Introduction to Bioinformatics

IMBB 2016

BecA-ILRI Hub, Nairobi

May 9 – 20, 2016

Joyce Nzioki

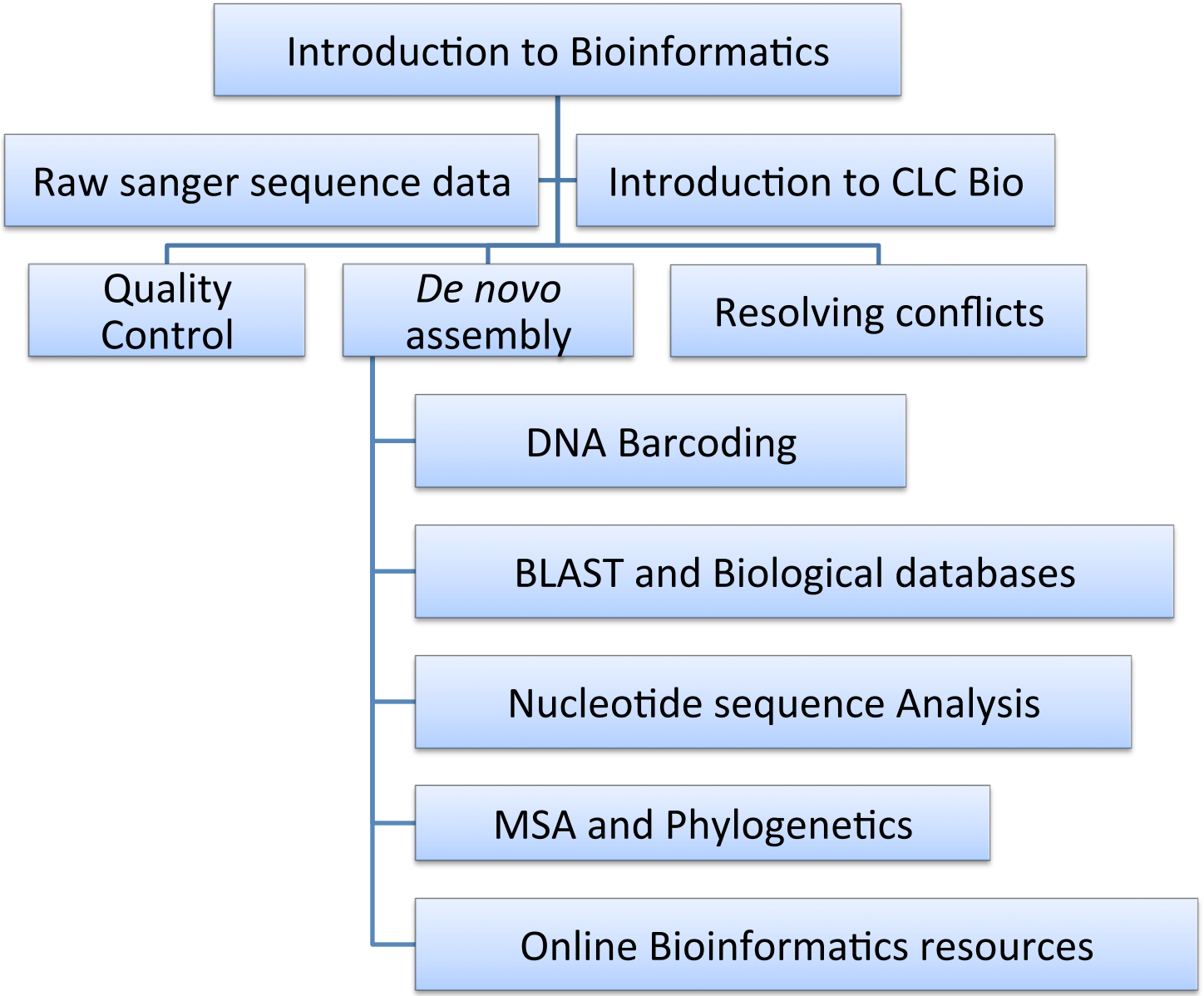
15th Anniversary biosciences eastern and central africa

NEPAD TRANSFORMING AFRICA

NC STATE UNIVERSITY

ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

Plan for the Week



What is Bioinformatics

- **Bioinformatics** is an interdisciplinary science that develops and improves on methods of storing, retrieving, organizing and analyzing biological data.

- This computational techniques are to solve biological problems and discover the wealth of biological information hidden in biological data.

Bioinformatics

The **design**, **construction** and **use** of software tools to **generate**, **store**, **annotate** and **analyse** data and information relating to Molecular Biology.

Here we consider the **use** of bioinformatics tools rather than their design and construction.

Here we consider the **access** and **analysis** of data and information items rather than the generation, storage or annotation.

What is Bioinformatics

Experiment

DATA

Analysis

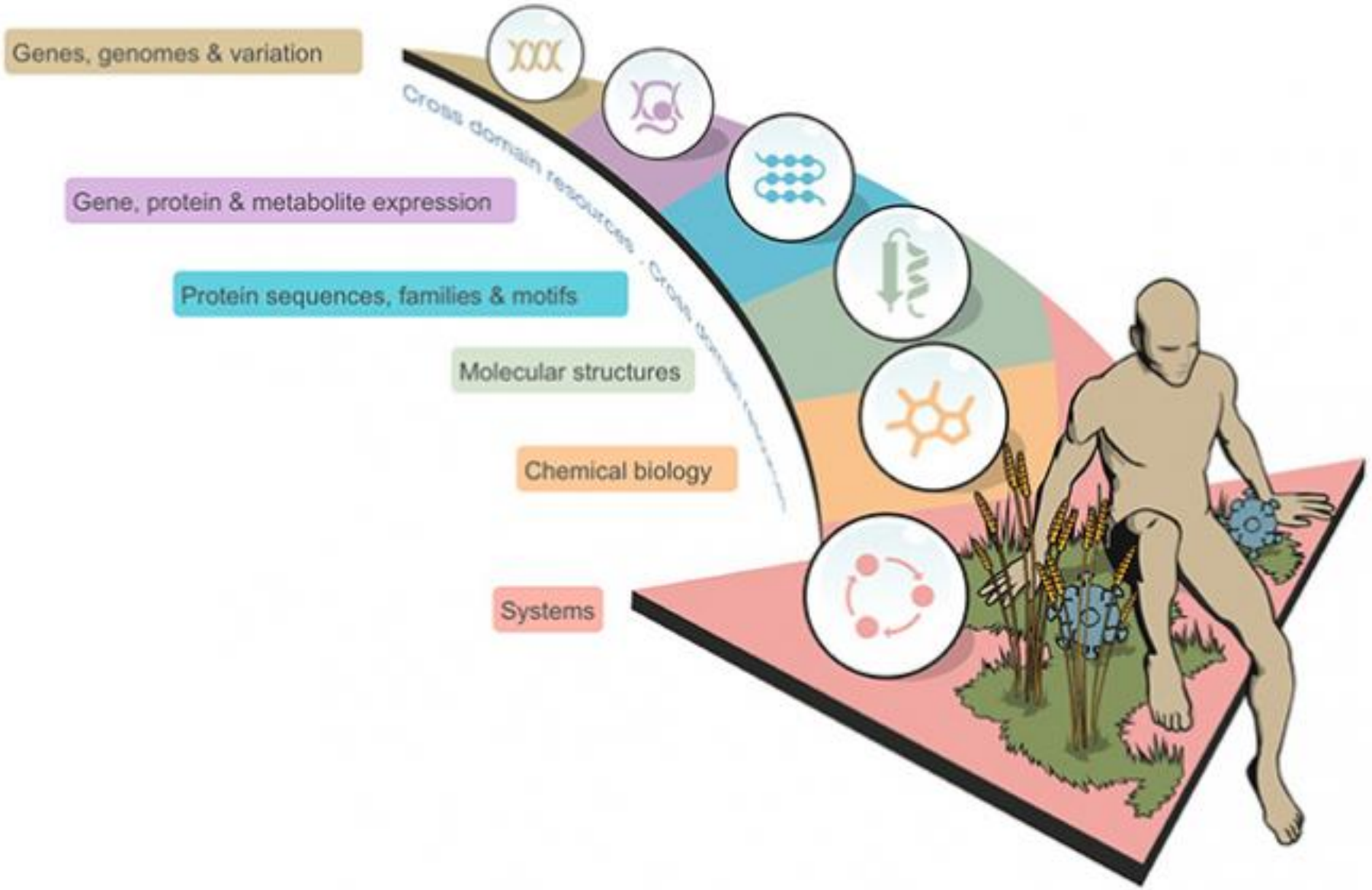
Sequence
Structure
Function
Evolution
Pathway
Interaction
Mutation
expression

RESULT

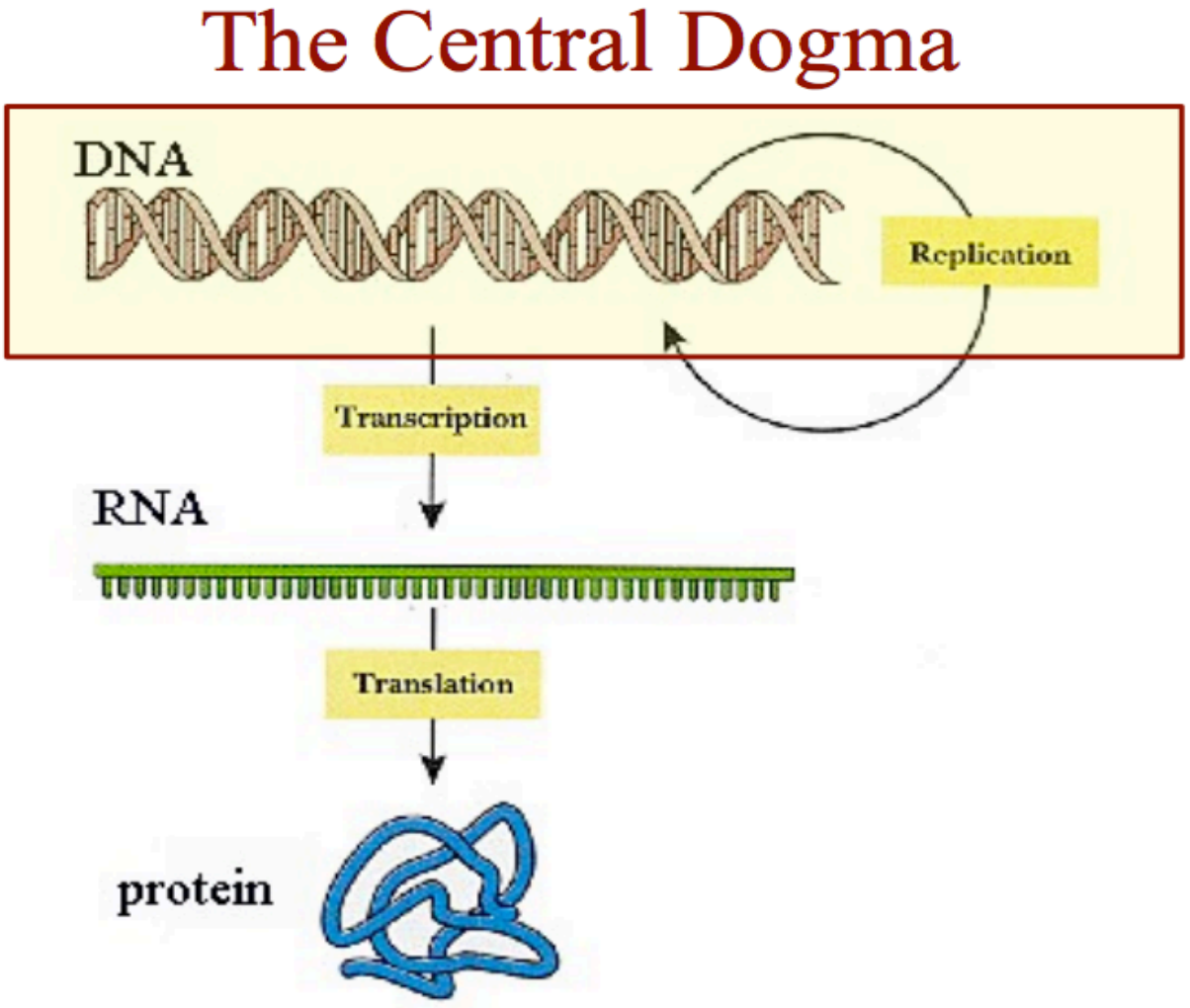
Hypothesis

Scope of Bioinformatics

1. **Storage and retrieval of biological data**.
2. **Sequence analysis**: Sequence alignments, database searches, genome assemblies, motif detection.
3. **Structural analysis**: protein / nucleic acid structure, visualization and analysis, classification, prediction.
4. **Genomics**: annotation, comparative genomics
5. **Functional genomics**: Transcriptome, proteome, interactome
6. **Analysis of biochemical networks**: metabolic networks, regulatory networks
7. **Phylogeny**


Genes, genomes & variation
Gene, protein & metabolite expression
Protein sequences, families & motifs
Molecular structures
Chemical biology
Systems
Cross domain resources

What “units of information” do we deal with in bioinformatics?



DNA

- A gene is a sequence of bases that caries information requires for constructing a particular protein.
- Bioinformatics is therefore applicable at the DNA level for:
    - Genes finding and their features
    - Gene expression
    - Genome Comparison
    - Database similarity searching
    - Multiple sequence alignment and Phylogeny
    - Primer design
    - Translation to protein

RNA

- Bioinformatics is applicable at the RNA level for:
  - Splice variants
  - Tissue specific expression
  - RNA structure
  - DNA chips, MicroArrays and expression array analysis

Protein

- Proteins are molecules composed of one or more polypeptides (polymer of amino acids).
- Bioinformatics is therefore applicable at the protein level for:
    - Protein sequence analysis
    - Protein structure prediction
    - Protein-protein interactions
    - Metabolic pathway analysis
    - Phylogeny

How does it look like on a computer

A cDNA sequence (reading frame)

>gi|14456711|ref|NM_000558.3| **Homo sapiens hemoglobin, alpha** 1 (HBA1), mRNA

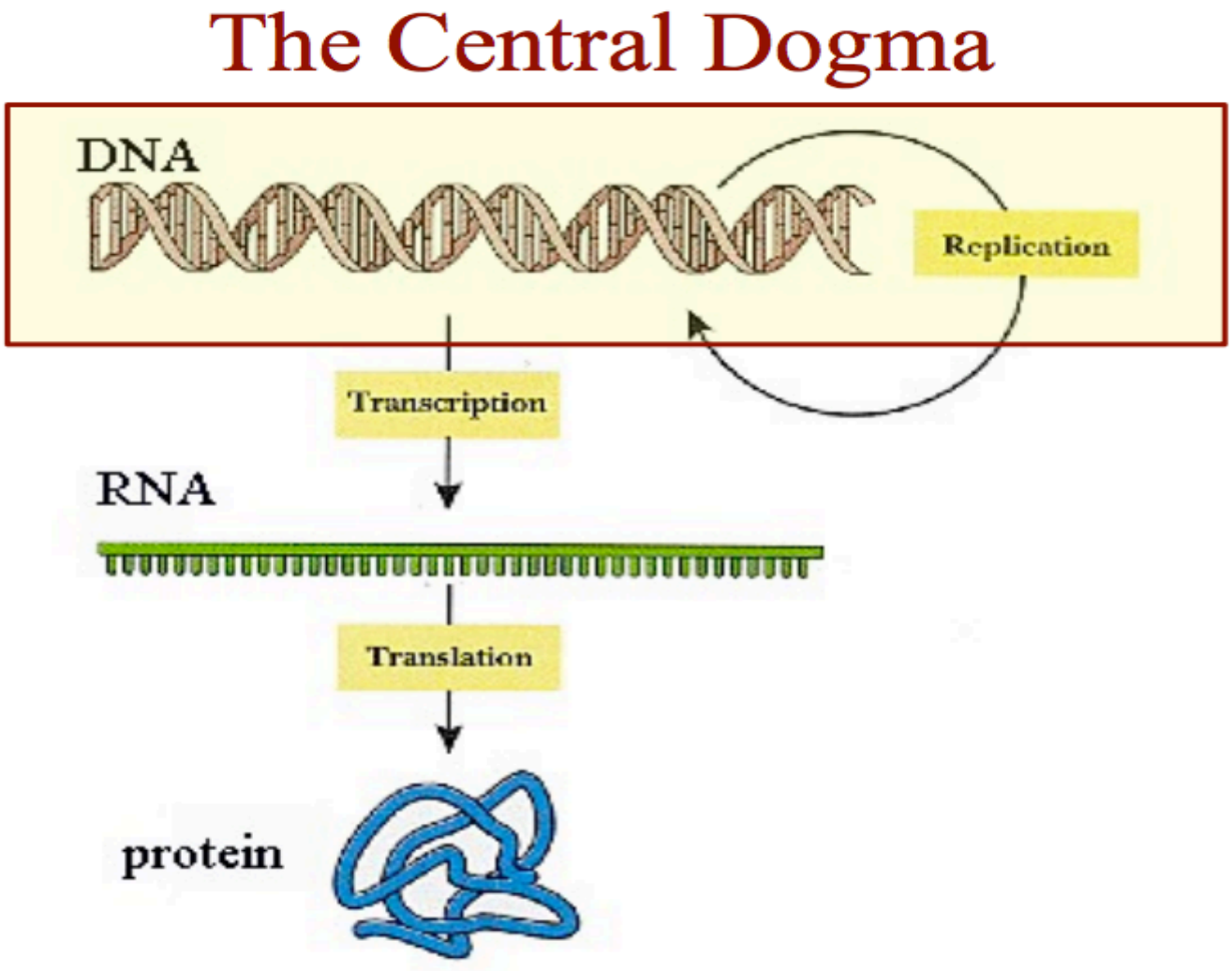
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACC**ATG**GTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCC
GCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCAC
CACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCG
ACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCAC
AAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGC
CGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACC
GT**TAA**GCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCC
GTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGC

A protein sequence

>gi|4504347|ref|NP_000549.1| **alpha 1 globin [Homo sapiens]**

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAH
VDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR

The integration of information learned about this three biological processes gives insight Into the biology of organisms



Who works in Bioinformatics

- Biologists
- Chemists
- Biochemists
- Computer scientists
- Mathematicians
- Statisticians
- Physicists
- Engineers

Why should molecular life scientist care about bioinformatics

Leon, postdoc

Goal: to understand what makes a normally harmless bacterium pathogenic in the lungs of people with cystic fibrosis.

Tasks: "I'm using a combination of transcriptomics, proteomics and metabolomics to understand these pathogenic changes better."

Barend, plant geneticist

Goal: to identify new crop strains resistant to drought, salt and fungal diseases.

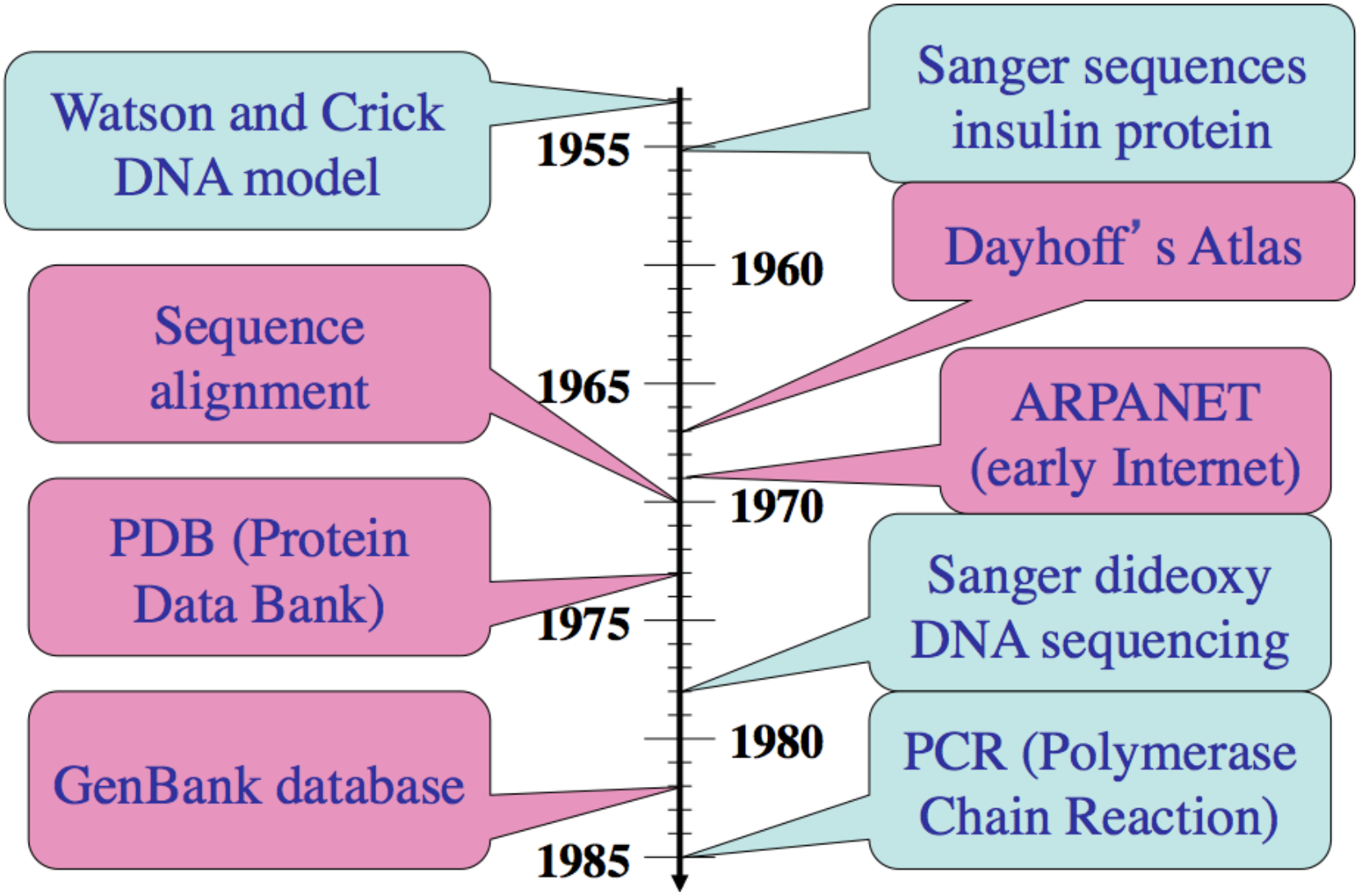
Tasks: "We're doing linkage studies to find out which genes are involved in resistance to different types of stress. We've got genomic and expression QTLs that we need to map on to well-characterised plants."

Ola, clinician–scientist

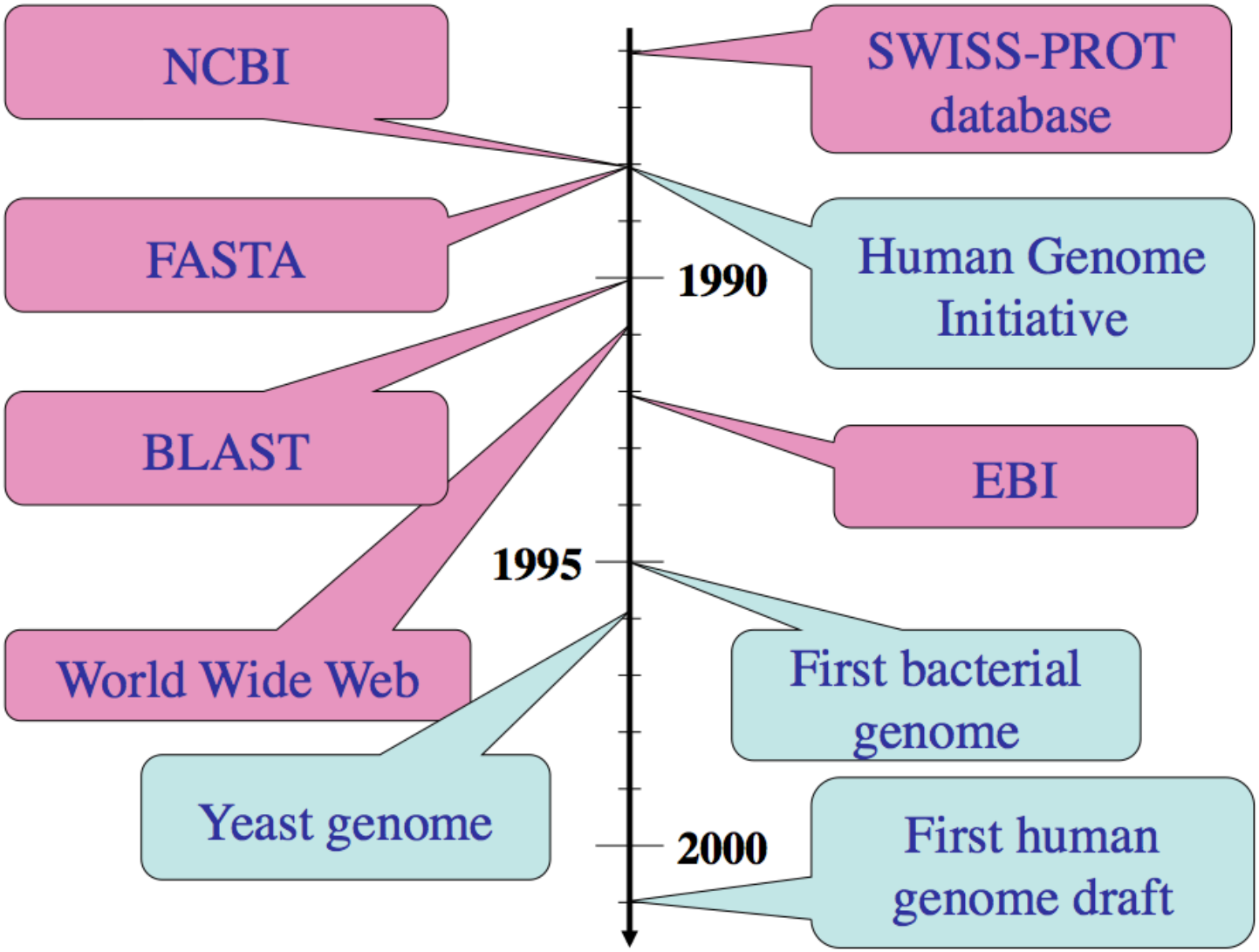
Goal: to identify proteomics-based biomarkers in urine for the early detection of bladder cancer

Tasks: "I do mass spectrometry of samples from patients coming in for biopsies. I've found a phosphoprotein that seems to be upregulated in some patients."

History of Bioinformatics



History of Bioinformatics



Applications of Bioinformatics

- Data driven biology
  - Functional genomics
  - Comparative genomics
  - Systems biology
- Molecular medicine
  - Diagnosis/prognosis from sequence expression
  - Gene therapy
  - Personalized therapy
  - Developing highly targeted drugs.
- Biotechnology
  - Bioengineering
  - Agriculture biotechnology.

Limitations of Bioinformatics

- Bioinformatics is a science of inference hence:
- Quality of bioinformatics predictions depends on the quality of data and sophistication of algorithms.
- Sequence data may have errors which subsequently leads to errors in downstream analysis.
- Many exhaustive algorithms cannot be used due to computational limitations.
- Trade-off between specificity and sensitivity

Why bioinformatics then

- In most cases biologics /wet lab is needed to validate bioinformatic predictions

- Bioinformatics can:
  - Reduce data to a small set of testable predictions
  - Assign a degree of confidence to each prediction

- The biologist will often have to choose the appropriate degree of confidence, depending on:
  - Cost of validating predictions.
  - Benefit expected from the right predictions.

- Data mining - the process by which testable hypothesis are generated regarding the function or structure of a gene or protein of interest by identifying homologs in better characterized organisms.

- Bioinformatics as *in sillico* biology:
  - Allows for exploration of domains that cannot be addressed manually e.g study of past evolutionary events / patterns.

Bioinformatics Revisited

- Representation/storage/retrieval/analysis of biological data concerning:
  - Sequence (DNA, protein, RNA)
  - Structures (protein, RNA)
  - Functions (protein, sequence)
  - Activity levels (mRNA, Protein, metabolites)
  - Networks of interactions (metabolic pathways, regulatory pathways, signaling pathways)
  - Textual information from biomedical literature

The End

Acknowledging EBI online courses for some slides