Fundamentals of Bioinformatics: computation, biology, computational biology

Vasilis J. Promponas
Bioinformatics Research Laboratory
Department of Biological Sciences
University of Cyprus
A short self-introduction

• Vasilis, pronounced: “Vass`ilis”
• A Frog Physicist turned into a Biologist
  – Coincidence: it all happened around 1995-96
  – Computational approach
• PhD in Biology (2004, Computational Biology/Bioinformatics), University of Athens, Greece
• 2005 – Moved to Cyprus
Cyprus?
Cyprus?
Cyprus?
Cyprus?
University of Cyprus
"A mathematician is a device for turning coffee into theorems"

Paul Erdos - 20th Century
"A mathematician is a device for turning coffee into theorems."

Paul Erdos - 20th Century
A mathematician is a device for turning coffee into theorems

Paul Erdos - 20th Century

A bioinformatician is a device for turning biological problems into a hot cup of coffee

Anonymous bioinformatician - 21st Century
Overview

• Introduction
  – Some definitions and concepts from (Molecular) Biology
  – The rapid growth of Biological Data

• The advent of the Genome Era (a paradigm shift in Biology?)

• Bioinformatics and Computational Biology: Fundamental Problems – Concepts – Applications

• Discussion
Introduction
Introduction

Bio::revenge

- Biology **IS** the science of the 21st century
  - Used to be a **QUALITATIVE** scientific domain
    - **Exceptions** have been the **Rule**
    - Turning into **QUANTITATIVE**
    - An **Information Rich** field
- Impact in every aspect of (human) lives
  - **Food** production and Quality Control
  - **Environment** (e.g. Ecology, Monitoring, Management)
  - **Human activities/welfare** (e.g. sports, cosmetics, health)
  - ...
Introduction

Key actors

- Genome(s)
- Chromosome(s)
- Gene(s)
- Protein(s)
Introduction

Key actors

- Genome(s)
- Chromosome(s)
- Gene(s)
- Protein(s)
Introduction

Key actors

- Genome(s)
- Chromosome(s)
- Gene(s)
- Protein(s)

But what about some viral genomes?
Introduction

Key actors

- Genome(s)
- Chromosome(s)
- Gene(s)
- Protein(s)

Contains heritable(?) information
Introduction

Key actors

- Genome(s)
- **Chromosome(s)**
- Gene(s)
- Protein(s)

- Tightly packaged (DNA/RNA/proteins)
- 3D-structural organization
- Contains “functional” regions (a.k.a. genes) and regions of (yet) unknown function
Introduction
Key actors

- Genome(s)
- Chromosome(s)
- Gene(s)
- Protein(s)

- A chromosomal region encoding mRNAs, tRNAs, etc.
- Useful keyword: Transcription
- mRNAs: non-terminal
Introduction
Key actors

- Genome(s)
- Chromosome(s)
- Gene(s)
- Protein(s)

- Main components of the (cellular) toolkit
- Linear polymers
- Interact with other biological molecules
- Useful keyword: Translation
Information Flow in Biological Systems
The “Central Dogma”

- Replication
- Transcription
- Translation
Coupled with a “Universal” genetic code

The Genetic Code is Degenerate!

- 1. The Standard Code
- 2. The Vertebrate Mitochondrial Code
- 3. The Yeast Mitochondrial Code
- 5. The Invertebrate Mitochondrial Code
- 6. The Ciliate, Dasyacean and Hexamitida Nuclear Code
- 9. The Echinoderm and Flatworm Mitochondrial Code
- 10. The Euglenid Nuclear Code
- 11. The Bacterial, Archaeal and Plant Plastid Code
- 12. The Alternative Yeast Nuclear Code
- 13. The Ascidian Mitochondrial Code
- 14. The Alternative Flatworm Mitochondrial Code
- 16. Chlorophycean Mitochondrial Code
- 21. Trematode Mitochondrial Code
- 22. Seenedesmus obliquus Mitochondrial Code
- 23. Thraustochytrium Mitochondrial Code
- 24. Porebranchia Mitochondrial Code
- 25. Candidate Division SR1 and Gracilibacteria Code
... some more complexity ...

**Diagram:**
- **Genome**
  - Chromosomes
  - DNA Sequence
  - Amino Acid Sequence
  - Gene (fine) Structure
- ORF1, ORF2, ORF3, ORF4
DNA stores information ...
Proteins get the dirty job done ...
Proteins

- Assembled from one or more polypeptide chains (homo-/hetero-polymers)
- The functional “toolkit”

  - Enzymes
  - Transport-Storage
  - Motion
  - Binding
  - Molecular Recognition
  - Signal Transduction

  - Structural Proteins
  - Energy Production
  - Cell Regulation and Differentiation
  - ... (…)
Yet, some more complexity (PTMs)

Pig Insulin Precursor

MALWTRLPLLALALWAPAPAQAFVNHLCGSHLVEALYLVCGERGFFYTPKARRAEEN
PQAGAVELOGGGGLGQLRALAGPPQKRGIIVEQCCTSICSLYQLENVYCN

Pig Insulin Dimer (PDB_ID: 4INS)

Chain A

Chain B
Back to the “Central doma”

For (almost) all proteins

Sequence

3D-structure

Function

Determines

Again, this “genetic code” is redundant

• Glucose Uptake Pathway
• Glycogen Synthesis
• Formation of triglycerides

..VEQCTSCSLYQL..
But, where is the computation in biology??
Bioinformatics

Biology

Statistics, Mathematics

Physics, Chemistry, Engineering, Linguistics, ...

Informatics, Computer Science
Bio – related fields

- Computational Molecular Biology
- Bioinformatics
- Theoretical Biology
- Biomedical Informatics
- ...

- Where are the limits?
A (fuzzy?) definition of Bioinformatics

• **Bioinformatics** is the “computational **handling** and **processing** of **genetic information**”

Ouzounis & Valencia, 2003
Handling Genetic Information

• Apply existing (or develop custom) efficient methods for
  – Describing and Visualizing
  – Storing
  – Retrieving
  – Integrating

• Large volumes of complex and inhomogeneous data*

*some still call it “Designing and Building Biological Databases”
Handling Genetic Information (part II)

- Particular attention:
  - Origin and Quality of Biological Data
  - Data Annotation [Expert-based, (semi-)automatic]
  - Interconnectivity
  - Friendly to the end-user
Processing Genetic Information

• Analysing biological data
  – **AIM I**: ADDRESSING BIOLOGICAL questions.
    • What makes **Frodo Baggins** (the Hobbit) differ from **Spiderman**? (consider that Spiderman's kitsch costume is not a valid answer)
    • Does molecule A **interact** with molecule B?
    • What is the **3D structure** adopted by X?
    • How does the 3D structure of a molecule specify its function?
  – **AIM II**: ADDRESSING other SCIENTIFIC or TECHNICAL questions
Processing Genetic Information (part II)

- Other questions???
  - Which is the **optimal way to store** genome data in a database?
  - How can I represent sequences belonging in a family with a **statistical model**?
  - How can I obtain the **optimal** pairwise DNA/RNA/Protein **sequence alignment**?
  - Is their any **statistical measure** for indicating the significance of a sequence comparison score?
A parenthesis (...) for solving a common misunderstanding

- Traditional biologists often see Bioinformatics as a “Black box”
  - i.e., predict, then go back in the lab to confirm with experiment …

- However,
  - the computational approach to addressing biological problems is an experimental field on its own
  - a single difference: experiments are performed *in silico*.
And finally ... what do you mean by “Genetic Information”?

- It can be quite generic
  - Nucleotide and amino acid sequences
  - Three dimensional molecular structures (proteins, DNA, RNA, sugars, drugs, ...)
  - Gene expression data
  - Molecular interaction networks
  - Complex biological systems (cells, tissues, organisms, ...)
  - ... even text in the biomedical literature ...
OMICS

- GenOMICS
- TranscriptOMICS
- ProteOMICS
- MetabolOMICS
- KinOMICS
- PhylogenOMICS
- EpitOMICS

- even more ...
  - BibliOMICS
  - DegradOMICS

??? cOMICS ???

Also be aware:

The Tree of Life

A comprehensive list may be found at the URL http://www.genomicglossaries.com/content/omes.asp
Importantly ...

- Freely available data
- Accessible software [free/open software]
thanks  

merci  

ευχαριστώ  

RSV  

20...