MEDICAL INFORMATICS

Lecture (4)

Bioinformatics in Medicine

Faisal Al-Madi **Reem Al jurayyad**

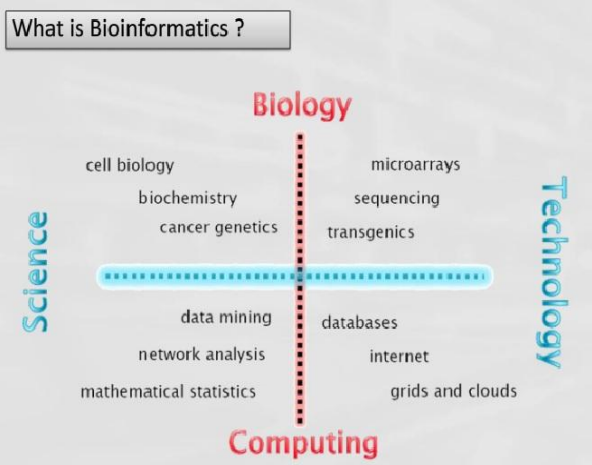
Aysha Al-mahboob **Areej Al-Qunaitir**

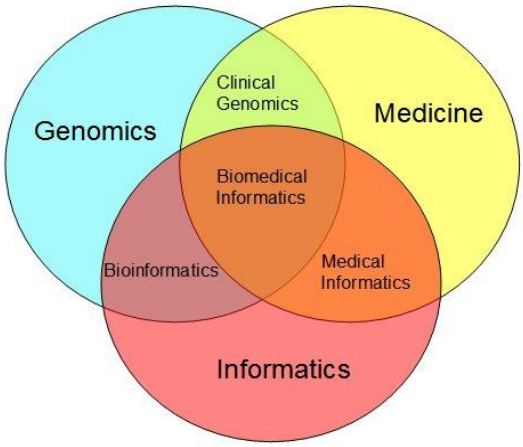
Faisal al-olayan **Ahmad Bogari**

Akeel Al Mahdaly

MEDICAL INFORMATICS

Part 1(Introduction to Bioinformatics):





What is bioinformatics?

Bioinformatics is the application of computer technology to the management of biological information.

-Some Definitions :

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health date, including those to acquire, store, organize, archive, analyze, or visualize such data.

Computational biology: the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study if biological, behavioral, and social systems.

-What is the difference between Bioinformatics and Computational Biology?

Bioinformatics tends to describe data analysis; Computational Biology tends to describe modeling. But these two terms are used interchangeably.

MEDICAL INFORMATICS

Another Definition:

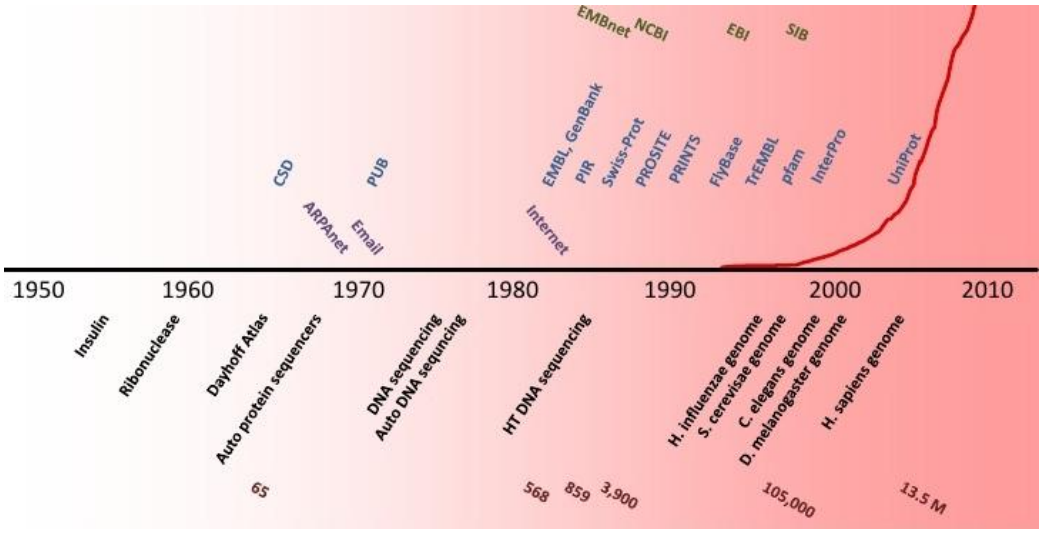
Bioinformatics: Is the field of science in which biology, computer science, and information technology merge into a single discipline. There are three important sub-disciplines within bioinformatics: the development of new algorithms and statistics with which to assess relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and the development and implementation of tools that enable efficient access and management of different types of information.

One More:

Bioinformatics: The use of computers by necessity to enable any study, in any field of the life sciences.

"Nurse, get on the internet, go to SURGERY.COM, scroll down and click on the 'Are you totally lost?' icon."

A Very Brief History:



MEDICAL INFORMATICS

1953: Watson and Crick propose the double helix model for DNA based on x-ray data. 1955: The sequence of the first protein to be anahlyzed, bovine insulin is announced by F. Sanger.

1970: The details of the Needleman-Wunsch algorithm for sequence comparison are published.

1973: The Brookhaven Protein Data Bank is announced.

1980: The first complete gene sequence for an organism (FX174) is published. The gene consists of 5,386 base pairs which code nine proteins.

-IntelliGenetics, Inc. founded in California. Their primary product is the IntelliGenetics Suite of programs for DNA and protein sequence analysis.

1981: The Smith-Waterman algorithm for local sequence alignment is published.

1988: The National Center for Biotechnology Information (NCBI) is established at the National Cancer Institute.

-The Human Genome Initiative is started

-The FASTA algorithm for sequence comparison is published by Pearson and Lupman.

-Des Higgins and Paul Sharpe announce the development of CLUSTAL

-A new program, an Internet computer virus designed by a student, infects 6,000 military computers in the US.

1990: The BLAST program (Altschul, et. al.) is implemented.

1995: The Haemophilus influenzea genome (1.8 Mb) is sequenced.

-The Mycoplasma genitalium genome is sequenced.

1996: The genome for Saccharomyces cerevisiae (baker's yeast, 12.1 Mb) is sequenced.

1997: The genome for E. coli (4.7 Mbp) is published.

MEDICAL INFORMATICS

1998: The genomes for Caenorhabditis elegans and baker's yeast are published.

2000: The genome for Pseudomonas aeruginosa (6.3 Mbp) is published.

-The A. thaliana genome (100 Mb) is secquenced.

-The D. melanogaster genome (180Mb) is secquenced.

2001: The human genome (3,000 Mbp) is published.

2007: Applied Biosystems started selling a new type of sequencer called SOLiD System. The technology allowed users to sequence 60 gigabases per run.

February 2009: Complete Genomics released a full sequence of a human genome that was sequenced using their service.

April 2009: Complete Genomics announced that it plans to sequence 1,000 full genomes between June 2009 and the end of the year and that they plan to be able to sequence one million full genomes per year by 2013.

June 2009: NABsys announced their goal of full genome sequencing for under US$100 per genome with a turnaround time of less than 15 minutes.

September 2009: The President of Halcyon Molecular announced that they will be able to provide full genome sequencing in under 10 minutes for less than US$100 per genome. This is, to date, the most ambitious promise of any full genome sequencing company.

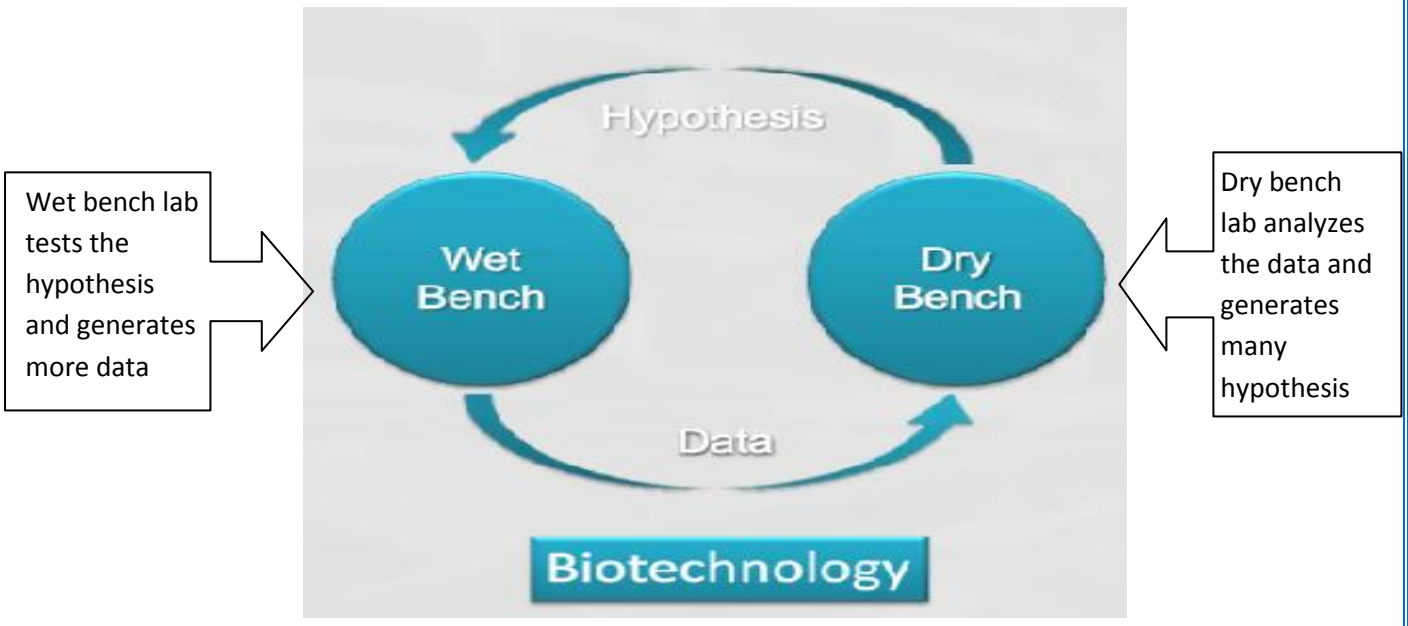
March 2010: Researchers from the Medical College of Wisconsin announced the first successful use of Genome Wide sequencing to change the treatment of a patient. This story was later retold in a pulitzer prize winning article and touted as a significant accomplishment in Nature and by the director of the NIH in presentations at congress.

2011: Knome provides full genome (98% genome) sequencing services for US$39,500 for consumers, or US$29,500 for researchers (depending on their requirements).

MEDICAL INFORMATICS

-Complete Genomics charges approximately US$10,000 to sequence a complete human genome (less for large orders).
May 2011: Illumina lowered its Full Genome Sequencing service to US$5,000 per human genome, or US$4,000 if ordering 50 or more.
January 2012: Life Technologies introduced a sequencer to decode a human genome in one day for $1,000.
-A UK firm spun out from Oxford University has come up with a DNA sequencing machine (the MinION) the size of a USB memory stick which costs $900 and can sequence simple genomes (but not full human genomes).

***Bioinformatics in the Lab :**



Biotechnology: The use of living systems and organisms to develop or make useful products, or "any technological application that uses biological systems, living organisms or derivatives thereof, to make or modify products or processes for specific use

MEDICAL INFORMATICS

***Key Concepts in Bioinformatics:**

1-Abstraction: reduce complexity to bare minimum.
2-Representation: how to put the data into the computer memory.
3-Databases: Organize, manage, store and retrieve.
4-Search:
Need to be fast and accurate:

-Needleman-Wunsch algorithm for global sequence alignment
-Smith-Waterman algorithm for local sequence alignment
-BLAST (Basic Local Alignment and Search Tool).

Advanced CS (Computer Science) techniques:

- Indexing
- Replication
- Distrbuted Access
-Google secret "recipe"

5-Integration: The total sum is greater than the parts.
6-Transitivity:
Implication by similarity:

-if we know gene A has certain properties
-And we know gene B is similar to gene A (by sequence, structure or composition)
-We may assume gene B has almost the same properties of gene A

Implication by guilt:

-if gene A seems to be strongly associated with a group of genes G (by expression profile or p-p interaction)
- And many of the genes in G are known to be involved in some function F
-We may assume gene A is involved in function F

7-Standards:
Interoperability is key:

-MIAME: Minimum Information about a microarray experiment.
-FASTA format for sequence data.
-BED and GFF genomic annotation formats
-HGNC gene naming standard

MEDICAL INFORMATICS

Still a long to go though

-Most tools use their own input and output formats

-Swapping formats can be tricky and time consuming

-It's mostly an open market dynamic

8-Open Access:

No one can do it alone

-open access journals: PLoS, BioMed central, …

-open access data: NCBI, EMBL, PubMed, …

-open access tools: BioPERL, R, MySQL, …

-open access servers: UCSC Browser, Gene modeling servers, Galaxy framework.

-open access knowledge: Wikipedia, the internet, …

9-Big data:

-Individual experiments are large

-many experiments are done on large samples

-data accumulation rates are huge

-acceleration is accelerating

10-Dynamic:

-new data is added every SECON

-some older data is deleted

-some older data is updated

-websites appear and disappear frequently

-new discoveries are made at a fast pace

-impossible to stand still

Part 2 (Bioinformatics in Action):

*****Bioinformatics in Medicine:

- Medical informatics is a related but different field. This interface is sometimes called (Clinical Bioinformatics).
- Bioinformatics has direct benefits in all areas of medicine and patient care:

❖ -Disease Basic Science

MEDICAL INFORMATICS

- -Disease Prevention
- -Disease Detection
- -Disease Treatment
- -Care Management

-Bioinformatics and Disease Basic Science:

- BI (Bio-Informatics) is now a major tool in the investigation of disease mechanisms, and especially complex diseases such as diabetes and cancer.
- Systems biology is an area in BI that connects the study of the micro to the macro.
- BI is key to the understanding of disease genetics, inheritance, detecting gene roles and risk factors, using tools like GWAS, LD, CNV, …
- BI is an important tool on investigating host/pathogen interaction.

*Bioinformatics and Disease Prevention

- Genetic testing and profiling can detect highly debilitating genetic diseases in the earliest fetal stages
- Neonatal screening can detect metabolic diseases before they are symptomatic.
- Genetic testing can detect high and control life style before disease strikes.
- All the above advances also bring huge ethical and social dilemmas.

*BI and Disease Detection/Diagnosis

- Computer image analysis can improve histology and some routine lab works.
- Microarray based profiling can detect the earliest stages of disease even before the symptoms are clinically detectable.

- Microarray based profiling can accurately classify complex samples (such as cancer) and allow very accurate treatment planning.

*BI and Disease Treatment

- Personalized medicine is on the horizon
- Genetic testing can allow selection of better drugs with fewer side effects for the patient
- BI is a key tool for better drug design, from target selection to development to efficacy prediction. This reduces cost and time to market
- BI analysis is key towards the design of treatments that utilize gene therapy

*BI and Care Management

- Microarray profiling alongside the treatment protocol allows very close monitoring of progress and plan modification.
- Genetic profiling and microarray testing can generate more accurate length improving overall care quality.
- Disease registries can track disease in the society and allow for effective health care planning and resource allocation.

*The Clinic of the Near Future

- Fast advances in sequencing technology will soon make full genome/transcriptome sequencing potentially a routine diagnostic.
- The doctor will have to deal with a flood of risk factors, red herrings and near misses.
- More accurate diagnostics mean bigger responsibility on the doctor to provide best of class and prevention.
- The computer (and the internet) will become an integral tool in the clinic.

The **transcriptome** is the set of all RNA molecules, including mRNA, rRNA, tRNA, and other non-coding RNA produced in one or a population of cells.