Biology + Computer Science = Bioinformatics?

Mario Inostroza-Ponta

mario.inostroza@usach.cl
Departamento de Ingeniería Informática
Universidad de Santiago de Chile

July, 2015

Disclaimers

Disclaimer #1

I am not a biologist, so I do apologize for biological **errors** and **horrors** as well...

Disclaimers

Disclaimer #1

I am not a biologist, so I do apologize for biological **errors** and **horrors** as well...

Disclaimer #2

I am a computer scientist “working” with biologists and people from other disciplines

Disclaimers

Disclaimer #1

I am not a biologist, so I do apologize for biological **errors** and **horrors** as well...

Disclaimer #2

I am a computer scientist “working” with biologists and people from other disciplines

Disclaimer #3

This talk is based on my own experience in the field

Common situations in a bioinformatic project

- Situation 1: a biologist asking for a system to help his/her research

Common situations in a bioinformatic project

- Situation 1: a biologist asking for a system to help his/her research
  - CS: What are you looking for?,
  - B: I am looking for X that has a shape Y,
  - CS: ok, using method M will give you all the matches of X with shape Y
  - B: that's cool!, **but sometimes** its shape changes to R,S,T and W
  - CS: ok, is there any rule that governs its shape?
  - B: No
  - FRUSTRATION!

Common situations in a bioinformatic project

- Situation 2: A CS presenting his/her solution for a given problem

Common situations in a bioinformatic project

- Situation 2: A CS presenting his/her solution for a given problem
  - CS: I built this system that implements algorithm 1 for your problem
  - B: cool!
  - CS: you have to set parameters A, B, C and D.
  - B: ok, but is there any **default** value that I can use?
  - CS: yes, I already put them in the algorithm, but you have to be aware of them
  - B: ok... (and never again the parameters were changed...)
  - POOR RESULTS!

Common situations in a bioinformatic project

- Situation 3: Two research groups trying to start a collaboration in bioinformatics

Common situations in a bioinformatic project

- Situation 3: Two research groups trying to start a collaboration in bioinformatics
- Situation 3.1:
  - CS: I am an expert in X, so you have a problem for me?
  - B: ok, maybe there is one situation, let me explain you the problem...
  - CS: No, **I don't need to understand everything**, just tell me where I need to apply X and...
  - FAILED!

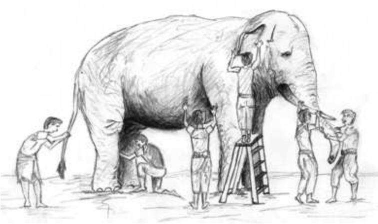
Common situations in a bioinformatic project

- Situation 3: Two research groups trying to start a collaboration in bioinformatics
- Situation 3.1:
  - CS: I am an expert in X, so you have a problem for me?
  - B: ok, maybe there is one situation, let me explain you the problem...
  - CS: No, **I don't need to understand everything**, just tell me where I need to apply X and...
  - FAILED!
- Situation 3.2:
  - B: I work in X and my data represents Y. I am looking for Z
  - CS: it sounds to me that we could model your problem as A or B. Let me explain both...
  - B: no, I don't have time to look into details, do you already have something to solve my problem?
  - FAILED!

Outline

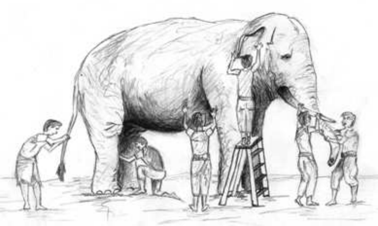
1. What is bioinformatics?
2. History
3. Examples of common mistakes
4. Education in Bioinformatics
5. Final comments

Bioinformatics?



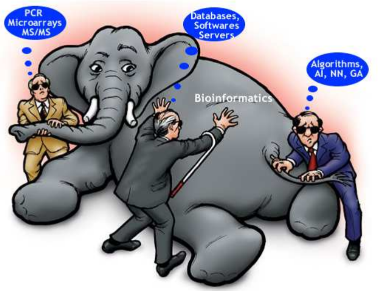
Indian tale about 5 blind men describing an elephant

Bioinformatics?



Indian tale about 5 blind men describing an elephant

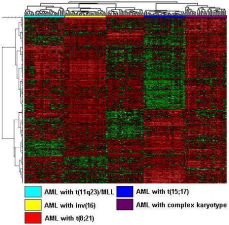
we will describe bioinformatics depending on our expertise and application field

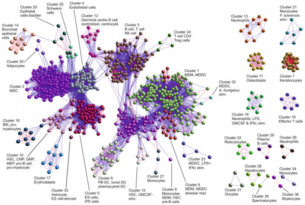


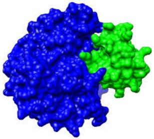
Bioinformatics

gene
phylogenetic
TyrRS
transfer
types
HGT
Archaea
bacterial
genome
genomes
evolutionary
evolution
errors
related
analysis
parasites
methods
homing
life
elements
biased
signal
ribosomal
sequencing
automated
introns
Group
free
horizontal
history
reconstruction
gene-calling
even
genetic
clades
help
may
alleles
expected
molecular
similar
show
positions
created
approaches
endonucleases
process
discuss
different
three
two
phylogeny
many
one
Eukarya
lineages
cycle
common
rather
Genome
+
Systems
Data
Algorithms
Automated
Learning
Information
Programming
Computational
Analysis
Software
Applications
Reasoning
Verification
Logic
Theory
Models
Complexity
Networks
Modelling
Machine
Engineering
Optimisation
Foundations
Embedded
Inference
Refinement
Multicore
Design
Games
Search
Neural
Constraints
Efficient
Checking
Support
Model
Techniques
Ontology
Methods
Spaces
Web
Network
Constraint
Stochastic
Resource
Problems
New
Control
Process
Flow
=
data
protein
gene
sequencing
expression
analysis
sequence
genome
prediction
structure
dna
methods
sequences
network
genomic
nucleotide
interaction
datasets
cancer
new
multiple
model
human
different
single
cell
alignment
next-generation
proteins
biological
genes
mirna
tool
detection
software
methylation
high-throughput
binding
genome-wide
association
algorithm
tools
differential
amino
phylogenetic
experimental
reads
annotation
metabolic
structural
computational
models
genetic
complex
regulatory
acid
variant
interactions
disease
web
networks
information
error
bioinformatics
specific
high
research
ngs
short
cells
sets
relevant
large-scale
novel
variants
similarity
algorithms
technologies
method
target
approach
selection
large
rna
peak
mass
drug
current
studies
standard
structures
functional
available
quality
sites
pipeline
reference
inference
variation
graph

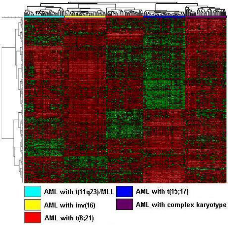
Bioinformatics?

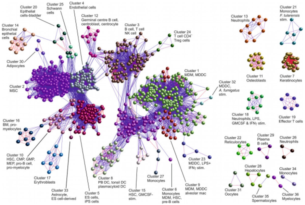


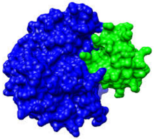

Cluster 20 Epithelial cells-bladder
Cluster 25 Schwann cells
Cluster 4 Endothelial cells
Cluster 12 Germinal centre B cell, centroblast, centrocyte
Cluster 14 Bronchial epithelial cells
Cluster 3 B cell, T cell NK cell
Cluster 24 T cell CD4⁺ Treg cells
Cluster 13 Neutrophils
Cluster 21 Monocytes F. tularensis stim.
Cluster 30 Adipocytes
Cluster 11 Osteoblasts
Cluster 7 Keratinocytes
Cluster 1 MDM, MDDC
Cluster 2 MSC
Cluster 32 MDDC, A. fumigatus stim.
Cluster 18 Neutrophils, LPS, GMCSF & IFNγ stim.
Cluster 19 Effector T cells
Cluster 16 BM, pro-myelocytes
Cluster 22 Reticulocytes
Cluster 29 Plasma B cells
Cluster 26 Neutrophils
Cluster 10 HSC, CMP, GMP, MEP, pro-B cell, pro-myelocyte
Cluster 23 MDDC, LPS+ IFNγ stim.
Cluster 28 Hepatocytes
Cluster 34 Monocytes
Cluster 17 Erythroblasts
Cluster 8 PB DC, tonsil DC plasmacytoid DC
Cluster 27 Monocytes
Cluster 9 MDM, MDDC alveolar mac
Cluster 31 Oocytes
Cluster 33 Astrocyte, ES cell-derived
Cluster 5 ES cells, iPS cells
Cluster 15 HSC, GMCSF-stim.
Cluster 6 Monocytes MDM, HSC, pre-B cells
Cluster 35 Spermatocytes
Cluster 36 Myelocytes

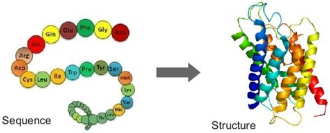


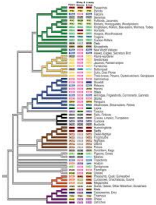
Bioinformatics?

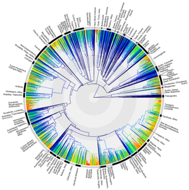












Definitions

- **Oxford Dictionary**: the science of collecting and analyzing complex biological data such as genetic codes.
- **Wikipedia**: an interdisciplinary field that develops methods and software tools for understanding biological data.
- First use was in 1970: “the study of information processes in biotic systems” in a similar way as
  - Biophysics: the study of physical processes in biological systems
  - Biochemistry: the study of chemical processes in biological systems

Historical perspective

A short historical review of biology and computer science

- Before 70's.
  - Biology:
    - DNA structure, 1953.
    - Molecular structural properties, 1953, 1957
    - Metabolic pathways, 1945.
    - Genetic regulation, 1969.
  - Computer Science:
    - Computer and Information Theory, 1966 y 1962.
    - Definition of grammars, Chomsky, 1959.
    - Game Theory, Neumann, 1953.
    - Celular Automata, Neumann, 1966.

A history review of biology and computer science

- 70's, Development of theoretical bases
  - Sequence alignment methods, 1974.
  - RNA structure prediction, 1971.
  - First successful application of ML to phylogenetics, 1974.
  - Secondary proteins structure prediction algorithm, Chou and Fasman, 1974
  - First public repositories of protein sequence (1978) and its structures (1977).

A history review of biology and computer science

- 80's, independent discipline definition
  - Efficient algorithm design to work with large data sets
  - First commercial software developed. First departments dedicated to this field.
  - Main developed areas in bioinformatics:
    - **Sequence analysis**: concept of evolution distance (1980), approximate string matching (1985). Smith-Waterman algorithm (1981), FASTA (1983, 1985)
    - **Molecular databases**: GenBank (1986), EMBL Data Library (1986). First network: EMBNET, BIONET. Search strategies in sequence databases (1985, 1987, 1988).
    - **Protein structure prediction**: representation and visualization of proteins (1980, 1982, 1983, 1984, 1986), visualization software (1985, 1988), structure comparison (1980, 1982, 1989), protein folding discovery strategies (1980, 1985, 1983).
    - **Molecular evolution**: relation between sequence and structure (1986), protein family analysis (1980, 1982, 1984), computation of evolution trees (1981, 1985, 1988), **PHYLIP** (1980)

A history review of biology and computer science

- 90's, technology and availability
  - Internet (ftp, gopher, email, first websites, Mosaic, Netscape)
  - Heterogeneous machines, resource and data distribution "by hand".
  - Introduction of perl v5 (1994) and python v1 (1994)
  - Develop of BLAST (1990), gene prediction algorithms (1990, 1991, 1992).
  - Sequence similarity using high performance systems.
  - First whole chromosome computationally annotated (yeast chromosome III, 1992)
  - Considered by some authors as the born of **"genome informatics era"**.

A history review of biology and computer science

- “new century”
  - High performance technology.
  - Large number of data.
  - Free online data-bases (which makes this a special area).
  - Whole-genome analysis.
  - Applications in several areas of health and others industry.
  - New technologies that allow to measure different gene products
  - Drop in cost of technologies ⇒ new ways of looking at data

Summary

- Combination of two well developed areas
- Availability of data and the lack of well known rules in several areas creates the need to work closely
- New questions have arose because of the new technologies
- Bioinformatics have been approached from several disciplines: mathematics, physics, statistics, computer science, biology, ecology, among others
- Some questions arise:
  - are we doing bioinformatics just by applying a computer program on a given dataset?
  - can biologists survive without the collaboration of other disciplines?
  - can other disciplines make real contributions in biology without really understand what the problems and the data are?

Examples of common mistakes

Sub areas of bioinformatics

- Sequence Analysis
- Genetic annotation
- **Gene expression analysis**
- Gene regulation
- Proteomics
- SNPs
- Genomic comparison
- **Structural bioinformatics**
- Biological system modelling
- Personalized medicine
- among others

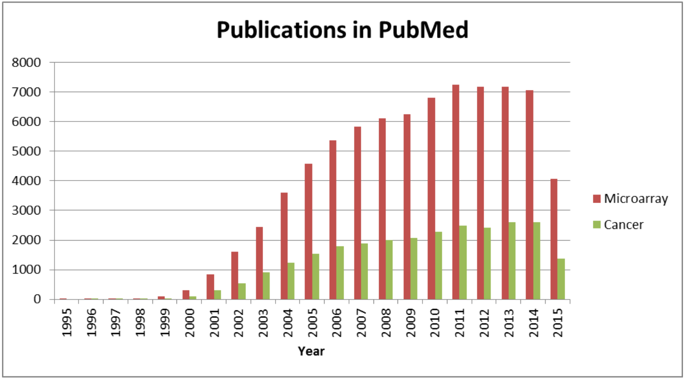
Examples of common mistakes

- In every sub area of bioinformatic it is possible to find examples of common mistakes that happens because of the lack or miscommunication between disciplines
- For example in sequence analysis depending if you are working with eukaryote or prokaryote, there will be presence or absent of the intron-exon gene structure.
- Two cases in detail:
  - Clustering expression data
  - Protein Structure Prediction using metaheuristics algorithms

Clustering gene expression data

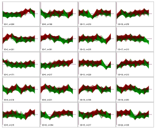
- Gene expression related technologies allows to measure the generation of gene products
- A microarray allows to measure the expression of several probes (genes) in several samples at the same time
- They are normally used to know the behaviour of genes under different conditions: disease/no disease, several subclasses of diseases, treatment/no treatment, etc
- This technology has been widely used in the quest to deal with several diseases, like cancer, neurodegenerative diseases, etc.
- One of the first works reported using microarray chips was in 1995
- Since then, thousands of papers have been written using this technology

Publications using microarray technology

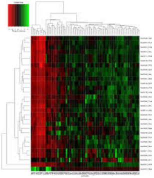


Computational analysis of Microarray data

- There are two main tasks when analysing gene expression data:
    - **Clustering**: find groups of genes/probes that have similar expression profiles



    - **Differential expression analysis**: find a subset of genes/probes that differentiate two or more classes of samples



Computational analysis of Microarray data

- There are several algorithms to performed these tasks:
  - **Clustering**: kMeans, Hierachical clustering, Self Organizing Maps, MSTkNN, model based clustering methods, fuzzy methods, among several others
  - **Differential Expression analysis**: fold change, ANOVA, SAM, combinatorial optimization based models ($\alpha - \beta$ feature set) among others

Computational analysis of Microarray data

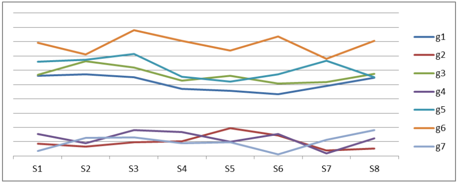
- Apart from the natural choice of the algorithms to use, there are other decisions to make:
    - Data normalization
    - Algorithm parameters
    - Number of clusters (if the method needs it)
    - Missing values
- In particular in clustering, the definition of the distance metric to use is a key decision to make

Simple example of the distance choice effect

- Two of the most common distances metrics used between gene expression profiles are Euclidean distance and Pearson correlation based distance
- Both distance look at very different characteristics to say when two genes are “close”

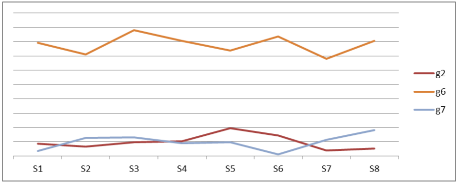
Simple example of the distance choice effect

- Two of the most common distances metrics used between gene expression profiles are Euclidean distance and Pearson correlation based distance
- Both distance look at very different characteristics to say when two genes are “close”



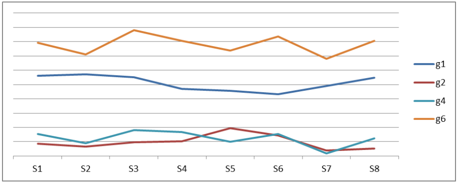
Simple example of the distance choice effect

- Using a euclidean distance:
  - Nearest genes: $d(g2, g7) = 0.01$
  - Most far genes: $d(g6, g7) = 27.35$



Simple example of the distance choice effect

- Using a pearson correlation base distance:
  - Nearest genes: $d(g4, g6) = 0.07$
  - Most far genes: $d(g1, g2) = 1.63$



Simple example of the distance choice effect

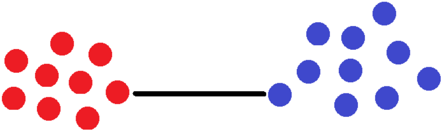
- Lessons:
  - The selection of the distance metric is important
  - It is closely related with the question that we are looking to answer
  - If we aim to find groups of genes that express in similar amounts, we should use Euclidean distance
  - If we aim to find groups of genes that express with similar patterns across the samples, we should use Pearson correlation based distance
  - Other distance metrics or combination of them can be also used

Another example in clustering: HC

- One of the most used clustering algorithms is Hierarchical clustering
- Apart from the distance choice already discussed we need to consider how to compute distances between groups
- At least three choices:
    - Single linkage
    - Complete linkage
    - Average linkage (UPGMA)

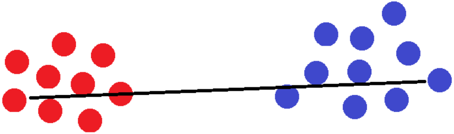
Another example in clustering: HC

- One of the most used clustering algorithms is Hierarchical clustering
- Apart from the distance choice already discussed we need to consider how to compute distances between groups
- At least three choices:
  - Single linkage
  - Complete linkage
  - Average linkage (UPGMA)



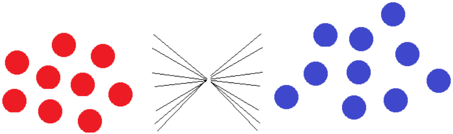
Another example in clustering: HC

- One of the most used clustering algorithms is Hierarchical clustering
- Apart from the distance choice already discussed we need to consider how to compute distances between groups
- At least three choices:
  - Single linkage
  - Complete linkage
  - Average linkage (UPGMA)



Another example in clustering: HC

- One of the most used clustering algorithms is Hierarchical clustering
- Apart from the distance choice already discussed we need to consider how to compute distances between groups
- At least three choices:
    - Single linkage
    - Complete linkage
    - Average linkage (UPGMA)



Another example in clustering: HC

- Lessons:
  - Knowing the parameters of the algorithm can tell us more about the results
  - Taking decision based on real information about the algorithm is more likely to produce better results
  - It allows to know the capacities and shortcomings of the methods
  - It is not necessary to become an expert on the algorithm, but at least to know the main parameters that it has

Protein Structure Prediction

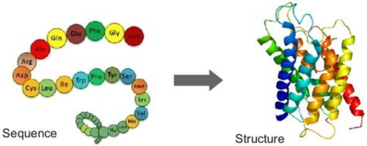
- It corresponds to one of the problems that are found in Structural Bioinformatics
- The goal is to predict the three dimensional shape of a given sequence of aminoacid
- This problem has challenged researchers from different disciplines and there is no single method that can solve it
- Computationally speaking this problem is hard to solve. It has been classified as NP-Complete problem
- This problem is important since the function of the protein is closely related with the three dimensional structure
- Other problems in Structural bioinformatics are: principles of molecular folding, evolution, binding interactions, structure/function relationships.

Protein Structure Prediction

- Computational methods to deal with the 3D-PSP are classified in four groups:
  - First principle methods without database information
  - Fold recognition
  - Comparative modelling
  - First principle methods with database information
- One of the approaches to deal with this problem is to model it as an optimization problem and use metaheuristics to deal with it
- There is a rich knowledge already accumulated in data bases like Protein Data Bank (*http://www.rcsb.org*)
- There is also a competition call CASP (Critical Assessment of protein Structure Prediction)
- At their website (*http://www.predictioncenter.org*) it is possible to find information of the latest advances in computational methods for the PSP problem.

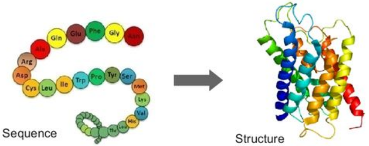
Protein Structure Prediction

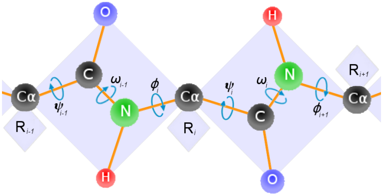
- The problem:



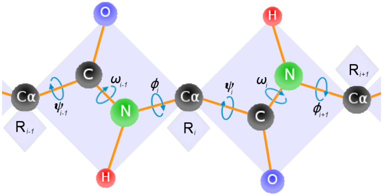
Protein Structure Prediction

- The problem:





Protein Structure Prediction



- We need to determine the values of angles ϕ and ψ
- In theory, these angles can take values in the range of $[-180, 180]$
- With this information, a CS can propose a solution using a metaheuristic like GA (others can be used as well)

Protein Structure Prediction

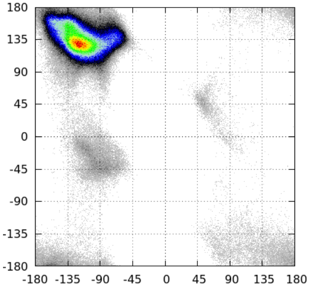
- First it needs to define the solution representation, genetic operators and fitness function, among other things
- The fitness function will guide the GA towards good solutions
- The most common fitness function used is the energy: AMBER (several version), ROSETTA (several version), among others.
- What is the CS thinking: “If I reached small energies I will find better results”
- However, after the first experiments even that the algorithm is reaching good solutions in terms of energy, they are not good in terms of structure

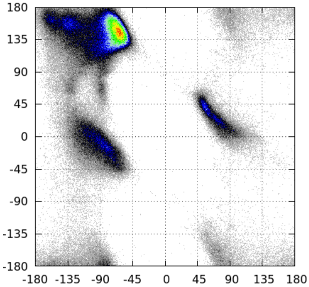
Protein Structure Prediction

- If we take a look at the data stored in the PDB, it is possible to collect some information to help the algorithm (β-*Sheet* and *Coil*)

Protein Structure Prediction

- If we take a look at the data stored in the PDB, it is possible to collect some information to help the algorithm (β-*Sheet* and *Coil*)





- What can be done with this information?

Protein Structure Prediction

- It is possible to reduce the search space by creating knowledge based operators
- An angles probability list (APL) can be created to help the search
- This list can be specialized in several ways so to help even further the algorithm
- But the question naturally rises: does it really matter for the algorithm?

Protein Structure Prediction

- It is possible to reduce the search space by creating knowledge based operators
- An angles probability list (APL) can be created to help the search
- This list can be specialized in several ways so to help even further the algorithm
- But the question naturally rises: does it really matter for the algorithm?

Protein	APL		Without APL	
	Energy	RMSD	Energy	RMSD
2EVQ	-94.2 (-70.2)	3.58 (2.87)	-92.6 (-40.8)	3.76 (2.76)
1K43	-558.6 (-515.2)	2.50 (2.71)	-447.8 (-405.0)	5.05 (4.77)
1DEP	-304.2 (-272.7)	1.43 (1.03)	-377.3 (-239.2)	4.12 (4.28)
1E0Q	-280.9 (-236.7)	7.08 (4.77)	-141.2 (-49.4)	5.04 (5.41)
1RPV	-1027.9 (-937.3)	2.15 (1.88)	-1075.1 (-947.3)	5.66 (5.66)
1L2Y	-261.9 (-225.7)	5.43 (4.04)	-187.4 (-23.8)	5.01 (5.39)

Protein Structure Prediction

- Further improvements can be incorporated to the algorithms
- Knowing the biology behind the problems, it is possible to design more ad-hoc algorithms
- Algorithms and models that work for similar problems in bioinformatics, not always will work well straight away
- Current developments incorporate machine learning techniques to take more advantage of the biological knowledge

Other common mistakes and bad practices

- Stick to what is already in used
- Do not give an opportunity to new ways of looking at data
- Do not understand the meaning of the parameters of the algorithms
- Not taking time to analyse and discuss the partial results of the algorithms
- Not taking time to really understand the biology behind the problems that are being faced
- Thinking that small steps are not real contributions:
  - Some discoveries are not a cure for a certain disease, but they can increase the understanding of the disease, leading for example to more accurate prognoses

Education in Bioinformatics

Education in Bioinformatics

- How can we cooperate to really increase the successful rate of interdisciplinary collaboration in bioinformatics?

Education in Bioinformatics

- How can we cooperate to really increase the successful rate of interdisciplinary collaboration in bioinformatics?
  - Education
  - Acceptance of alternatives
  - Listening to each other (not just hearing)
  - And something that is taught in most of the CS courses: learning the problem that need to be solve

Bioinformatic programs

- Around the world there are several bioinformatic graduate programs (UCLA, Boston University, Georgia Tech, MIT, etc)
- In other disciplines the formation of bioinformaticians comes from taking some courses in their respective programs
- However, what is a bioinformatician?
  - Is you are a biologist and use bioinformatic tools, are you a bioinformatician?
  - If you area an IT engineer and maintain a web server with bioinformatic tools and biological data, are you a bioinformatician?
- The formation in the area must be interdisciplinary.
- Universities created specific programs that teach both worlds but not as different entities
- Much work need to be done, specially since new technology is increasing the gap between data generated and the knowledge that have been extracted.

Bioinformatics roles

- It is clear that depending on the tasks we are performing a different opinion about Bioinformatics will exist.
- According to the work presented by Searls, 2012 ("An Online Bioinformatics Curriculum") it is possible to identify five types of bioinformatic practitioners:

Bioinformatics roles

- It is clear that depending on the tasks we are performing a different opinion about Bioinformatics will exist.
- According to the work presented by Searls, 2012 ("An Online Bioinformatics Curriculum") it is possible to identify five types of bioinformatic practitioners:
  1. Bioinformatics Analysis (BA)
  2. Data Mining (DM)
  3. Bioinformatics Tools (BT)
  4. Bioinformatics Systems (BS)
  5. Computational Biology (CB)
- Which name best describe your practice?

Bioinformatics roles

Bioinformatics Analysis (BA): **Goal:** interpretation or prediction of biological data. It involves **Skills:** sequence, expression, and functional analysis using standard bioinformatics tools, to write computational scripts, database queries, and simple programs.

Bioinformatics roles

Bioinformatics Analysis (BA): **Goal:** interpretation or prediction of biological data. It involves **Skills:** sequence, expression, and functional analysis using standard bioinformatics tools, to write computational scripts, database queries, and simple programs.

Data Mining (DM): **Goal:** enable for more sophisticated analyses of datasets (very large scale, noisy, high-dimensional, semantically rich, poorly organized or integrated, among others) **Skills:** deeper in mathematical knowledge and programming skills.

Bioinformatics roles

Bioinformatics Analysis (BA): **Goal:** interpretation or prediction of biological data. It involves **Skills:** sequence, expression, and functional analysis using standard bioinformatics tools, to write computational scripts, database queries, and simple programs.

Data Mining (DM): **Goal:** enable for more sophisticated analyses of datasets (very large scale, noisy, high-dimensional, semantically rich, poorly organized or integrated, among others) **Skills:** deeper in mathematical knowledge and programming skills.

Bioinformatics Tools (BT): **Goal:** develop standalone tools of significant sophistication for bioinformatics analysis, visualization, presentation, and local data management. **Skills:** programming skills in a variety of languages and the ability to **implement complex algorithms efficiently**, based on solid **biological domain knowledge**.

Bioinformatics roles

Bioinformatics Systems (BS): **Goal:** development and lead of major bioinformatics systems and/or products, for instance supporting data management and analysis from novel technological platforms through complex downstream analysis pipelines. **Skills:** software engineering knowledge

Bioinformatics roles

Bioinformatics Systems (BS): **Goal:** development and lead of major bioinformatics systems and/or products, for instance supporting data management and analysis from novel technological platforms through complex downstream analysis pipelines. **Skills:** software engineering knowledge

Computational Biology (CB): **Goal:** prepare individuals to do original research in biological modelling and analysis by way of advanced mathematical and computational techniques. **Skills:** a deeper grounding in computer science and engineering disciplines relevant to the sciences of complexity, information, and systems.

Final comments

Final comments

- Education
  - Select or design key subjects in a bioinformatics program
  - Early immersion and participation in real research projects
  - Biology subjects taught by biologist who uses bioinformatics tools
  - Computer science subjects taught by computer scientist who have participated in bioinformatic projects
- Collaborative research

Useful Links

- PLOS Computational Biology Collections
  - The Roots of Bioinformatics
  - Education
  - Ten Simple Rules
- Courses and programs in bioinformatics
  - GeorgiaTech Bioinformatics. Interdisciplinary Graduate Programs
  - Boston University Interdisciplinary programs. Bioinformatics
  - Computational and Systems biology at MIT
  - UCLA Bioinformatics Program
  - Biolinux
  - NCBI
- Other
  - DNA seen through the eyes of a coder

Acknowledgments

Asociación de Universidades GRUPO MONTEVIDEO Programa Asociación de Universidades del Grupo de Montevideo

.inf INSTITUTO DE INFORMÁTICA UFRGS Instituto de Informática, UFRGS

EGB Organizing committee of the **Escola Gaucha de Bioinformática**

UdeSantiago

DIINF

Thank you for your attention!