E

TWC/34/24
ORIGINAL: English
DATE: May 20, 2016

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
Geneva

TECHNICAL WORKING PARTY ON AUTOMATION AND COMPUTER PROGRAMS

Thirty-Fourth Session
Shanghai, China, June 7 to 10, 2016

BIOINFORMATICS

Document prepared by an expert from the Netherlands

Disclaimer: this document does not represent UPOV policies or guidance

The Annex to this document contains a copy of a presentation on “Bioinformatics” that will be made at the thirty-fourth session of the Technical Working Party on Automation and Computer Programs (TWC).

Fleur Gawehns, Researcher, Naktuinbouw

[Annex follows]

Bioinformatics

Dr. Fleur Gawehns

(Kees van Ettekoven)

Overview

- What is Bioinformatics?
  - and what not…
- Why do we need Bioinformatics?
  - Modern biology
  - ~omics revolution
- Examples of typical challenges in Bioinformatics

Bioinformatics is not…

… fixing a computer.

Microsoft® Office

… analysing excel sheets or compare Word documents.

Definition

Bioinformatics /ˌbaɪ.oʊ.ɪnfɚˈmætɪks/ is an interdisciplinary field that develops methods and software tools for understanding biological data.

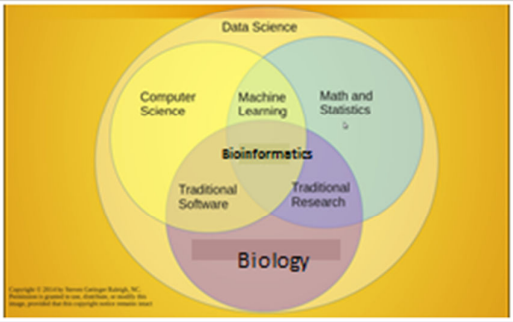
WIKIPEDIA
The Free Encyclopedia

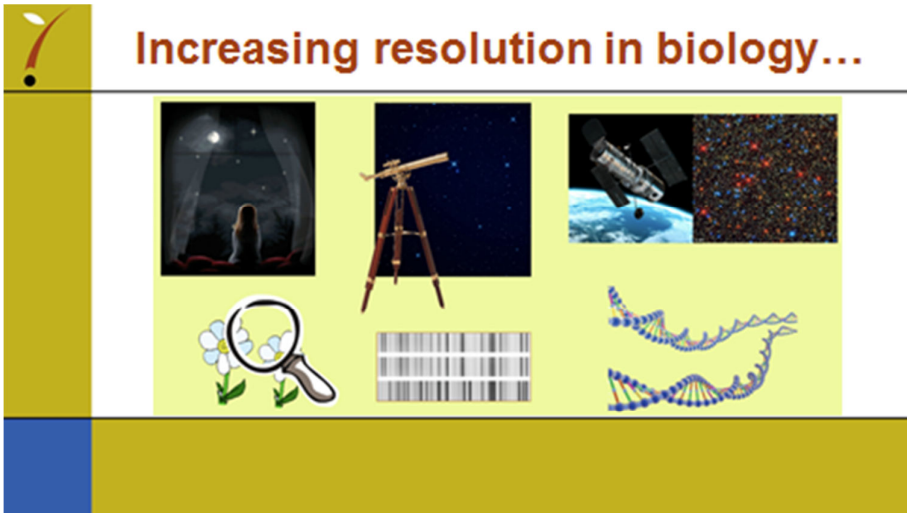
(Molecular) **bio** – informatics: bioinformatics is conceptualising biology in terms of molecules (in the sense of physical chemistry) and applying "***informatics techniques***" (derived from disciplines such as applied maths, computer science and statistics) to ***understand*** and ***organise*** the ***information*** associated with these molecules, on a ***large scale***. In short, bioinformatics is a management information system for molecular biology and has many ***practical applications***.

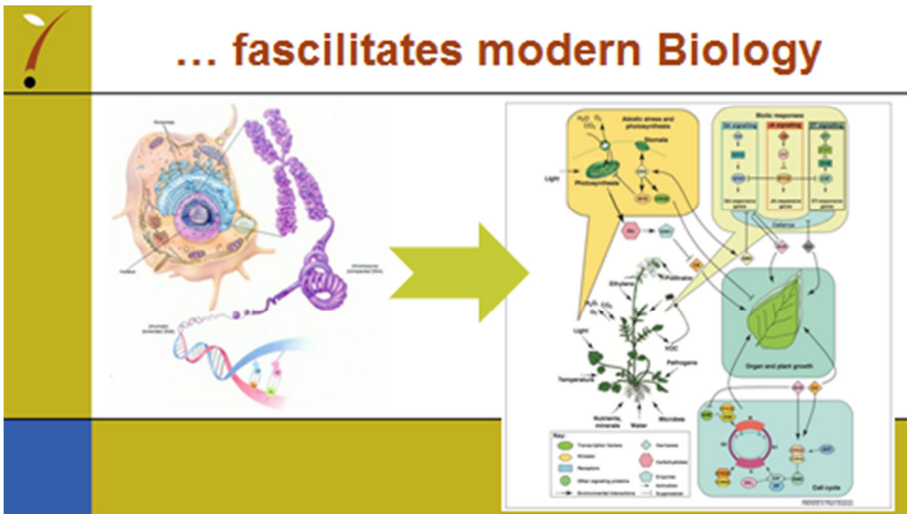
Luscombe, et al. 2001

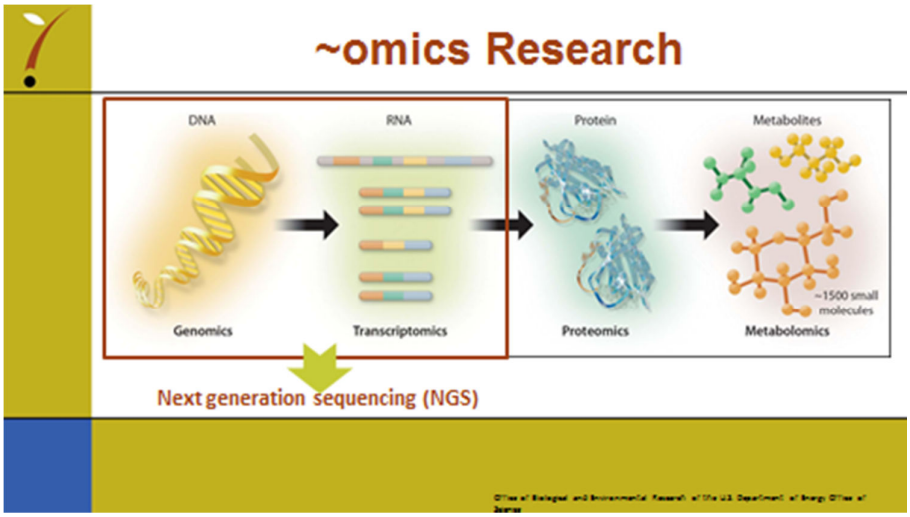
Bioinformatics: Biological data science

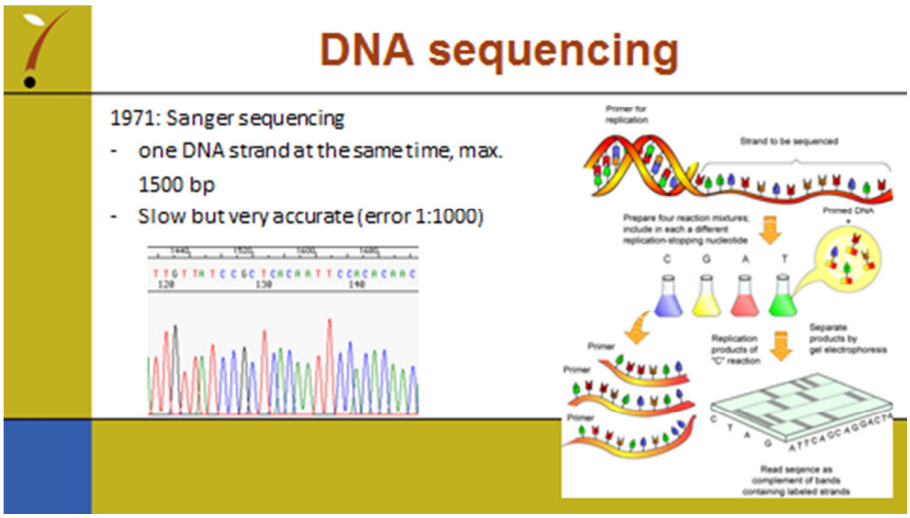
Bioinformatics: Biological data science

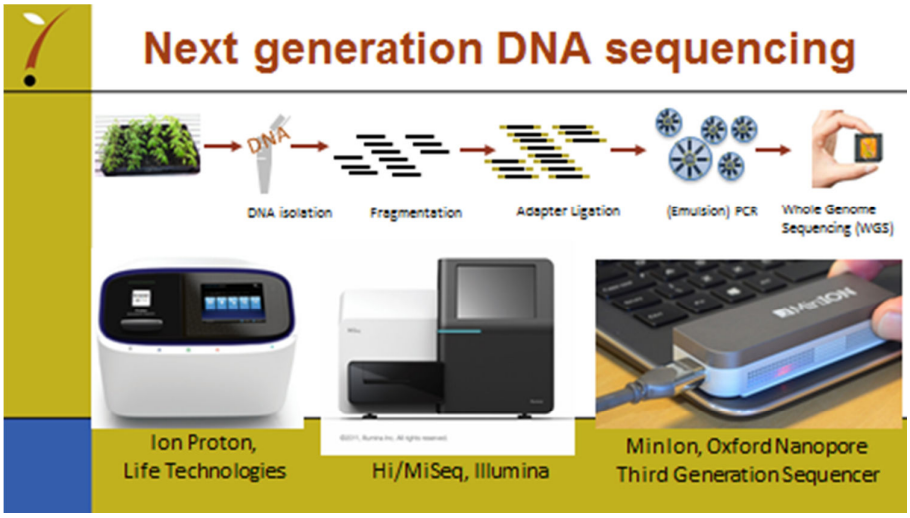


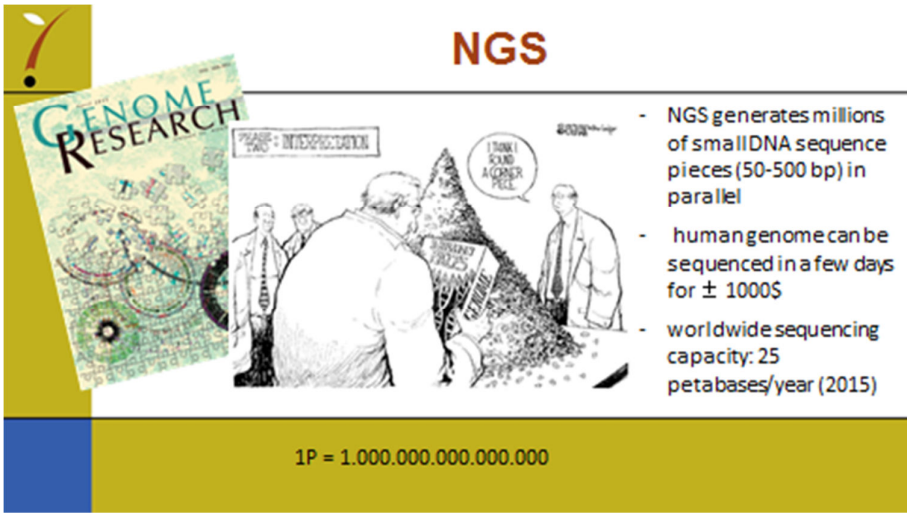

Increasing resolution in biology…


… fascilitates modern Biology

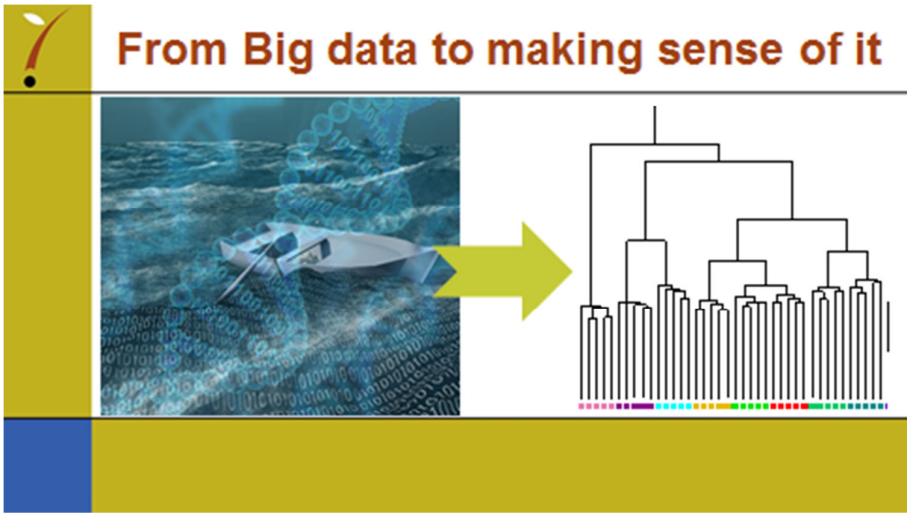

~omics Research
DNA
RNA
Protein
Metabolites
Genomics
Transcriptomics
Proteomics
Metabolomics
~1500 small molecules
Next generation sequencing (NGS)


DNA sequencing
1971: Sanger sequencing
- one DNA strand at the same time, max. 1500 bp
- Slow but very accurate (error 1:1000)


Next generation DNA sequencing
DNA
DNA isolation
Fragmentation
Adapter Ligation
(Emulsion) PCR
Whole Genome Sequencing (WGS)
Ion Proton, Life Technologies
Hi/MiSeq, Illumina
MinIon, Oxford Nanopore Third Generation Sequencer


NGS
GENOME RESEARCH
- NGS generates millions of small DNA sequence pieces (50-500 bp) in parallel
- human genome can be sequenced in a few days for ± 1000$
- worldwide sequencing capacity: 25 petabases/year (2015)
1P = 1.000.000.000.000.000


nature genetic
nature
THE POTATO GENOME
nature
THE TOMATO GENOME
Pineapple genome
Germatin of T2D loci
Wheat rust resistance


From Big data to making sense of it

How to assemble the puzzle?

- Puzzle piece = 1 sequence read
- Assembled puzzle = genome
- Picture on the puzzle box = reference genome
- 1 sequencing run generates ~20 mio reads!
- Reads from one or several puzzels in the same box (e.g. mix of plant and bacterial DNA in infected plants)

Reference-based mapping

Reads are aligned against a reference genome

Reference-based mapping

Important parameters used in common mapping algorithms

100%

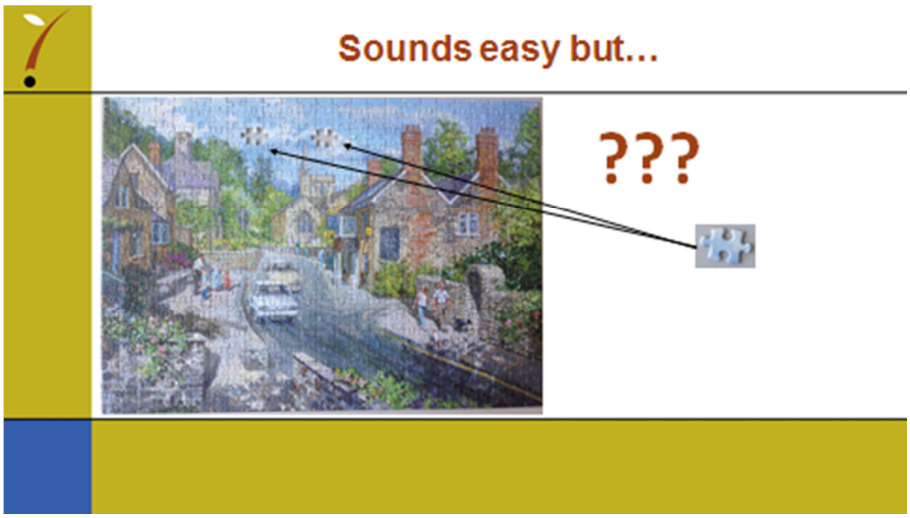
50%

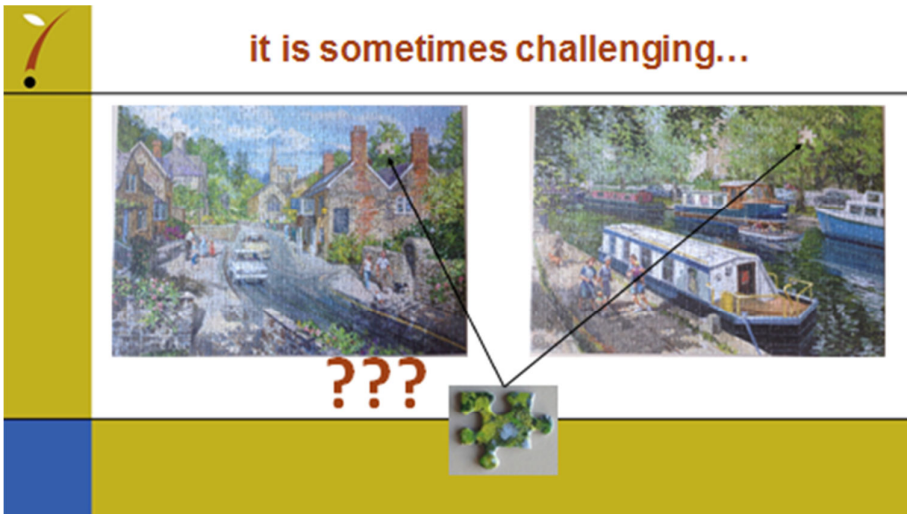
Length fraction: How much of the total read length should match the reference?

100%

50%

Similarity fraction: How many mismatches are allowed between the read and the reference?


Sounds easy but…
???

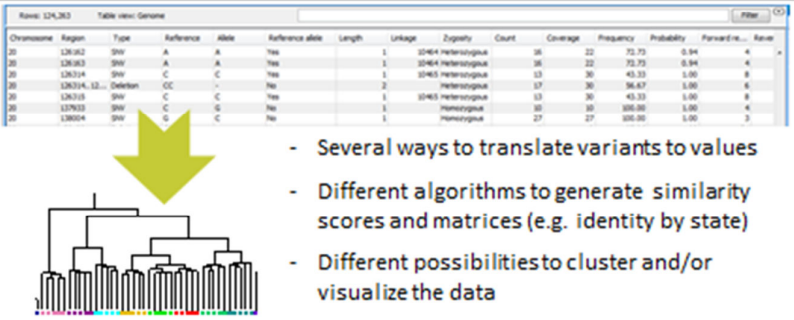

it is sometimes challenging…
???

Genome assembly accomplished but…

Sequencing error vs. sequence variant

- Error rate for NGS still relatively high
- Illumina error rate 0,1-1%
- Both random and context specific errors
- Coverage is not evenly distributed over the genome (some variants are covered less than others)
- Polyploidy!

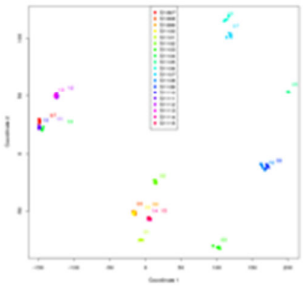
From variants to information



- Several ways to translate variants to values
- Different algorithms to generate similarity scores and matrices (e.g. identity by state)
- Different possibilities to cluster and/or visualize the data

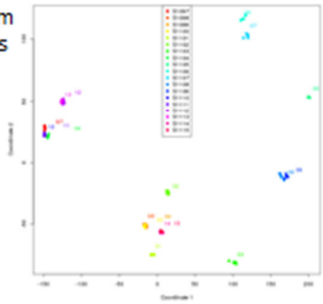
Examples of Statistics in Bioinformatics

- How many varieties do you need to cover the genomic space of a plant species?
- How many individuals do you need to define a variety?
- Which distance is allowed between individuals to define them the same/ a new variety?
- Mapping: Where is the best match for a certain read? How to deal with insertions/deletions (gap open)

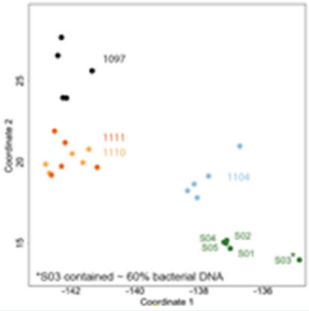


Examples of Statistics in Bioinformatics

- Variant Calling: Probability calculations/Maximum Likelihood (how big is the chance that a variant is caused by a nucleotide difference and not by a sequencing error?), Slide 22
- How much information content does a specific variant have?
- How to calculate distances/similarities between varieties on the basis of SNPs?, Slide 23



Information example



- Naktuinbouw tested the variety identity of the winning tomatoes in a Giant Veg growing competition on behalf of a Flower Show
- NGS and a Bioinformatics pipeline were performed from samples S01-S05
- Distance relative to the samples in a tomato database was visualized in an MDS
- S01 to S05 belonged to the same variety
- But: In one sample (S03) DNA present from a growth-promoting bacterium

Summary

- Due to the ~omics revolution, modern biology is turning into a data science
- Bioinformatics delivers algorithms and models to get the most out of this data
- Bioinformatics is a relative young science that develops fast; so does NGS
- Quality of the Bioinformatics analysis determines the reliability of the obtained information

Where to place a bioinformatician?

[End of Annex and of document]