Medical Informatics

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www.yildiz.edu.tr/~naydin

1

Databases

2

Databases

- Computers serve four interdependent functions in biomedical informatics:
  - communications,
  - computation,
  - control,
  - storage
- Embedded computer controllers in sequencing machines, fermentation tanks, and bioreactors direct the programmable robotic arms that automate intricate processes and markedly decrease the need for human operators.

3

Databases

- When time is of the essence, computer-controlled devices are superior to manual operations,
  - in part because they can operate virtually unattended around the clock
- As a communications device,
  - not only has the computer helped researchers craft more journal articles in less time than at any other point in history,
  - but an increasingly large proportion of academic research information appears online.

4

Databases

- As computational devices in biomedical informatics, computers are used
  - for tasks that range from searching for nucleotide sequences and visualizing protein folding patterns to simulating complex 3D protein-protein interactions,
  - for applications ranging from drug discovery to biomaterials research and development.

5

Databases

- All of these activities revolve around database technology.
- For example,
  - both communications and computation operations in bioinformatics depend on data that have to be maintained.
- Electronic databases maintain data in a persistent, non-volatile form that allows operations to be repeated and compared with other operations, with the results communicated to other researchers and developers.

6

Databases

- The electronic database
  - a file composed of records,
    - each containing fields together with a set of operations for
      - searching, sorting, recombining, and other functions
  - is the silicon, plastic, and ironoxide equivalent of the experimenter's private notebook, and
  - is the basis for electronic publishing to the scientific community.
- As an illustration of how central databases are to the molecular biology research and development, consider a sampling of the public bioinformatics databases listed in the following table

7

Public Bioinformatics Databases Accessible via the Internet

Database Type	Example	Note
Nucleotide Sequence	GenBank	One of the largest public sequence databases
	DDBJ	DNA DataBank of Japan
	EMBL	European Molecular Biology Laboratory
	MGDB	Mouse Genome Database
	GSX	Mouse Gene Expression Database
	NDB	Nucleic Acid Database
Protein Sequence	SWISS-PROT	Swiss Institute for Bioinformatics and European Bioinformatics Institute
	TrEMBL	Annotated supplement to SWISSPROT
	TrEMBLnew	Weekly, pre-processed update to TrEMBL
	PIR	Protein Information Resource

8

Public Bioinformatics Databases Accessible via the Internet

Database Type	Example	Note
3D Structures	PDB	Protein DataBank
	MMDB	Molecular Modeling Database
	Cambridge Structural Database	For small molecules
Enzymes and Compounds	LIGAND	Chemical compounds and reactions
Sequence Motifs (Alignment)	PROSITE	Sequence motifs
	BLOCKS	Derived from PROSITE
	PRINTSA	superset of BLOCKS
	Pfam	Protein families database of alignments and hidden Markov models
	ProDOM	Protein Domains
Pathways and Complexes	Pathway	Metabolic and regulatory pathway maps

9

Public Bioinformatics Databases Accessible via the Internet

Database Type	Example	Note
Molecular Disease	OMIM	Online Mendelian Inheritance in Man
Biomedical Literature	PubMed	Contains Medline
	Medline	Medical Literature
Vectors	UniVec	Used to identify vector contamination
Protein Mutations	PMD	Protein Mutant Database
Gene Expressions	GEO	Gene Expression Omnibus
Amino Acid Indices	Aaindex	Amino Acid Index Database
Protein/Peptide Literature	LITDB	Literature database for proteins and peptides
Gene Catalog	GENES KEGG	Genes Database

10

Databases

- Database technology is most valuable in the biotech industry when it enables the integration of
  - research,
  - development,
  - clinical activity,
  - manufacturing,
  - selling and marketing.
- Data take on added value when they leave the confines of a workstation and become incorporated into shared public and private databases, applications, and products.

11

Definitions

- Databases, which provide the long-term memory of computer operations, take on a variety of names, depending on their
  - structure,
  - contents,
  - use,
  - amount of data

  they contain.
- Two technologies often confused with databases are disk servers and file servers.

12

Definitions

- A disk server
  - a node in a local area network that acts as a remote disk drive.
  - can be divided into multiple volumes,
    - some of which are shared by all users on the server,
    - others of which can be accessed only by a specific user, as defined by username and password login.
- The file server,
  - can be thought of as a disk server with intelligence.
    - store files, manages the network requests for them and maintains order as users request and modify files.

13

Definitions

- The file server, supports movement and cataloging of files,
  - but, unlike a true database, the contents of a file server are unavailable without the use of some other application.
- With both disk servers and file servers, separate applications must be used to open documents for reviewing and editing.
- In this regard, most disk and file servers work like extensions to the computer operating system.
  - Files can be identified, copied, deleted, and otherwise managed at a very high level.
- File servers and disk servers can be considered as extensions to the internal workstation hard drive that may be configured as a shared volume so that
  - collaborators on the same network can share data stored on the server.

14

Definitions

- The file server, supports movement and cataloging of files,
  - but, unlike a true database, the contents of a file server are unavailable without the use of some other application.
- With both disk servers and file servers, separate applications must be used to open documents for reviewing and editing.
- In this regard, most disk and file servers work like extensions to the computer operating system.
  - Files can be identified, copied, deleted, and otherwise managed at a very high level.
- File servers and disk servers can be considered as extensions to the internal workstation hard drive that may be configured as a shared volume so that
  - collaborators on the same network can share data stored on the server.

15

Definitions

- A database is a collection of one or more
  - related tables
- A table is a collection of one or more
  - rows of data
- A row is a collection of one or more
  - data items, arranged in columns
- A database system is a computer program (or group of programs)
  - that provides a mechanism to define and manipulate one or more databases

16

Definitions

- Available Database Systems
  - Personal database systems
    - Designed to run on PCs
      - Access, Paradox, FileMaker, dBase
  - Enterprise database systems
    - Designed to support efficient storage and retrieval of vast amount of data
      - Interbase, Ingres, SQL Server, Informix, DB2, Oracle
  - Open source database systems
    - Free!!! (Linux!!!)
      - PostgreSQL, MySQL

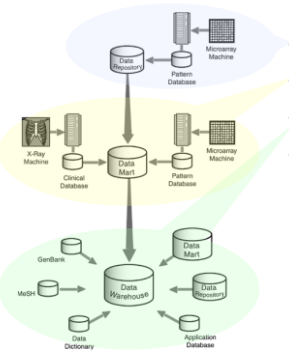
17

Definitions

- SQL
  - The Language of Databases
  - Structured Query Language
- SQL provides two facilities:
  - A database definition Language (DDL)
    - provides a mechanism whereby databases can be created
  - A Data Manipulation Language (DML)
    - provides a mechanism to work with data in tables

18

Database Nomenclature



- Data repositories
- Data marts,
- Data warehouses
- differ primarily in the diversity of data sources that contribute to their contents

19

Database Nomenclature

- The data repository,
  - a database used as an information storage facility,
    - with minimal analysis or querying functionality.
  - a structured, systematically collected storehouse of data distilled or mirrored from a single application,
    - such as a sequencing machine, microarray analyzer, or clinical system
- Advantages of using a data repository:
  - longitudinal studies are possible because all data in the host application are mirrored and stored in the repository
  - it offloads the query functions that are available through native applications to the database management system that enables efficient control and management of the data repository

20

Database Nomenclature

- The data mart,
  - a searchable database system,
    - organized according to the user's likely needs
  - contains a subset of the data contained in other databases as opposed to an indiscriminate mass copying of all the data from another database.
- The major difference between a data mart and a data repository is that
  - a data mart contains data extracted or mirrored from multiple application databases.

21

Database Nomenclature

- The data warehouse,
  - a central database, frequently very large,
  - can provide authenticated researchers with access to all of an institution's information.
    - A data warehouse is usually populated with data from a variety of non-compatible sources, such as
      - sequencing machines, clinical systems, or national genomic databases.
- Because a data warehouse combines data from a variety of application-oriented databases into a single system,
  - data from disparate sources must be cleaned, encoded, and translated so that a standard set of analytical tools can be used with the data.

22

Database Nomenclature

- Note that data repositories, data marts, and data warehouses are simply databases.
  - The three architectures share the usual issues of
    - database design,
    - provision for maintenance,
    - security,
    - periodic modification.
  - Data repositories, data marts, and data warehouses are built with some form of a database management system
    - a program that allows researchers to store, process, and manage data in a systematic way.

23

Data mining

- One of the uses of a fully functional data warehouse or data mart is that it supports data mining
  - The process of extracting meaningful relationships from usually very large quantities of seemingly unrelated data.
  - Specialized data-mining tools allow researchers to perform complex analyses and predictions on data.
- A prerequisite to data mining and the archiving process in general is the availability of a controlled vocabulary

24

Data dictionary

- This controlled vocabulary is most often implemented as part of a data dictionary
  - a program that maps or translates identical concepts that are expressed in different words, phrases, or units into a single vocabulary.
    - A popular controlled vocabulary is the Medical Subject Heading (MeSH), maintained by the U.S. National Library of Medicine, and used with the government-sponsored PubMed biomedical literature database.

25

Data archive

- Data archive
  - a non-volatile holder for data that are infrequently accessed
  - optimized for data recovery and data longevity
  - made on multi-gigabyte tape cartridges
  - are stored offsite in environmentally controlled conditions to minimize the chances of data loss
- An archive needn't be a database.

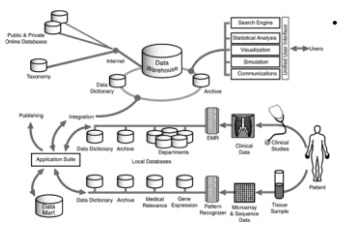
26

Data Management

- Main principle in applied information technology
  - process should drive technology.
    - If there is an obvious need that is only partially or inefficiently addressed, it's much easier to introduce a technology to address the need than it is to eradicate the need through technology alone.
- In biomedical informatics, database technology is the means to handling the enormity of data and information that is
  - created, manipulated, and communicated every day.
- Consider the various components in the biological data-management scenario in the pharmacogenomic laboratory depicted in the next slide

27

Data Management



- In this data-management scenario for a pharmacogenomic laboratory, data of various types are
  - acquired from a variety of sources,
  - incorporated into the data warehouse,
  - used by a variety of applications,
  - archived for future use.
- Data created locally
  - may be published electronically,
  - serve as the basis for a paper publication,
  - may be used in a variety of applications,
    - from drug discovery to genetic engineering.

28

Data Management

- In this scenario, patient medical records are combined with genomic data in order to associate genes with particular diseases.
  - Researchers in the laboratory also have access to the public and private online databases,
    - such as those from the National Center for Biological Information and Celera Genomics, respectively.
  - In addition to numerous application-specific databases in the clinical departments and local databases associated with the sequencing machines,
    - researchers query local data repositories of aggregated data, data marts, and a data warehouse.

29

Typical Electronic Medical Record (EMR) Contents

- The EMR contains both objective signs, such as physical examination findings, as well as subjective patient symptoms, including chief complaint and review of systems.

Data Category	Description
Chief Complaint	Patient's primary reason for the medical visit
History of Present Illness	History of onset of clinical signs and symptoms
Medications	Current list of medications the patient is using

30

Typical Electronic Medical Record (EMR) Contents

Past Medical History	Relevant past medical history, including hospital admissions, surgeries, and diagnoses
Family History	History of family diseases, such as diabetes, cancer, heart disease, and mental illness
Social History	Use of drugs, smoking, job stability, housing, living conditions, incarceration
Review of Systems	Patient's recollection of symptoms and current medical problems, such as trouble sleeping at night or panic episodes, and results of tests

31

Typical Electronic Medical Record (EMR) Contents

Physical Examination	The clinician's hands-on examination of the patient, including head, eyes, ears, nose, throat, chest, and extremities
Labs	Includes blood glucose, cholesterol, and drug levels
Studies	X-ray, MRI, CT, and EKG
Progress notes	Record of temporal progression of signs and symptoms, labs, and studies for the length of the study or admission

32

Data Management

- The components of the EMR report rarely exist in a single, unified database, but reside in the separate, domain-specific databases that may exist within a single hospital or clinic or be dispersed geographically across a region or country.
- Regardless of their relative proximity to each other, laboratory, radiology, cardiology, hematology, internal medicine, and other clinical departments typically maintain their own medical-record systems.

33

Data Management

- each application may be supported by a different operating system, use a different underlying database
  - some of which may be outdated and execute on a completely different hardware platform.
- The traditional method of creating a composite view of a patient's clinical status is to generate custom reports, which is time-consuming and expensive.
- The modern approach to the EMR is to create one or more central databases derived from, and yet completely independent of, each of the application databases, and to optimize these databases for research and analysis

34

Data Management

- In order to create a comprehensive record that can be queried, the data from the various clinical systems have to be integrated,
  - usually with the assistance of a data dictionary that translates various clinical databases to common formats so that the data can be more easily combined.
- The data dictionary is a collection of information about naming, classification, structure, usage, and administration of data that originates from a variety of sources.

35

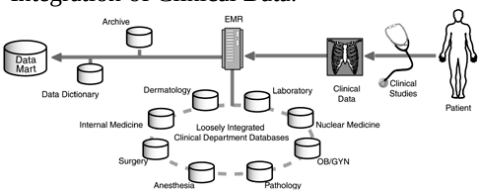
Data Management

- The data dictionary can also impose a standard vocabulary on the system so that clinical findings can be identified unambiguously.
  - For example, one clinical system might refer to heart attack as "M.I.,« another as "Myocardial Infarction," and yet another as "Heart Attack."
- By imposing a standard vocabulary, the data dictionary allows data from the various systems to be combined into a unified view of the patient that can be more easily mined for patterns.
  - This view is typically maintained in a data mart, as illustrated in the next slide.
- The data mart contains a subset of the data that resides in the individual databases combined with contents from these databases translated into a standard format that can be efficiently mined for data.

36

Data Management

- Integration of Clinical Data.



- To create an EMR capable of supporting efficient data mining, a data dictionary is used to impose a standard format and vocabulary on data stored in the clinical data mart.

Data Management

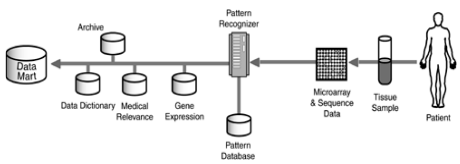
- A similar situation exists in the bioinformatics component of the patient data management.
- Patients provide DNA source material for analysis in the form of tissue samples, which are processed for microarray analysis, generating thousands of data points.
- These data are then processed by a pattern-recognizer program to identify significant patterns.

Data Management

- Researchers rely on local databases of gene expression, medical relevance, and a data dictionary to provide a common language and format for the data.
- Links to the large public genomic databases provide additional reference material.
- As with the clinical data, the composite genomic data are stored in a data mart for efficient manipulation and analysis through a suite of applications.
- Ideally, relevant data from clinical applications are combined in the data mart as well.

Data Management

- Integration of Bioinformatics Data.

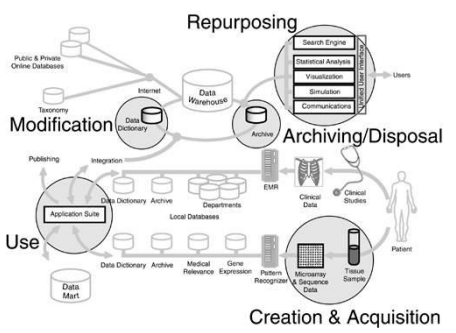


- Like clinical data, bioinformatics data from a variety of sources and in numerous formats are combined in a data mart to enhance data management.

Data Life Cycle

- In the data-management process, data are
  - authored by clinicians and researchers and generated directly by research and test equipment,
  - used by a variety of applications,
  - repurposed or modified for other uses,
  - archived for future study.
- Eventually, the data are disposed of, freeing the data warehouses and other hardware from the overhead of maintaining low-value data.
- The overall process, from data creation to disposal, is normally referred to as
  - the data life cycle, as depicted in the next slide.

Data Life Cycle



Data Life Cycle

- Key steps in the process include
  - data creation and acquisition,
  - data use,
  - data modification,
  - data repurposing,
  - data archiving and disposal.
- The same process applies to data in a desktop workstation or, to a large pharmacogenomic operation with multiple, disparate systems.

43

Data Life Cycle

- **Data Creation and Acquisition**
- The process of data creation and acquisition is a function of the source and type of data.
  - For example, in the scenario depicted in previous slide, data are generated by sequencing machines and microarrays in the molecular biology laboratory, and by clinicians and clinical studies in the clinic or hospital.
- Depending on the difficulty in creating the data and the intended use, the creation process may be trivial and inexpensive or extremely complicated and costly.
  - For example, recruiting test subjects to donate tissue biopsies is generally more expensive and difficult than identifying patients who are willing to provide less-invasive (and painful) tissue samples.

44

Data Life Cycle

- In addition to cost, the major issues in the data-creation phase of the data life cycle include
  - tool selection, data format, standards, version control, error rate, precision, and accuracy.
    - These metrics apply equally to clinical and genomic studies.
    - In particular, metrics such as error rate, precision, and accuracy are more easily ascribed to machine-generated data, whether from clinical laboratory studies or microarray analysis.
      - For example, optical character recognition (OCR), which was once used extensively as a means of acquiring sequence information from print publications, has an error rate of about two characters per hundred, which is generally unacceptable.

45

Data Life Cycle

- **Use**
- Once clinical and genomic data are captured, they can be put to a variety of immediate uses,
  - from simulation, statistical analysis, and visualization to communications.
- Issues at this stage of the data life cycle include
  - intellectual property rights, privacy, and distribution.
    - For example, unless patients have expressly given permission to have their names used, microarray data should be identified by ID number through a system that maintains the anonymity of the donor.

46

Data Life Cycle

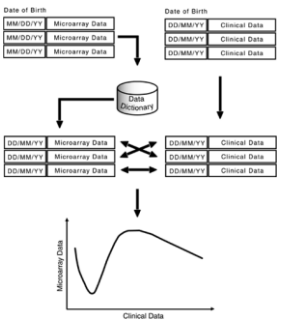
- **Data Modification**
- Data are rarely used in their raw form, without some amount of formatting or editing.
- In addition, data are seldom used only for their originally intended purpose, in part because future uses are difficult to predict.
  - For example, microarray data may not be captured expressly for comparison with clinical pathology data, but it may serve that purpose well.
- The data dictionary is one means of modifying data in a controlled way that ensures standards are followed.
- A data dictionary can be used to tag all microarray data with time and date information in a standard format so that they can be automatically correlated with clinical findings (next slide).

47

Data Life Cycle

- **Archiving**
- Archiving, is concerned with making data available for future use.
- An archive is a container for data that is infrequently accessed, with the focus more on longevity than on access speed.
- In the archiving data are named, indexed, and filed in a way that facilitates identification later.
- While university or government personnel archive the large online public databases, the archiving of locally generated data is a personal or corporate responsibility.
- Regardless of who takes responsibility for the process, the issues associated with archiving are numerous (next slide)

48



- Data Dictionary-Directed Data Modification.
- The time and date header for microarray data can be automatically modified so that it can be easily correlated with clinical findings.
- Data that are modified or transformed by the data dictionary are normally stored in a data mart or data warehouse so that the transformed data are readily available for subsequent analysis without investing time and diverting computational resources by repeatedly reformatting the data.

49

Archiving Issues

Issue	Description
Indexing	Vocabulary, metadata, language, completeness, efficiency
Space Requirements	Index space versus data space
Hardware Requirements	Hard drives, network
Scalability	Ability to expand functionality without investing in new hardware and software
Database Design	Data model
Archival Process	Responsibilities for overseeing the process
Space Requirements	Current and projected archival capacity
Completeness	Relative quantity of total data that are archived
Media Selection	Compatibility, speed, capacity, data density, cost, volatility, durability, and stability
Location	Local, server-based, or network
Infrastructure Requirements	Network and computer hardware
Relative Value	Value of data vs. archival overhead
Hardware Configuration	RAID and other configurations
Longevity	Technical obsolescence of media and MTBF rating of related equipment
Security	Limited access to data

50

Data Life Cycle

- **Repurposing**
- One of the major benefits of having data readily available in an archive is the ability to repurpose it for a variety of uses.
  - For example, linear sequence data originally captured to discover new genes are commonly repurposed to support the 3D visualization of protein structures.
- One of the major issues in repurposing data is the ability to efficiently locate data in archives.
  - The difficulty in locating data once it's been incorporated into a storage system depends on the volume of data involved.
- Efficient retrieval is a function of
  - the hardware and database management software,
  - the effectiveness of the user interface, and
  - the granularity of the index.
    - For example, nucleotide sequence data indexed by chromosome number would be virtually impossible to locate if the database contains thousands of sequences indexed to each chromosome.

51

Data Life Cycle

- Issues in the repurposing phase of the data life cycle include
- the sensitivity, specificity, false positives, and false negatives associated with searches.
- The usability of the user interface is also a factor, whether free-text natural language, search by example, or simple keyword searching is supported.
- In addition, the provisions for security can affect the ease with which data can be located and repurposed.
  - An overly complex security procedure that requires revalidation of user identity every five minutes could deter even the most well-intentioned researcher.

52

Data Life Cycle

- **Disposal**
- The duration of the data life cycle is a function of
  - the perceived value of the data,
  - the effectiveness of the underlying process,
  - the limitations imposed by the hardware, software, and environmental infrastructure.
- Eventually, all data die, either because they are intentionally disposed of when their value has decreased to the point that it is less than the cost of maintaining it, or because of accidental loss.
- Often, data have to be archived because of legal reasons, even though the data is of no intrinsic value to the institution or researcher.

53

Managing the Life Cycle

- Managing the data life cycle is an engineering exercise that's a compromise between speed, completeness, longevity, cost, usability, and security.
  - For example, the media selected for archiving will not only affect the cost, but the speed of storage and longevity of the data.
- Similarly, using an inhouse tape backup facility may be more costly than outsourcing the task to networked vendor, but the in-house approach is likely to be more secure.
  - These tradeoffs are reflected in the implementation of the overall data-management process.

54