Biostatistics and informatics

Clinical databases

Miklós Kellermayer

Outline

1. Informatics summary
2. Coding, efficiency, redundancy, the genetic code
3. Databases. Important databases in biomedical sciences
4. Bioinformatics
5. Medical knowledge database
6. Healthcare informatics

Informatics summary

Information:
In technical sense - an ordered set of symbols
As concept - various meanings (communication, semantics, physics, sensory input, etc.)
Informare (Lat): "to give form to the mind", "to instruct", "to discipline"

Information content (H):
Related to uncertainty, disorder (entropy) - news that alleviates uncertainty

Information content in case of equal-probability outcomes:

$$H = \log_b n$$

H=entropy, information content
n=number of outcomes of an experiment, the probability of which are identical (e.g., coin-toss, chess table, dice rolling)
b=usually 2; then the unit of H is the "bit"

Informatics summary

Information content in case of unequal-probability outcomes:

$$H = -\sum_{i=1}^{n} P(A_i) \log_b P(A_i)$$

H = entropy, information content
i = 1, 2, 3 ...
n = number of possible experimental outcomes
P = probability
A_i = ith experimental outcome
$P(A_i)$ = probability of the ith experimental outcome
b = usually 2; then the unit of H is the "bit"

Information content increases if:
n increases, P decreases, $P(A_i)$ approaches identical 1/n values.

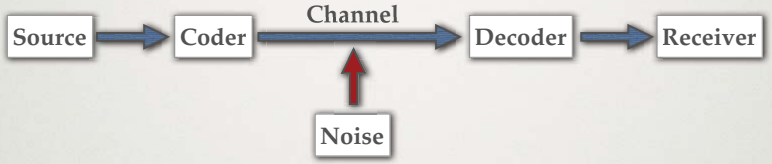
Note:
if $P(A_1) = P(A_2) = ... = P(A_n)$, then $H = \log_b n$


Claude Shannon
(1916-2011)

Information content of a million-base-long DNA:
Number of outcomes at each sequence position = 4
Total number of outcomes: $n = 4^{(10^6)}$ $H = 10^6 \log_2 4 = 2 \cdot 10^6$ *bit*

Coding of information

Transmission of information:



Coding efficiency: $\eta = \frac{H}{H_{\max}} = \frac{H}{\log_2 n}$ (*H* is maximal in the case of equal-probability outcomes.)

Redundancy: $R = 1 - \eta$

Coding of genetic information:

4 bases in triplets. Assuming equal probability of occurence, maximal information content (Hmax) is $4^3 = 64$. Because actual probabilities are unequal (i.e., not all triplets are used), the genetic code is redundant. Nevertheless, it is quite sufficient for coding the 20 amino acids. Other examples - alphabet.

Databases

Organized collection of data (coded information).

Classification of databases: usually by content (e.g., bibliographic, document-text, statistical)

Digital databases: managed using database management systems (DBMS), which store database contents

DBMS allows:
- Data creation
- Database maintenance
- Search
- Data access

DBMS consists of **software** that operates databases, providing storage, access, security, backup and other facilities.

Biomedically important databases

- Bioinformatics
- Medical knowledge
- Healthcare informatics

Bioinformatics

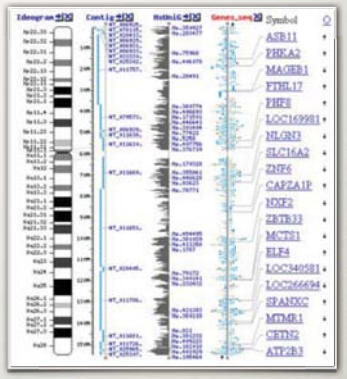
Bioinformatics: application of statistics and computer science to the field of molecular biology.

History:

The term *bioinformatics* was coined by Paulien Hogeweg and Ben Hesper in 1978 for the study of informatic processes in biotic systems.

Its primary use since at least the late 1980s has been in genomics and genetics, particularly in those areas of genomics involving large-scale DNA sequencing.

Human Genome Project (1990-2003) - international research project for identifying the map of the 20 - 25 thousand genes of the human genome both physically and functionally.



Map of the human X chromosome

The primary objective of bioinformatics:

To increase the understanding of biological processes by developing and applying computationally intensive techniques (e.g., pattern recognition, data mining, machine learning algorithms, and visualization).

Sequence analysis

History:

1977: First organismal genome sequenced (Phage Φ-X174)
1995: First free-living organismal genome sequenced (Haemophilus influenzae)
2003: Human genome sequenced

Genomic sequences are stored in sophisticated databases.

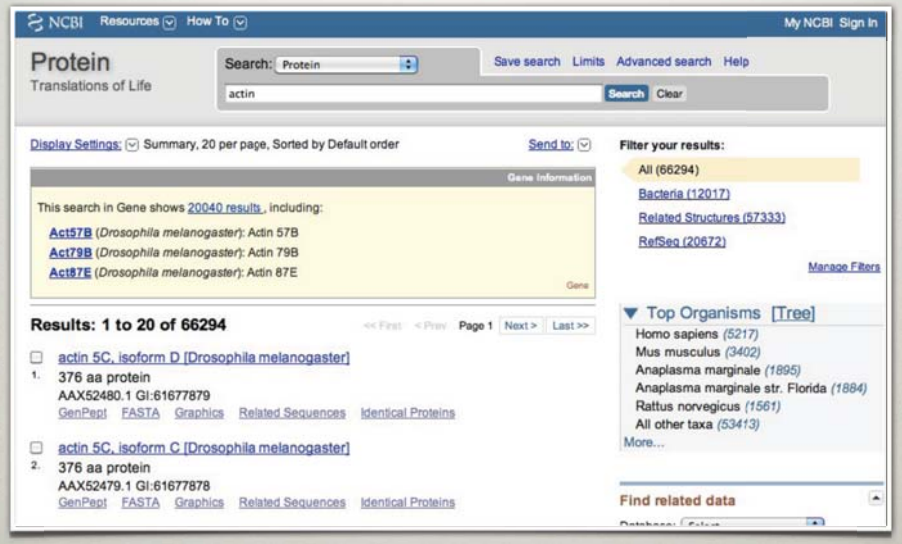
Sequence databases allow:

Searches
Sequence alignment (finding similarities) - BLAST
Gene finding
Annotation (marking features)

Major bioinformatics databases

- European Bioinformatics Institute databases
- NCBI completely sequenced genomes
- Stanford Saccharomyces Genome Database
- Protein

Sequence analysis



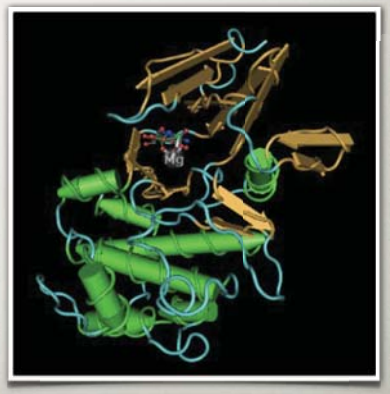
STRUCTURAL DATABASES

Biomolecular structural information is stored in sophisticated databases:

- NMR structure
- X-ray diffraction structure
- Cryo-electron microscopic structure

Structural bioinformatics aim at modeling:

Predicted protein structure
Structural homologies
Protein-ligand interactions
Protein-protein docking



Actin

STRUCTURAL DATABASES

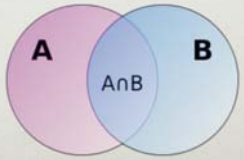
Medical knowledge database

Structured bibliographic database of life sciences and biomedical information.

National Library of Medicine (USA)
- 1879-2004: ***Index medicus***, a monthly guide to ~5000 selected journals
- ***MEDLINE***: ~18 million records (articles) of biomedicine and healt.

Accessible via the free PubMed

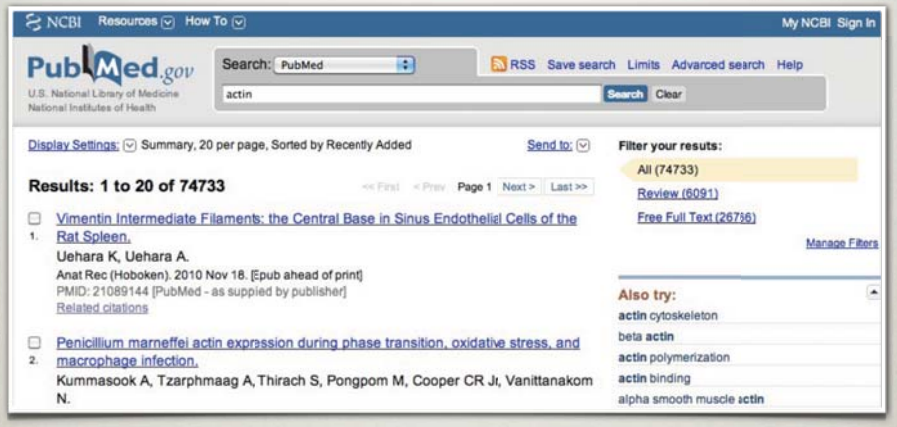
MEDLINE uses Medical Subject Headings (MeSH) for information retrieval. Engines designed to search MEDLINE (such as Entrez and PubMed) generally use a **Boolean expression** combining MeSH terms, words in abstract and title of the article, author names, date of publication, etc.



Venn diagram, set theory

Boolean operators:
A OR B: union
A AND B: intersection
A XOR B: union-intersection

Medical knowledge database



Healthcare informatics

Healthcare informatics (**medical informatics** or **biomedical informatics**): discipline at the intersection of information science, computer science, and health care.

Healthcare informatics deals with the resources, devices, and methods required to optimize the acquisition, storage, retrieval, and use of information in health and biomedicine.

Health informatics tools include not only computers but also clinical guidelines, formal medical terminologies, and information and communication systems.

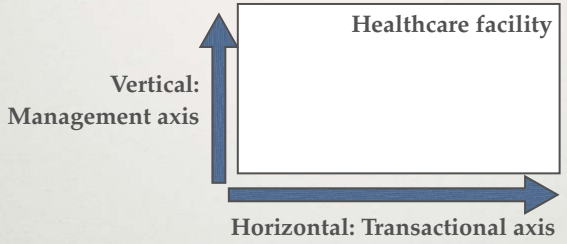
It is applied to the areas of nursing, clinical care, dentistry, pharmacy, public health and (bio)medical research.

Aspects:

- architectures for electronic medical records
- decision support systems in healthcare
- standards (e.g. DICOM, HL7) and integration profiles (e.g. Integrating the Healthcare Enterprise) to facilitate the exchange of information between healthcare information systems
- controlled medical vocabularies (CMVs) - used to allow a standard, accurate exchange of data content between systems and providers
- use of hand-held or portable devices to assist providers with data entry/retrieval
- Molecular bioinformatics and clinical informatics - translational bioinformatics.

Healthcare informatics

Information flow within healthcare systems



Transactional information flow: patient information (records) exchanged between participants of health care.

Management information flow: transfer of non-medical (aggregated) information relevant for operation of the healthcare facility (billing, etc.).

Healthcare informatics

Types of data generated within a healthcare system:

1. Data attached to the patient:
patient information related to the medical treatment of the patient (medical records).

2. Data attached to the institution:
information related to the management and administration of the healthcare institution.

Healthcare informatics

History of Hospital Information Systems:

1960-1970s:
-introduction of first, simple, computerized protocols (billing)
-appearance of mainframe computers
-first patient admission systems
-first laboratory database systems

1980s:
-management-based systems
-direct medical assisting systems (radiology records, intensive care records, etc.)
-integration

1990s:
-integrated patient care
-computerized health record

Healthcare informatics

History of the patient record:
From Hippocrates through Medieval times:
-Time-oriented patient record
-Documentation of the temporal evolution of the disease (disease-centered medicine)

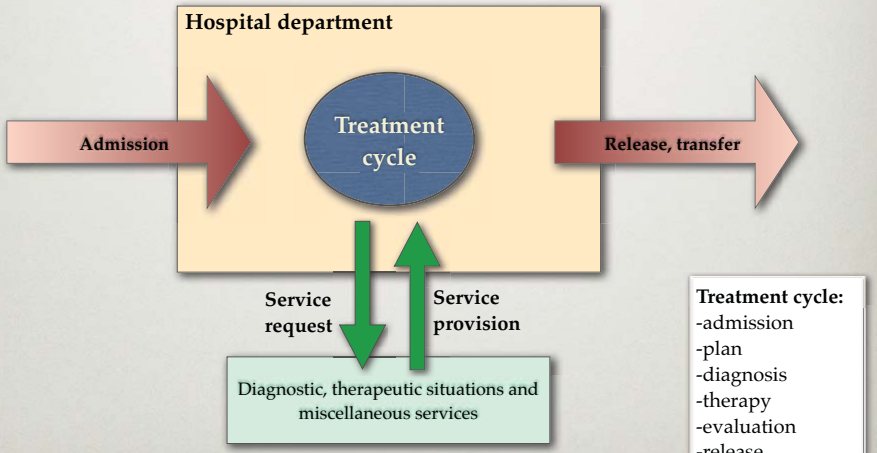
William Mayo, 1880:
-Patient data in chronological order

Plummer, 1907:
-Organization of patient data as separate unit. Beginnings of patient-centered medicine. Identification of "minimally required set of data".

1960s:
-Problem-oriented SOAP data structure
S: subjective data (patient history)
O: objective data (physical examination)
A: assessment (complex test results with interpretation and conclusion)
P: plan (diagnostic and therapeutic protocols)

Healthcare informatics

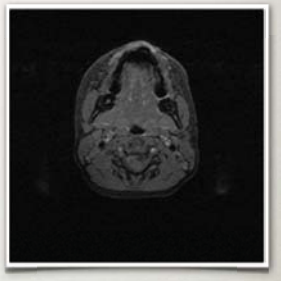
Structure of hospital patient care

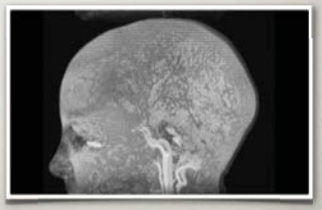


Healthcare informatics

Digital Imaging and Communications in Medicine

Digital **I**maging and **Co**mmunications in **M**edicine (**DICOM**) is a standard for handling, storing, printing, and transmitting information in medical imaging. It includes a file format definition and a network communications protocol. The communication protocol is an application protocol that uses TCP/IP to communicate between systems. DICOM files can be exchanged between two entities that are capable of receiving image and patient data in DICOM format.





DICOM enables the integration of scanners, servers, workstations, printers, and network hardware from multiple manufacturers into a picture archiving and communication system (PACS). The different devices come with DICOM conformance statements which clearly state the DICOM classes they support. DICOM has been widely adopted by hospitals and is making inroads in smaller applications like dentists' and doctors' offices.