Computational bioinformatics

Madalina Sas[1]

[1]Centre for Complexity science, Imperial College London

Talk given at St. Paul's Girls School, London

23.01.2023

Computational Bioinformatics

Methods

Contributions and Applications

A complexity perspective

Methods

Data analysis

Data visualisation

Modelling and Simulation

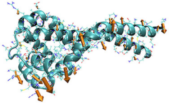
I'm Madalina Sas, a complexity science PhD student at Imperial College London.

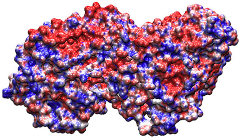
My interest is in **complex collective phenomena**: when many work together to create something greater than themselves. Flocks of birds, schools of fish, humans synchronising behaviour, group improvisation, neurons in brains. Often overlapping with bioinformatics.

I run *experiments* and *simulations*, develop *models* and analysis pipelines, and work with *data*, to investigate the universal features of emergent collective behaviour. You can see some of my work on my website, `https://mis.pm`

Computational Bioinformatics

Computational bioinformatics is an interdisciplinary field of science concerned with developing **methods and software tools** for storage, manipulation and processing of *large* and *complex* biological datasets using computers.

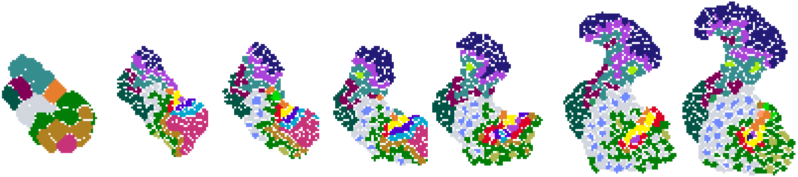




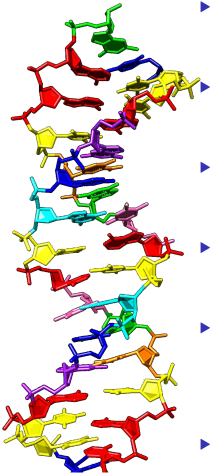
It's goal is to develop tools that help understand biology. The use of such tools to investigate nature is mostly the concern of **computational biology**, and often intersects with **complexity science**.

Methods

- database software for biological data
- analysis methods and algorithms
  - pattern recognition, clustering, prediction, statistics, networks
- visualisation tools
- data mining of biological texts
- modelling biological systems



Contributions and Applications



- gene finding and sequencing: how genes store information
- drug discovery and design: relating cellular chemistry to organism health
- genetics of disease: relating genes to disease, , analysing mutations in cancer
- modelling evolution: from genes to cells to individuals to species
- protein interactions and gene expression: relating chemical properties of proteins to physical properties and information
- learning and information processing in biological systems

A COMPLEXITY PERSPECTIVE

Bioinformatics originally referred to **the study of information processes in biological systems**, as coined by Paulien Hogeweg and Ben Hesper (1970) in 1970.

The discipline has been growing since 1990s due to the Human Genome Project, and became more focused on methods of analysing data.

Paulien Hogeweg still combines modelling with experimental and simulation data to investigate evolutionary processes: for example, simulating how does a unicellular organism become multicellular.

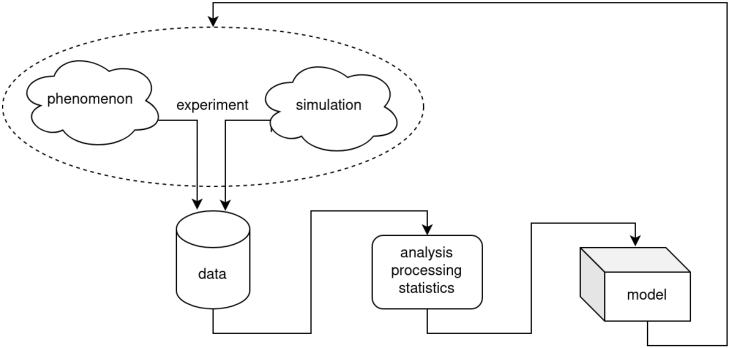
Complexity science is the investigation of systems of many interacting agents who manifest emergent patterns and phenomena at the system level that are not present at the individual level.

"... the whole is something besides the parts ..."

- Aristotle, Metaphysics, Book VII

Methods

To understand a **phenomenon**, we can isolate in it a reproducible manner as an **experiment**. We gather **data** from the experiment, **store** it and **process** it in preparation for **analysis**. Insights from this process can be used to construct **models** of the phenomenon, which can then inform new experiments and **simulations**.



Data analysis

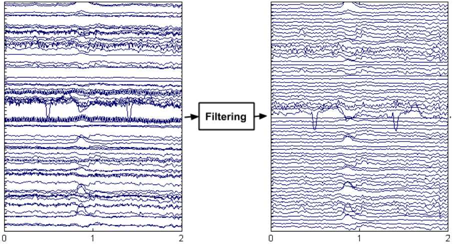
Our data is often collected **raw**, such as photographs, sound recordings, or tissue samples.

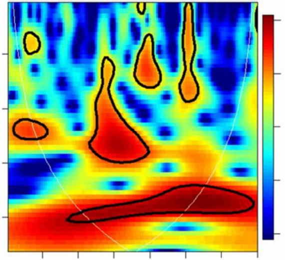
Processing and analysis is required to find *patterns* in this data and extract *information* that we can test our research hypotheses against.

SIGNAL PROCESSING

Using mathematical methods to extract information from time series and images

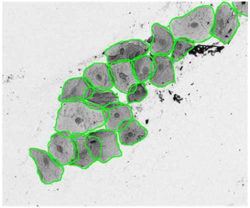
- often involves translation from the time to the frequency domain
- methods for removing noise from data and signal separation

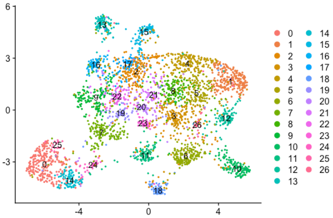


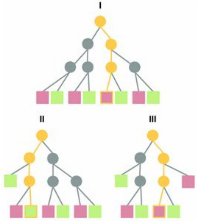


Machine learning

- methods include pattern recognition, feature extraction, clustering, neural network and decision tree classification
- applicable to genomics, drug discovery, systems biology, evolution, text mining

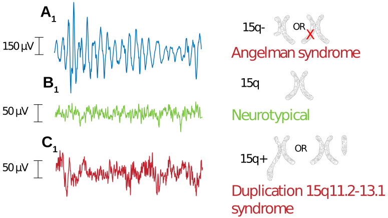


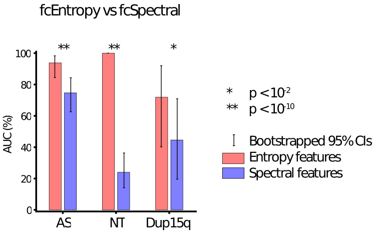




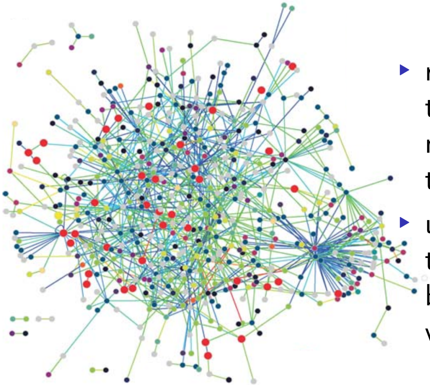
Information theory

- scientific study of the quantification, storage, compression, communication of information
- **Entropy**: the average amount of information conveyed by an event, when considering all possible outcomes.
- such information-based quantities can be used to improve classifiers or assess models





Network science

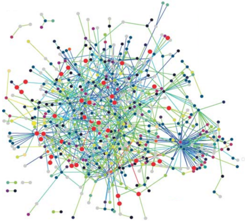


- reasoning about relationships between individuals, such as information transfer, and the evolution of these relationships
- used for studying protein interactions, metabolism, relationships between species in ecosystems, vascular systems, brains

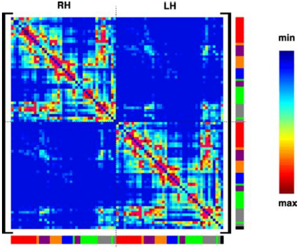
Data visualisation

The same data can be visualised in different ways to reveal different insights about the patterns and relationships existent in it.

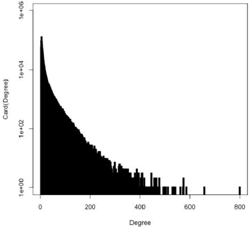
Network



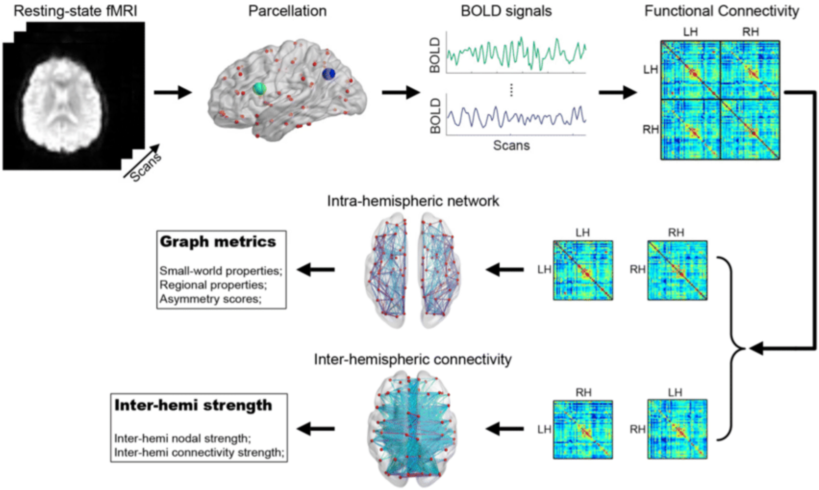
Adjacency Matrix



Degree distribution



Example: neuroscience



Graph metrics

Small-world properties;
Regional properties;
Asymmetry scores;

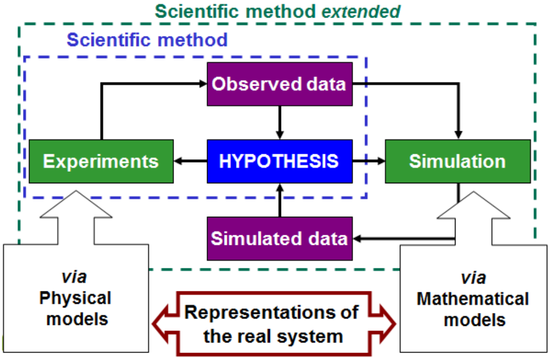
Inter-hemi strength

Inter-hemi nodal strength;
Inter-hemi connectivity strength;

Modelling & Simulation

Model: a system of hypotheses, data, and conclusions presented as a mathematical description of a phenomenon.

Simulation: using a model to create a computer program that iteratively produces data described by the model.


Scientific method *extended*
Scientific method
Observed data
Experiments
HYPOTHESIS
Simulation
Simulated data
via Physical models
Representations of the real system
via Mathematical models