Introduction to Bioinformatics

a Data Scientist perspective

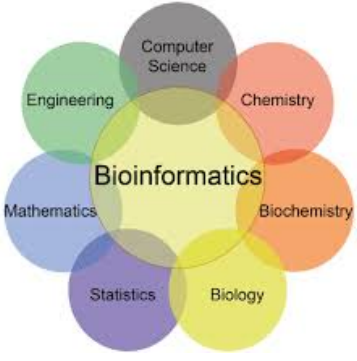
Dina Machuve (PhD)

PythontzMLS2018 @ IndabaX-Tanzania
dmachuve@python.or.tz

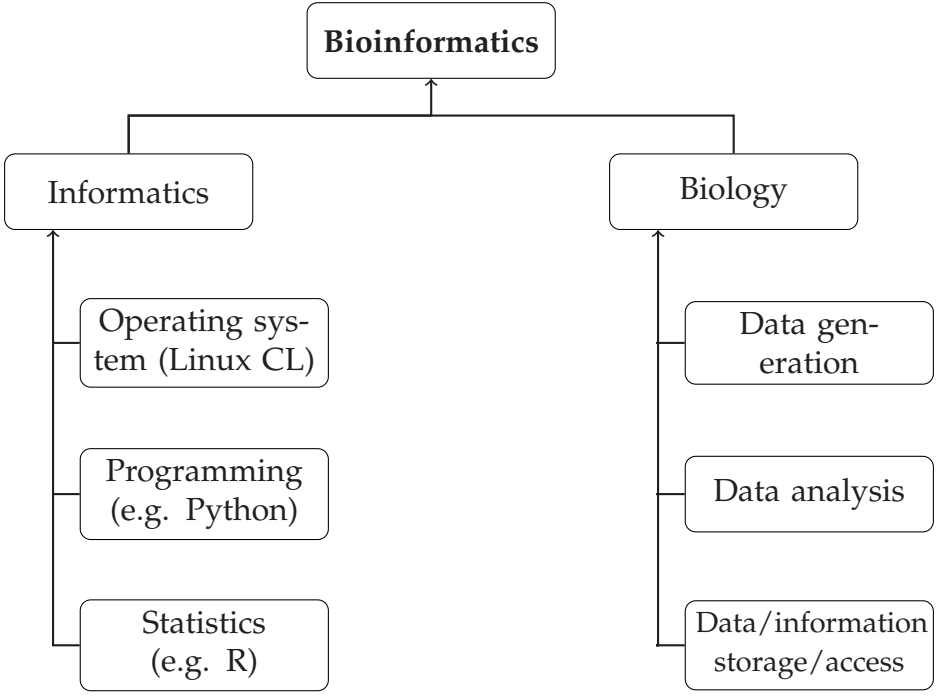
04 April 2018

What is Bioinformatics?

the science of storing, retrieving and analysing large amounts of biological information (EMBL-EBI,2018)



- highly interdisciplinary field
- different types of specialists: biologists, molecular life scientists, computer scientists and mathematicians


Bioinformatics
Informatics
Biology
Operating system (Linux CL)
Programming (e.g. Python)
Statistics (e.g. R)
Data generation
Data analysis
Data/information storage/access

Molecular life scientists

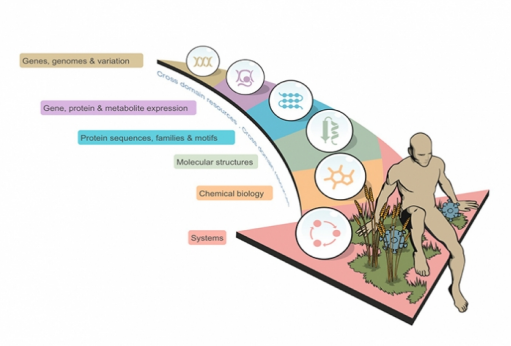
Leon, postdoc
Goal: to understand what makes a normally harmless bacterium pathogenic in the lungs of people with cystic fibrosis.
Tasks: "I'm using a combination of transcriptomics, proteomics and metabolomics to understand these pathogenic changes better."

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."

- Life sciences have become increasingly data driven
- Molecular life scientists work with bioinformatics experts to design, analyse and interpret their experiments (EMBL-EBI,2018)

Big data explosion

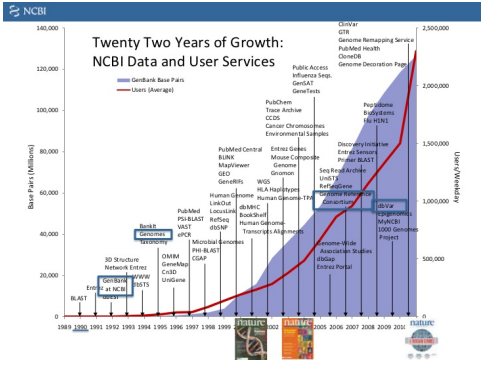


Human Genome

- 3.3 billion base pairs
- 23 chromosome pairs of size ranging up to a few 100 millions of base pairs
- Working 8 hours a day and 200 days a year, it would take
- $3,300,000,000/60$ x 60 x 8 x 200 $=$ 572 years to analyse a single human DNA molecule

Biological data is big data

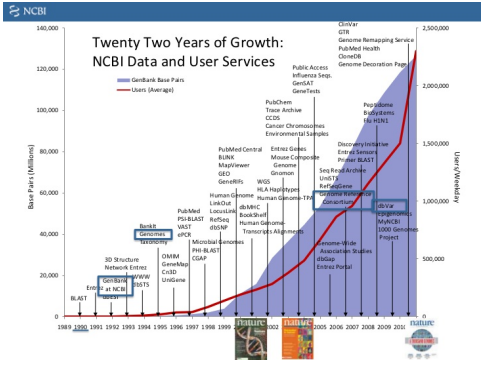
European Bioinformatics Institute (EMBL-EBI): total disk capacity 75 petabytes (Dec 2015)



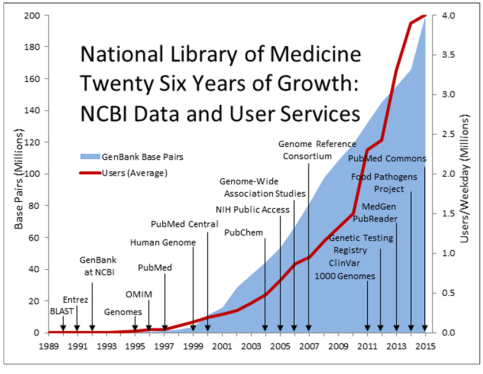
In 2010

Biological data is big data

European Bioinformatics Institute (EMBL-EBI): total disk capacity 75 petabytes (Dec 2015)



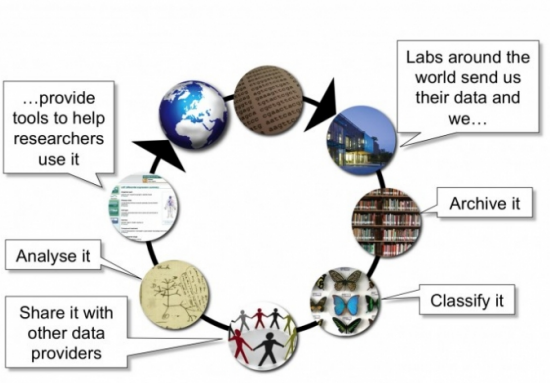
In 2010



In 2015

Bioinformatics centres of excellence

- A few in number worldwide with responsibility to collect, catalogue and provide open access to published biological data.
- Among these centers include:
  - The EMBL-European Bioinformatics Institute (EMBL-EBI)
  - The US National Center for Biotechnology Information (NCBI)
  - The National Institute of Genetics in Japan (NIG)



Data sharing collaborations

- Established to manage the public record of different biological data types
- researchers submit their sequences to one of the public databases before submitting the paper, use accession number for citation in the paper
- examples of global collaborations established to manage the public record of different biological data types:

Data type	**Collaboration**
Nucleotide sequences	International Sequence Database Collaboration
Protein sequences	UniProt Consortium
Macromolecular structures	Worldwide Protein Data Bank
Molecular interactions	The International Molecular Exchange Consortium
Protein identifications	The ProteomeXchange Consortium
Metabolomics data	Coordination of Standards in Metabolomics
Genomic and clinical data	Global Alliance for Genomics and Health

Who owns the Data?

- the bioinformatics community has championed open data sharing
- if public money is being spent on research (in any field), the data from that research should be made publicly available for others to make use of
- researchers contribute to the public record by submitting data to appropriate public databases through dedicated submission tools.

There are two major reasons that you might frequently come across for not openly sharing data:

1. To protect the individual: any data relating to identifiable individuals is sensitive and should be protected by ethical policies
2. To protect intellectual property or other competitive information. If data are potentially commercially applicable

Value of big data in life sciences

- Novel discoveries for healthcare, agriculture, food security ($>$ 1.2 million species of plants & animals)
- Disease surveillance and response
- Management of health data (EHRs and experimental data) can inform diagnosis and treatment –Precision Medicine
- Data can save lives!
- Challenges: volume, velocity, variety

Big Data Analytics in Biology

BIOINFORMATICS

Big data versus the big C

The torrents of data flowing out of cancer research and treatment are yielding fresh insight into the disease.

BY NEIL SAVAGE

In 2013, geneticist Stephen Elledge answered a question that had puzzled cancer researchers for nearly 100 years. In 1914, German biologist Theodor Boveri suggested that the abnormal number of chromosomes — called aneuploidy — seen in cancers might drive the growth of tumours. For most of the next century, researchers made little progress on the matter. They knew that cancers often have extra or missing chromosomes or pieces of chromosomes, but they did not know whether this was important or simply a by-product of tumour growth — and they had no way of finding out.

S66 | NATURE | VOL 509 | 29 MAY 2014

OUTLOOK BIG DATA IN BIOMEDICINE

PERSPECTIVE

Sustaining the big-data ecosystem

Organizing and accessing biomedical big data will require quite different business models, say **Philip E. Bourne**, **Jon R. Lorsch** *and* **Eric D. Green**.

Biomedical big data offer tremendous potential for making dis-

PHOTOGRAPHS: NATIONAL HUMAN GENOME RESEARCH INSTITUTE

These caveats notwithstanding, more details of data usage are needed to inform funding decisions. Over time, such data usage patterns could tell us how best to target annotation and curation efforts, establish which data should receive the most attention and therefore incur the largest cost, and determine which data should be kept in the longer term. The cost of data regeneration can also influence decisions about keeping data.

Funders should encourage the development of new metrics to ascertain the usage and value of data, and persuade data resources to provide such statistics for all of the data they maintain. We can learn here from the private sector: understanding detailed data usage patterns through data analytics forms the basis of highly successful companies

TECHNOLOGY FEATURE

THE BIG CHALLENGES OF BIG DATA

As they grapple with increasingly large data sets, biologists and computer scientists uncork new bottlenecks.

Extremely powerful computers are needed to help biologists to handle big-data traffic jams.

BY VIVIEN MARX

Biologists are joining the big-data club. With the advent of high-throughput genomics, life scientists are starting to grapple with massive data sets, encountering challenges with handling, processing and moving information that were once the domain of astronomers and high-energy physicists[1].

With every passing year, they turn more often to big data to probe everything from the regulation of genes and the evolution of genomes to why coastal algae bloom, what and how the genetic make-up of different cancers influences how cancer patients fare[2]. The European Bioinformatics Institute (EBI) in Hinxton, UK, part of the European Molecular Biology Laboratory and one of the world's largest biology-data repositories, currently stores 20 petabytes (1 petabyte is 10^{15} bytes) of data and back-ups about genes, proteins and small molecules. Genomic data account for 2 petabytes of that, a number that more than doubles every year[3] (see 'Data explosion').

This data pile is just one-tenth the size of the data store at CERN, Europe's particle-physics

Data Science challenges in Africa - **The opportunity!**

- Work on unique fauna and flora –African data underrepresented
- Have to deal with ever-increasing data size and complexity
- IT challenges include:
  - Data transfer from generation site
  - Internet access
  - Adequate IT infrastructure for storage and processing
  - Long term secure storage
  - Training people at different levels on the use of this
- Researchers usually don't budget for data
- Not enough bioinformaticians or data scientists
- Meta data is not well curated, data quality and accuracy is not a high priority for clinicians and some researchers

Sequence Comparison

Types of Sequences

- Genetic data is digital.
- DNA, *deoxyribonucleic acid* is a medium to transmit information from generation to generation.
- DNA is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms.
- Nucleic accids: DNA, RNA and Proteins are sequences. These biomolecules actually depict the information content in the living system.
- DNA sequences are associated with the four-letter DNA alphabet A, C, G, T, where they stand for the nucleic acids or nucleotides adenine, cytosine, guanine and thymine respectively.

Sequence Comparison

- Sequence comparison/alignment is a way of arranging the sequences of DNA, RNA and protein to identify regions of similarity that be a consequence of structural, functional and evolutionary relationships between the sequences.
- Sequence alignment in particular, enables the following tasks:
  1. assigning functions to unknown proteins
  2. determining relatedness of organisms
  3. identifying structurally and functionally important elements
  4. making predictions about the 3D structure

Why compare sequences?

Consider a long sequenced string of DNA, exact string matching applications are of secondary interest in bioinformatics.

- measurements obtained in the laboratory are reliable only up to a certain degree
- even the process of typing data into computer storage must be seen as error prone up to a certain degree
- nature itself is a source of non-exact matching between e.g. homologous strings in a sense that evolution may mutate bases, insert fresh bases, or delete existing bases

Sequence Alignment

Look at two short nucleotide sequences below. The sequences are of the same length, and there is only one way to align them, if one does not allow gaps in alignments

x:	T	A	C	C	A	G	T
y:	C	C	C	G	T	A	A

Que. 1: What does matching of strings mean in case that not only exact coincidence of characters counts?

Que. 2: How to exactly measure the quality of an alignment?

Que. 3: How to efficiently compute alignments with maximum score?

Quality of an alignment

If we allow gaps, there are many possible alignments. In particular, the following alignment seems to be much more informative than the preceding one

x:	T	A	C	C	A	G	T	-	-
y:	C	-	C	C	-	G	T	A	A

Another possible alignment that also looks reasonable is

x:	T	A	C	C	A	G	T	-	-
y:	-	-	C	C	C	G	T	A	A

- The two alignments indicate that the subsequence *CCGT* may be an evolutionarily conserved region
- How can one choose between the two alignments? To answer these questions we need to be able to score any possible alignment. Then the alignments that have the highest score are by definition the best or optimal ones (there may be more than one such alignments).

Sequence Alignment Challenge

- The introduction of a spacing symbol in one of the strings might be interpreted as a deletion of a formerly present character from this string, but also as an insertion of a fresh character into the other string.
- Quality refers to a convention on how strongly to reward identities between characters and how to score *mutations, insertions, and deletions*.
- Counting the frequencies for each pair (x, y) of characters or spacing symbol does not help, not even for relatively short strings to find whether the first string better aligns to the second or the third one.
- Therefore, we must develop efficient algorithms that compute optimal alignments

Types of alignment

- The procedure of comparing two (pair-wise alignment) or more multiple sequences is to search for a series of individual characters/patterns that are in the same order in the sequences.
- There are two types of alignment: *Local* and *Global*
- **Global sequence alignment** approach spans the entire length of all query sequences. The tool for global alignment is based on Needleman-Wunsch algorithm.
- **Local sequence alignment** span on identified regions of similarity within a long sequence. The tool for local alignment is based on Smith-Waterman algorithm.
- Both algorithms are derivatives from **Dynamic Programming** algorithm, an optimal alignment algorithm

Dynamic Programming: Global Alignment

- Considerations only to DNA sequences, thus assuming that $Q = A, C, G, T$
- Assume a *linear gap model* that is, $s(-, a) = s(a, -) = -d$ for $a \epsilon Q$, with $d > 0$, so that the score of a gap region of length L is equal to $-dL$
- Suppose we are given two sequences $x = x_1 x_2 \dots x_i \dots x_n$ and $y = y_1 y_2 \dots y_j \dots y_m$. We construct an $(n + 1)$ x $(m + 1)$ matrix F.
- Its $(i, j)th$ element $F(i, j)$ for $i = 1, \dots, n, j = 1, \dots, m$ is equal to the score of an optimal alignment between $x_1 \dots x_i$ and $y_1 \dots y_j$.
- The element $F(i, 0)$ for $i = 1, \dots, n$ is the score of aligning $x_1 \dots x_i$ to a gap region of length i.

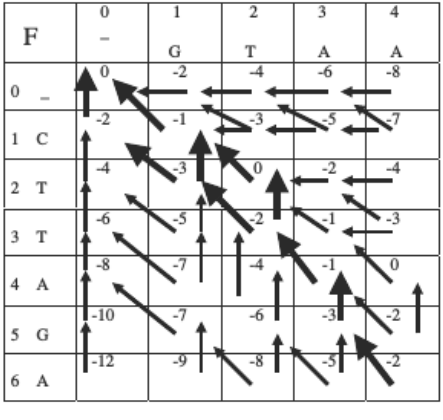
Global Alignment: Score

If $F(i-1, j-1), F(i-1, j)$ and $F(i, j-1)$ are known, $F(i, j)$ is clearly calculated as follows

$$F(i,j) = max \begin{cases} F(i-1,j-1) + s(x_i, y_j), \\ F(i-1,j) - d, \\ F(i,j-1) - d. \end{cases}$$

Global Alignment: Example

Let $x = CTTAGA$, $y = GTAA$, and suppose that we are using the scoring scheme: $s(a, a) = 1, s(a, b) = -1$, if $a \neq b$, and $s(-, a) = s(a, -) = -2$. The corresponding matrix F with pointers is shown



Global Alignment: Optimal alignments

Tracing back the pointers gives the following three optimal alignments

```
x: C T T A G A
y: G - T A - A

x: C T T A G A
y: G T - A - A

x: C T T A G A
y: - G T A - A
```

with score -2. The corresponding paths through the matrix F are shown with thick arrows.

Dynamic Programming: Local Alignment

- To find all pairs of subsequences of two given sequences that have the highest-scoring alignments
- Any such subsequence of a sequence $x_1x_2 \dots x_n$ has the form $x_ix_{i+1} \dots x_{i+k}$ for some $1 \leq i \leq n$ and $k \leq n - i$.
- We construct an $(n+1)$ x $(m+1)$ -matrix as in the previous section, but the formula for its entries is slightly different

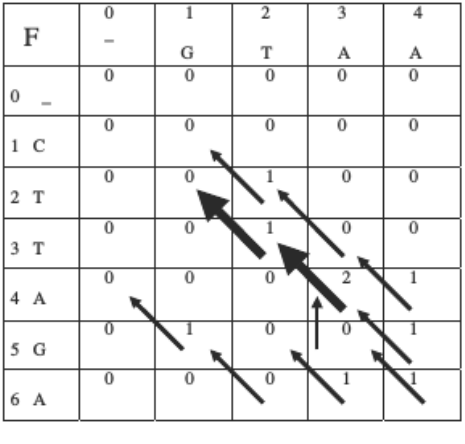
$$F(i,j) = max \begin{cases} 0, \\ F(i-1,j-1) + s(x_i, y_j), \\ F(i-1,j) - d, \\ F(i,j-1) - d. \end{cases}$$

Local Alignment: Example

For the sequences from previous example, the only best local alignment is

x: T A
y: T A

and its score is equal to 2



References

1. Brooksbank, C. and Cowley, A.(2018), Bioinformatics for the terrified, EMBL-EBI Train Online
2. Isaev, A. (2006) Introduction to Mathematical Models in Bioinformatics, Springer
3. Marx, V. (2013). Biology: The big challenges of big data.
4. Mulder, N. (2017), "H3ABioNet –enabling bioinformatics and big data research in Africa", CHPC Conference
5. Savage, N. (2014). Big data versus the big C. Nature, 509(7502), S66.