). Within each entry are a number of fields separated by tabs or commas.

- Except for the raw values in each field, the entire text file does not contain any hidden instructions for computers to search for specific information or to create reports based on certain fields from each record.

- **The text file can be considered a single table**.
- To search a flat file for a particular piece of information, a computer has to read through the entire file, an obviously inefficient process which is manageable for a small database, but as database size increases or data types become more complex, this database style can become very difficult for information retrieval.
- Searches through such files often cause crashes of the entire computer system because of the memory-intensive nature of the operation.

- To facilitate the access and retrieval of data, sophisticated computer software programs for organizing, searching, and accessing data have been developed which are called **database management systems**.

- These systems contain not only raw data records but also operational instructions to help identify hidden connections among data records.

- The purpose of establishing a data structure is for easy execution of the searches and to combine different records to form final search reports.

- Depending on the types of data structures, these database management systems can be classified into **two types**: **relational database management systems** and **object-oriented database management systems**.
- Databases employing these management systems are known as relational databases or object-oriented databases, respectively.

Relational Databases

- Instead of using a single table as in a flat file database, relational databases use a set of tables to organize data.
- Each table, also called a **relation**, is made up of **columns** and **rows**. Columns represent individual fields. Rows represent values in the fields of records.
- The columns in a table are indexed according to a common feature called an **attribute**, so they can be cross-referenced in other tables.

- To execute a query in a relational database, the system selects linked data items from different tables and combines the information into one report.
- Therefore, specific information can be found more quickly from a relational database than from a flat file database.

Flat File

Name, States, Course number, Course name|John Smith, Texas, Biol 689, Bioinformatics|Jane Doe, Kansas, Bich 441, Biochemistry|William Brown, Illinois, Chem 289, Organic Chemistry|Jennifer Taylor, New York, Hort 201, Horticulture|Howard Douglas, Texas, Math 172, Calculus

Table A

Student #	Name	State
1	John Smith	Texas
2	Jane Doe	Kansas
3	William Brown	Illinois
4	Jennifer Taylor	New York
5	Howard Douglas	Texas

Table B

Student #	Course #
1	Biol 689
2	Bich 441
3	Chem 289
4	Hort 201
5	Math 172

Table C

Course #	Course name
Biol 689	Bioinformatics
Bich 441	Biochemistry
Chem 289	Organic chemistry
Hort 201	Horticulture
Math 172	Calculus

Example of constructing a relational database for five students' course information originally expressed in a flat file. By creating three different tables linked by common fields, data can be easily accessed and reassembled.

- Example; student course information expressed in a flat file which contains records of five students from four different states, each taking a different course.
- Each data record, separated by a vertical bar, contains four fields describing the name, state, course number and title.
- A relational database is also created to store the same information, in which the data are structured as a number of tables.
- In each table, data that fit a particular criterion are grouped together and different tables can be linked by common data categories, which facilitate finding of specific information.

- For example, if one is to ask the question, which courses are students from Texas taking? The database will first find the field for “State” in Table A and look up for Texas. This returns students 1 and 5.
- The student numbers are colisted in Table B, in which students 1 and 5 correspond to Biol 689 and Math 172, respectively.
- The course names listed by course numbers are found in Table C. By going to Table C, exact course names corresponding to the course numbers can be retrieved.
- A final report is then given showing that the Texans are taking the courses Bioinformatics and Calculus.

- However, executing the same query through the flat file requires the computer to read through the entire text file word by word and to store the information in a temporary memory space and later mark up the data records containing the word *Texas.*
- This is easily accomplishable for a small database. To perform queries in a large database using flat files obviously becomes an onerous task for the computer system.

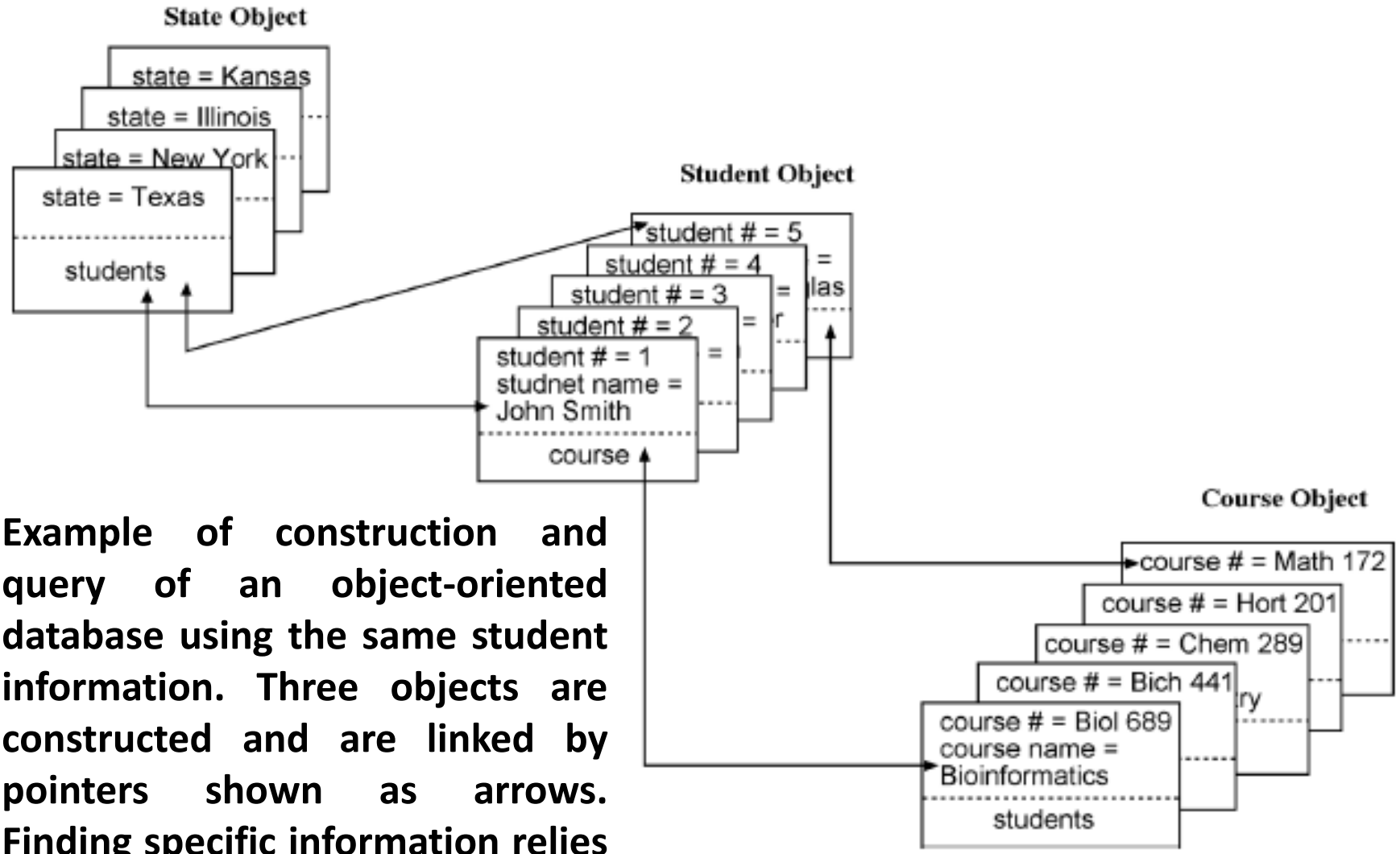
Object-Oriented Databases

- One of the problems with relational databases is that the tables used do not describe complex hierarchical relationships between data items.
- To overcome the problem, object-oriented databases have been developed that store data as objects.
- In an object-oriented programming language, an object can be considered as a unit that combines data and mathematical routines that act on the data.

- The database is structured such that the objects are linked by a set of pointers defining predetermined relationships between the objects.
- Searching the database involves navigating through the objects with the aid of the pointers linking different objects.
- The object-oriented database system is more flexible; data can be structured based on hierarchical relationships.

- However, this type of database system lacks the rigorous mathematical foundation of the relational databases and there is also a risk that some of the relationships between objects may be misrepresented.

- Some current databases have therefore incorporated features of both types of database programming, creating the **object–relational database management system**.



Example of construction and query of an object-oriented database using the same student information. Three objects are constructed and are linked by pointers shown as arrows. Finding specific information relies on navigating through the objects by way of pointers.

- The students' course information can be used to construct an object-oriented database. Three different objects can be designed: student object, course object, and state object. Their interrelations are indicated by lines with arrows.
- To answer the same question – which courses are students from Texas taking – one simply needs to start from Texas in the state object, which has pointers that lead to students 1 and 5 in the student object. Further pointers in the student object point to the course each of the two students is taking.
- Therefore, a simple navigation through the linked objects provides a final report.

BIOLOGICAL DATABASES

- Current biological databases use all three types of database structures: flat files, relational, and object oriented.
- Despite the obvious drawbacks of using flat files in database management, many biological databases still use this format.
- The justification for this is that this system involves minimum amount of database design and the search output can be easily understood by working biologists.

- Based on their contents, biological databases can be roughly divided into three categories: **primary databases**, **secondary databases**, and **specialized databases**.
- **Primary databases** contain original biological data.
- They are archives of raw sequence or structural data submitted by the scientific community. **GenBank** and **Protein Data Bank** (PDB) are examples of primary databases.

- **Secondary databases** contain computationally processed or manually curated information, based on original information from primary databases.

- Translated protein sequence databases containing functional annotation belong to this category. Examples are SWISS-Prot and Protein Information Resources (PIR).

- **Specialized databases** are those that cater to a particular research interest. Examples TAIR, Flybase, ect.

Primary Databases

- There are three major public sequence databases that store raw nucleic acid sequence data produced and submitted by researchers worldwide: **GenBank, the European Molecular Biology Laboratory (EMBL) database** and the **DNA Data Bank of Japan** (DDBJ), which are all freely available on the Internet.
- Most of the data in the databases are contributed directly by authors with a minimal level of annotation.

- These three public databases closely collaborate and exchange new data daily. They together constitute the International Nucleotide Sequence Database Collaboration.
- This means that by connecting to any one of the three databases, one should have access to the same nucleotide sequence data.
- Although the three databases all contain the same sets of raw data, each of the individual databases has a slightly different kind of format to represent the data.

- For the three-dimensional structures of biological macromolecules, there is only one centralized database, the PDB.
- This database archives atomic coordinates of macromolecules (both proteins and nucleic acids) determined by x-ray crystallography and NMR.
- It uses a flat file format to represent protein name, authors, experimental details, secondary structure, cofactors, and atomic coordinates.

Secondary Databases

- Secondary databases contain computationally processed sequence information derived from the primary databases.
- The amount of computational processing work varies greatly among the secondary databases; some are simple archives of translated sequence data from identified open reading frames in DNA, whereas others provide additional annotation and information related to higher levels of information regarding structure and functions.

- A prominent example of secondary databases is SWISS-PROT, which provides detailed sequence annotation that includes structure, function, and protein family assignment.

- The sequence data are mainly derived from TrEMBL, a database of translated nucleic acid sequences stored in the EMBL database. The annotation of each entry is carefully curated by human experts and thus is of good quality.

- The protein annotation includes function, domain structure, catalytic sites, cofactor binding, posttranslational modification, metabolic pathway information, disease association, and similarity with other sequences.

- A recent effort to combine SWISS-PROT, TrEMBL, and PIR led to the creation of the **UniProt database**, which has larger coverage than any one of the three databases while at the same time maintaining the original SWISS-PROT feature of low redundancy, cross-references, and a high quality of annotation.

Specialized Databases

- Specialized databases normally serve a specific research community or focus on a particular organism and the content of these databases may be sequences or other types of information.
- The sequences in these databases may overlap with a primary database, but may also have new data submitted directly by authors and because they are often curated by experts in the field, they may have unique organizations and additional annotations associated with the sequences.

- Many genome databases that are taxonomic specific fall within this category. Examples include Flybase, WormBase, AceDB, and TAIR.

- In addition, there are also specialized databases that contain original data derived from functional analysis. For example, GenBank EST database and Microarray Gene Expression Database at the European Bioinformatics Institute (EBI) are some of the gene expression databases available.