Bioinformatics in Medicine

Dr. Anas Al-Halees, Scientist

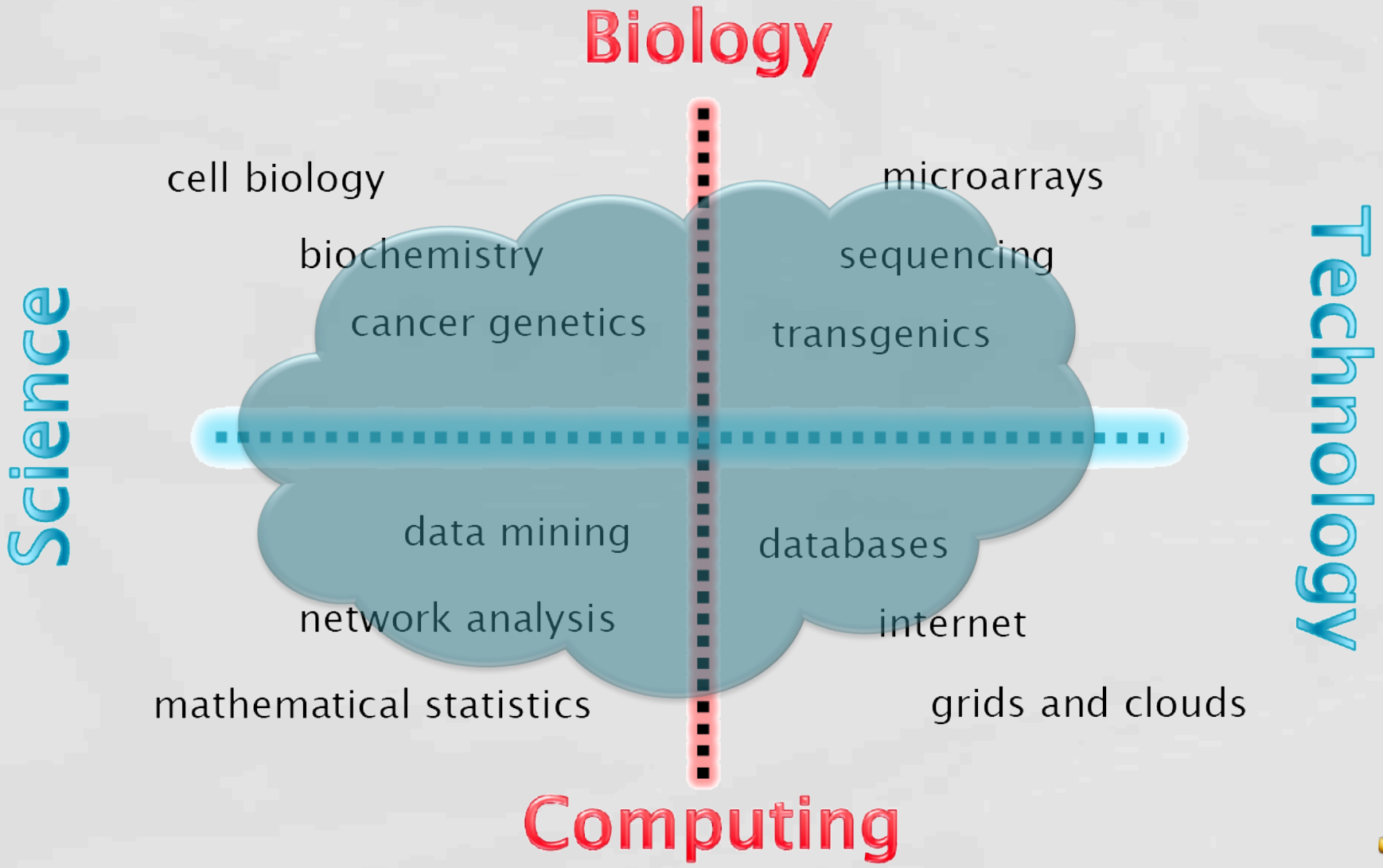
King Faisal Specialist Hospital
and Research Center

ahalees@kfshrc.edu.sa

Part I

Introduction to Bioinformatics

What is Bioinformatics ?



Bioinformatics Defined

- **Bioinformatics:** Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, 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 of biological, behavioral, and social systems.

NIH Biomedical Information Science and Technology Initiative Consortium, 2000

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.

The National Center for Biotechnology Information (NCBI), 2001

One More !

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

My Definition !

- Even more at http://bioinformaticsweb.net/definition.html

A VERY Brief History

- 1953: Watson and Crick propose the double helix model for DNA.
- 1955: The sequence of the first protein to be analyzed, bovine insulin, is announced by F. Sanger.
- 1970: The details of the Needleman-Wunsch algorithm for global sequence comparison are published.
- 1973: The Brookhaven Protein Data Bank is announced.
- 1980: The first complete genome sequence for an organism (a bacteriophage) is published. It 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.

A VERY Brief History

- 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.
- 1990: The BLAST program (Altschul, et. al.) is implemented.
- 1995: The Haemophilus influenzea genome (1.8 Mbp) 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.
- 1998: The genomes for Caenorhabditis elegans (100 Mbp) and baker's yeast are published.

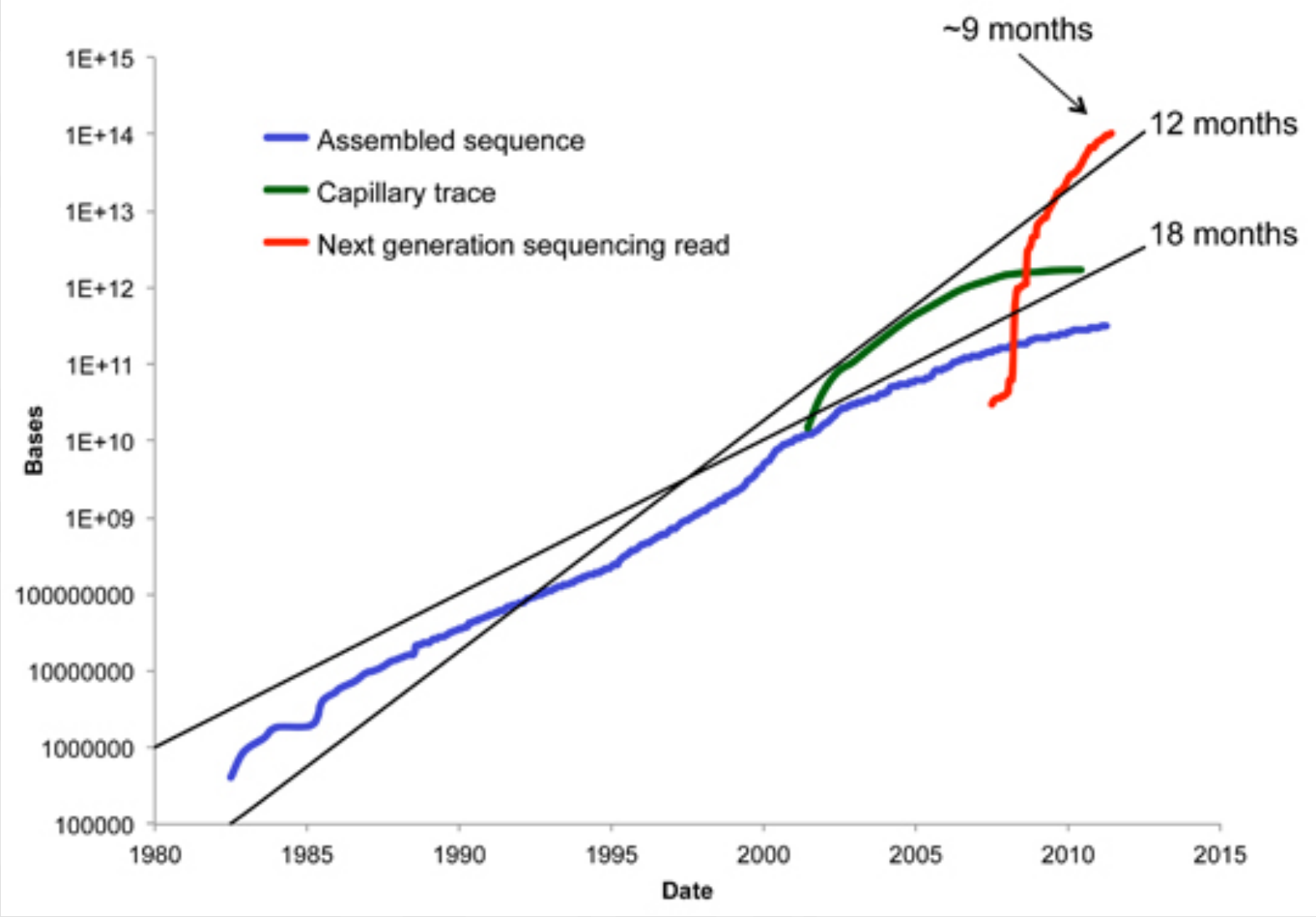
A VERY Brief History

- 2000: The genome for Pseudomonas aeruginosa (6.3 Mbp) is published.
  - The A. thaliana genome (100 Mbp) is sequenced.
  - The D. melanogaster genome (180Mbp) is sequenced.
- 2001: The human genome (3,000 Mbp) is published.
- 2007: Applied Biosystems started selling a new type of sequencer called SOLiD System that can 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 $100 per genome with a turnaround time of less than 15 minutes

A VERY Brief History

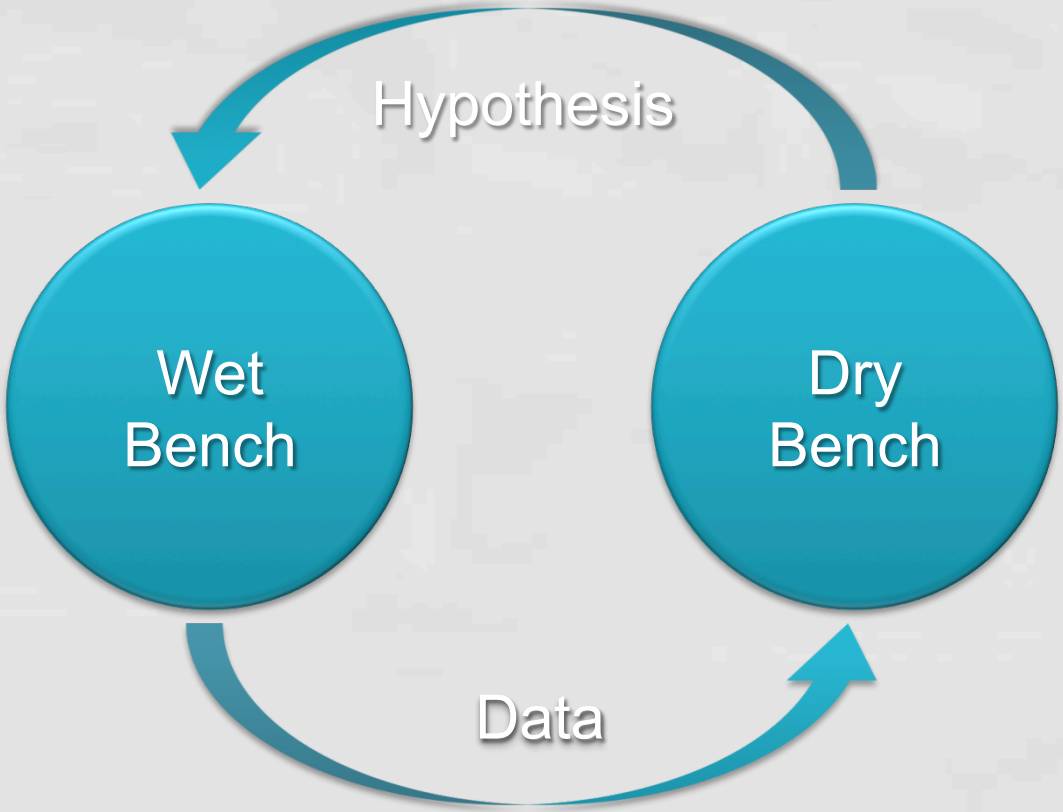
- September 2009: Halcyon Molecular announced that they will be able to provide full genome sequencing in under 10 minutes for less than $100 per genome. This is, to date, the most ambitious promise of any full genome sequencing company.
- March 2010: Researchers from the Medical College of Wisconsin 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 sequencing (98%) services for $39,500 for consumers, or $29,500 for researchers (depending on their requirements).
  - Complete Genomics charges approximately $10,000 to sequence a complete human genome (less for large orders).
- May 2011: Illumina lowered its Full Genome Sequencing service to $5,000 per human genome, or $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).

A Clear Trend



Bioinformatics in the Lab

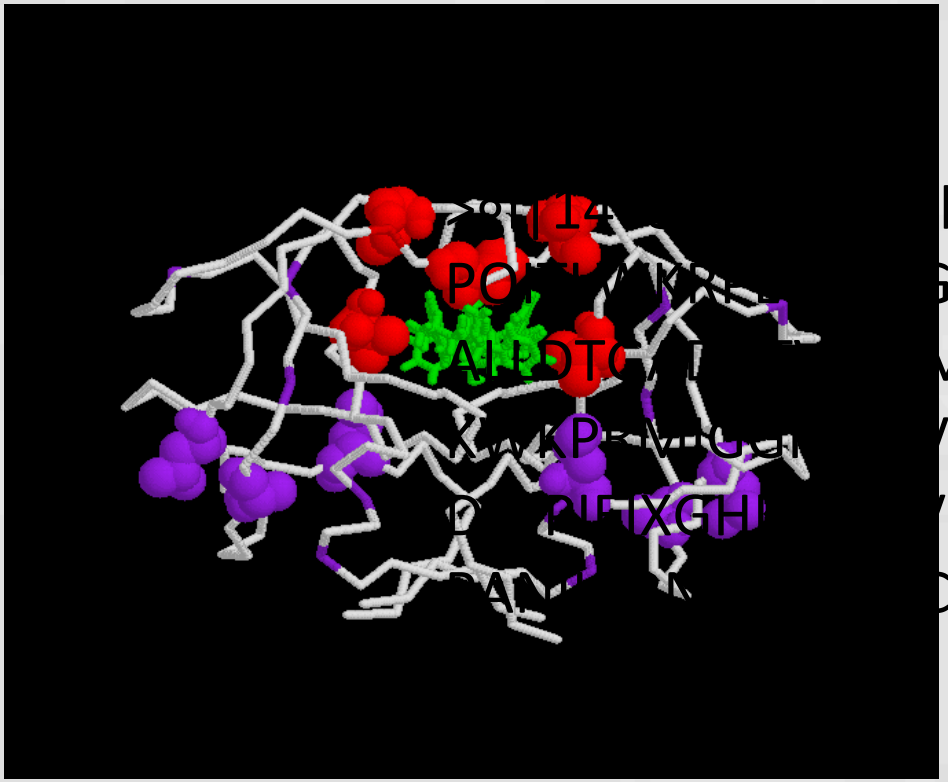
- Wet bench lab generates big data
- Dry bench lab analyzes the data and generates many hypothesis
- Wet bench lab tests the hypothesis and generates more data
- Keep on until ?



Biotechnology

Key Concepts: Abstraction

- Reduce complexity to bare minimum



b|2P3A
GGQLKE
MALPG
VKVRQY
VLVGPT
CTLNF

Key Concepts: Representation

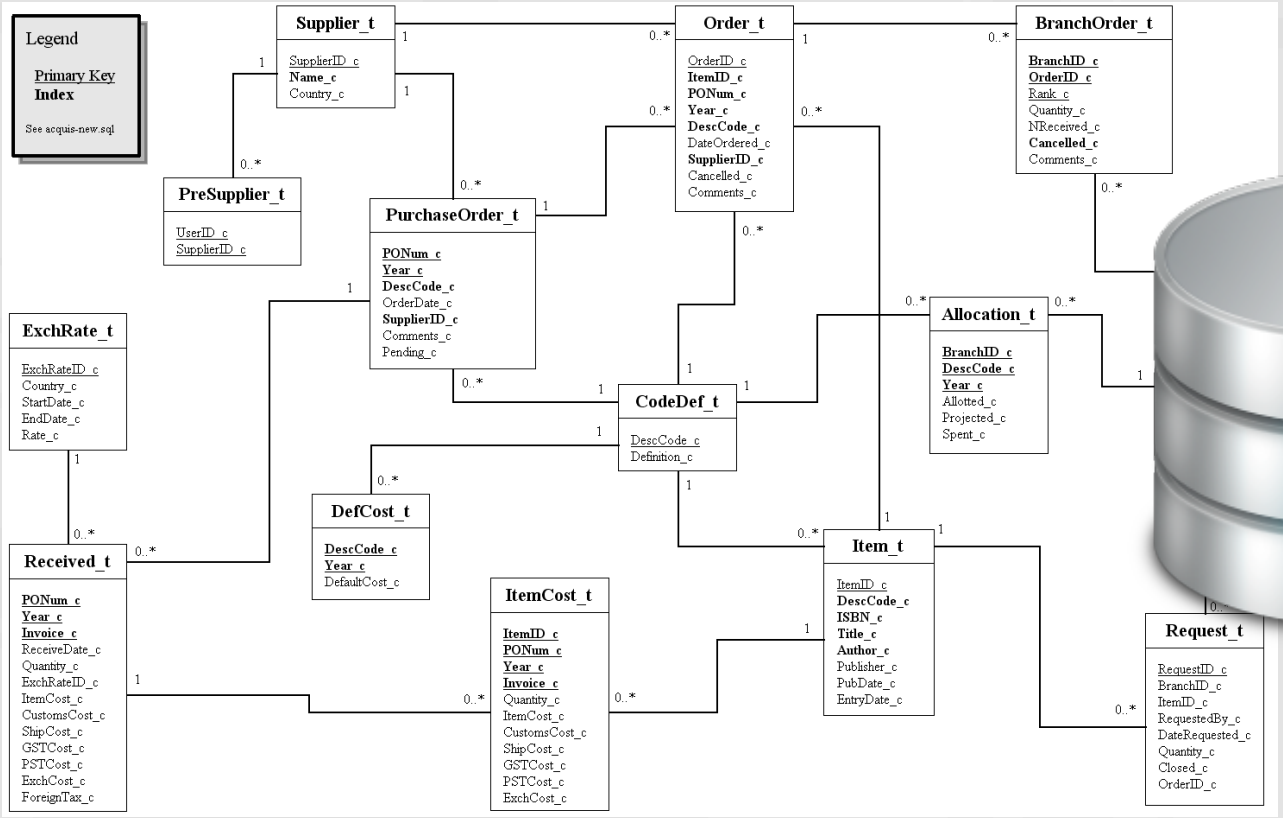
- How to put the data into the computer memory

```
LOCUS       2P3A_B                 99 aa            linear   VRL 10-OCT-2012
DEFINITION  Chain B, Crystal Structure Of The Multi-Drug Resistant Mutant
            Subtype B Hiv Protease Complexed With Tl-3 Inhibitor.
ACCESSION   2P3A_B
VERSION     2P3A_B  GI:146387072
DBSOURCE    pdb: molecule 2P3A, chain 66, release Aug 28, 2012;
            deposition: Mar 8, 2007;
            class: HydrolaseHYDROLASE INHIBITOR;
            source: Mmdb_id: 45722, Pdb_id 1: 2P3A;
            Exp. method: X-Ray Diffraction.
KEYWORDS    .
SOURCE      Human immunodeficiency virus 1 (HIV-1)
  ORGANISM  Human immunodeficiency virus 1
            Viruses; Retro-transcribing viruses; Retroviridae;
            Orthoretrovirinae; Lentivirus; Primate lentivirus group.
REFERENCE   1  (residues 1 to 99)
  AUTHORS   Sanches,M., Martins,N.H., Calazans,A., Brindeiro Rde,M., Tanuri,A.,
            Antunes,O.A. and Polikarpov,I.
  TITLE     Crystallization of a non-B and a B mutant HIV protease
  JOURNAL   Acta Crystallogr. D Biol. Crystallogr. 60 (PT 9), 1625-1627 (2004)
   PUBMED   15333937
COMMENT     1 Protease.
FEATURES             Location/Qualifiers
     source          1..99
                     /organism="Human immunodeficiency virus 1"
                     /db_xref="taxon:11676"
     SecStr          9..16
                     /sec_str_type="sheet"
                     /note="strand 11"
     Region          11..91
                     /region_name="HIV_retropepsin_like"
                     /note="Retropepsins, pepsin-like aspartate proteases;
                     cd05482"
                     /db_xref="CDD:133149"
     SecStr          17..26
                     /sec_str_type="sheet"
                     /note="strand 12"
     Site            order(25,27,29,46..48,84)
                     /site_type="inhibition"
                     /note="inhibitor binding site"
                     /db_xref="CDD:133149"
     SecStr          51..57
                     /sec_str_type="sheet"
                     /note="strand 15"
     SecStr          58..61
                     /sec_str_type="sheet"
                     /note="strand 16"
     SecStr          62..67
                     /sec_str_type="sheet"
                     /note="strand 17"
     SecStr          68..72
                     /sec_str_type="sheet"
                     /note="strand 18"
     SecStr          73..79
                     /sec_str_type="sheet"
                     /note="strand 19"
     SecStr          82..85
                     /sec_str_type="sheet"
                     /note="strand 20"
     SecStr          86..93
                     /sec_str_type="helix"
                     /note="helix 1"
ORIGIN
        1 pqitlwkrpl vtikiggqlk ealldtgadd tvleemalpg kwkprmiggi ggfvkvrqyd
       61 qipieixghk vigtvlvgpt paniigrnlm tqigctlnf
//
```

GO
(Gene Ontology)

Key Concepts: Databases

- Organize, manage, store and retrieve data

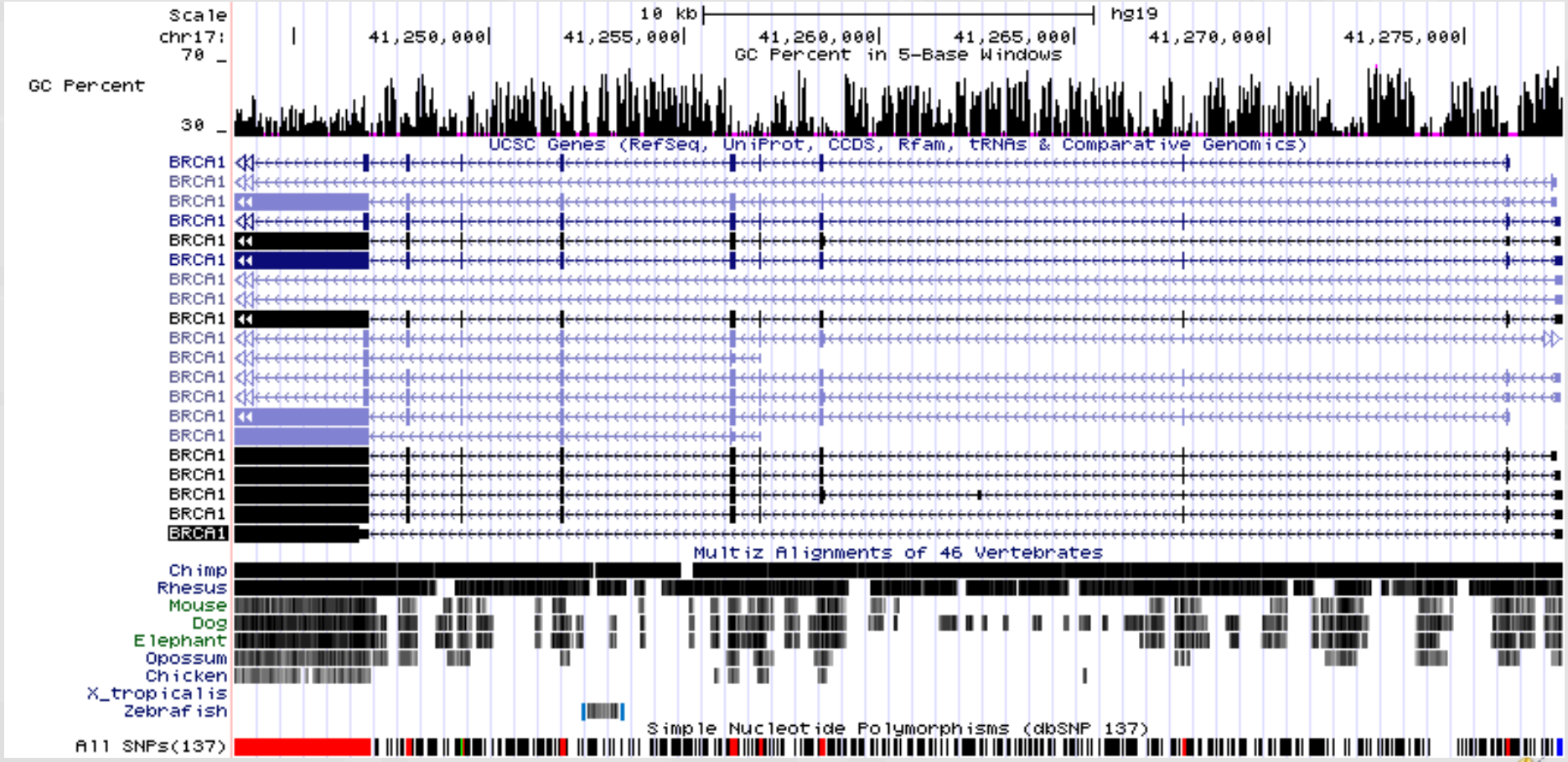


Key Concepts: Search

- Needs 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 techniques
  - Indexing
  - Replication
  - Distributed Access
  - Google secret “recipe”

Key Concepts: Integration

- The total sum is greater than the parts



Key Concepts: 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

Key Concepts: Standards

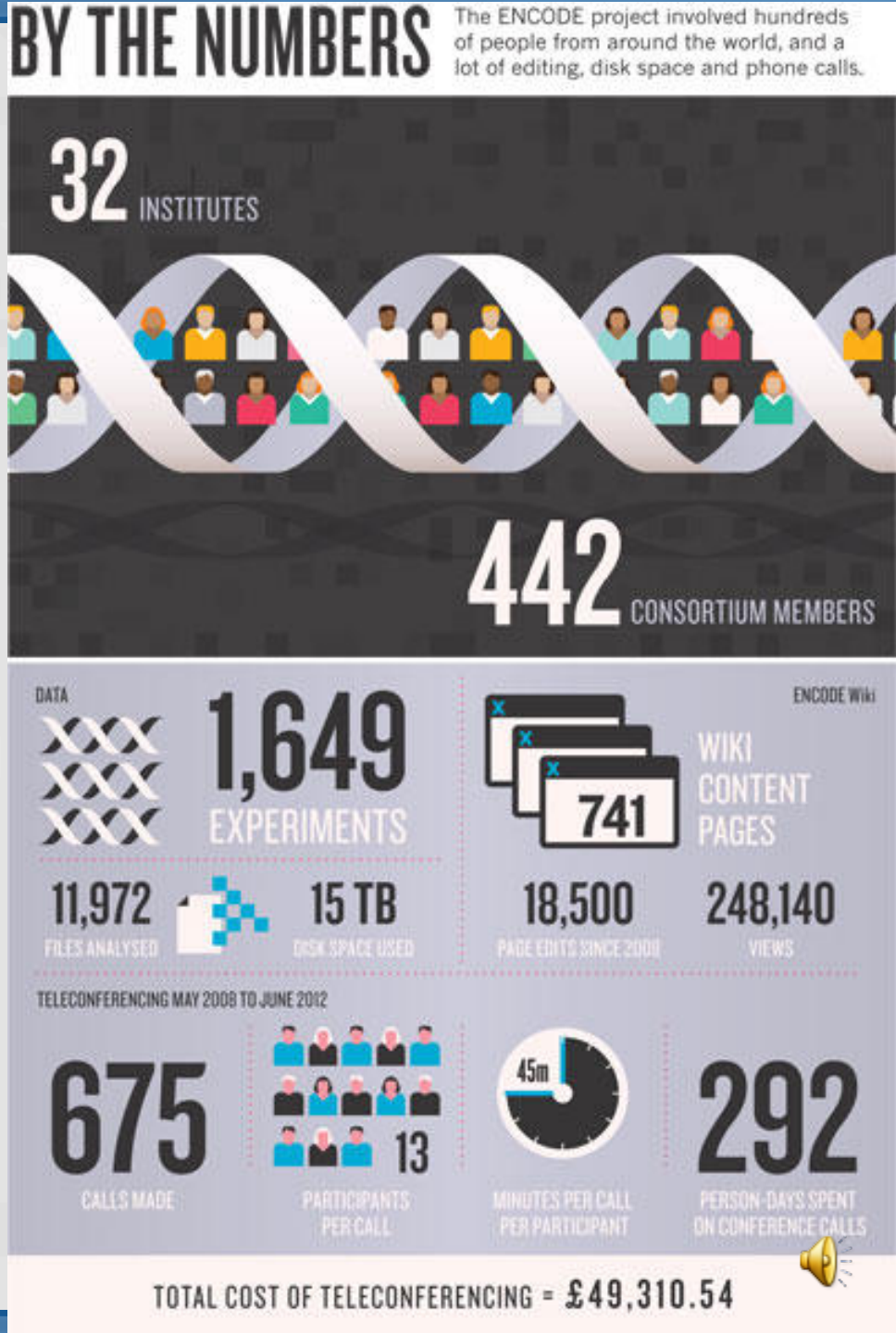
- Interoperability is key
  - MIAME: Minimum Information About a Microarray Experiment
  - FASTA format for sequence data
  - BED and GFF genomic annotation formats
  - HGNC gene naming standard
- Still a long way to go though
  - Most tools use their own input and output formats
  - Swapping formats can be tricky and time consuming
  - It's mostly an open market dynamic

Key Concepts: 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,…

Key Concepts: Big Data

- Individual experiments are large
- Many experiments are done on large samples
- Data accumulation rates are huge
- Acceleration is accelerating !



Key Concepts: Dynamic

- New data is added every SECOND !
- Some older data is deleted
- Some older data is updated
- Websites appear and disappear frequently
- New technology is introduced at a fast pace
- New discoveries are made at a fast pace
- Impossible to stand still !

Part II

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

BI and Disease Basic Science

- BI 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 in investigating host/ pathogen interaction

BI 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 risk profiles and allow patients to actively change 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 of survival predictions, reducing costs and improving overall care quality
- Disease registries can track diseases 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 care and prevention
- The computer (and the internet) will become an integral tool in the clinic

For more information

- The Web!
  - Wikipedia, start from “Bioinformatics” and explore, ad-infinitum !
  - Google
  - http://clinical-bioinformatics.com/
  - http://bioinformatics.ca/links_directory/index.php
- Tools:
  - The UCSC genome browser (http://genome.ucsc.edu/)
- Books:
  - Introduction to Bioinformatics by Arthur M. Lesk
- Journals:
  - Use PubMed
  - Read the accompanying article

In the End

Get Ready ! Bioinformatics is coming to your cl