Dreaming about models: a biologist’s perspective

raffaele.calogero@unito.it

Bioinformatics & Genomics unit
Agenda

- About the biology domain
- How a biologist designs an experiment
- What a biologist would like to get from a model
- What a biologist could really provide for model design and development
Agenda

- About the biology domain
- How a biologist designs an experiment
- What a biologist would like to get from a model
- What a biologist could really provide for model design and development
Models in daily life
Weather prediction

Keywords:
Large data input
Data standardization

1. Globally measured data and observations over a particular time period (e.g., 6 h)
2. Geophysical environment of ground, air, and water
3. Collect and process data
4. Analysis and assimilation
5. Mathematical-physical models of numerical weather forecast
6. Statistical interpretation by meteorologist
7. DMO (direct model output)
8. Statistical post-processing (e.g., using KALMAN filters)
9. Weather forecast for users
10. Subjective interpretation by meteorologist
Weather prediction

- Why is prediction successful?
  - We have good numerical models predicting the weather changes.
  - We have a significant set of measurable standardized parameters to feed the models.

- Type of prediction:
  - Weather development
Weather prediction

• Does weather forecast prediction work at any space scale?

• Space scale is important!
Can biological processes be modeled?

- Over the past 10-20 years, biology has become increasingly quantitative, and mathematical sciences have in turn been increasingly influenced by biology.

- However to be able to do really quantitative biology:
  - Questions need to be addressed in the **right biology space**
  - Data need to be provided in a **sufficient amount**
  - Analytical methodologies need to be **standardized**
Biology model

Body space
Biology model

Organ space
Biology model

Cell & tissue spaces
Biology model

- Signal transduction
- Protein-protein interaction
- Epigenetics
- Genome long-range interactions
- Expression regulation
- Post-transcription regulation
- Post-translational modification

Molecular spaces
Biology spaces

Phenotype

body

Microbiota

organs

cells

signal transduction

prot. interactions

prot. modifications

prot. expression

post transcription

transcription

DNA structure

long-range interactions

epigenetics

model development direction
Agenda

- About the biology domain
- **How a biologist designs an experiment**
- What a biologist would like to get from a model
- What a biologist could really provide for model design and development
Biological research strategies

- How a biological problem is defined?
  - A hypothesis is proposed.
  - Experiments are designed to validate hypothesis:
    - Budget phase
    - Technology phase
    - Discovery phase
    - Validation phase
Budget phase

- This issue is part of every biological project in daily life.
- Biological experiments are getting more and more expensive.
- Technology offers constantly new opportunities to have more robust and effective experiments, but not always it is possible to use new methodologies because of their intrinsic cost.
Technology phase

Serial approach
Technology phase

Parallel approach
Technology phase

- Since the serial and parallel approach have similar time frame, why not using always the parallel approach?
Technology phase

- Technologies have limitations!
- Methodological bias

**Correlation between cell type expression profiles**

- **Shared protein coding** (n = 12007)
- **Shared lncRNA** (n = 657)

*Shared lncRNA (n = 657)*

- Technologies have limitations!
- Methodological bias
Technology phase

- Technologies have limitations!
- Lack of standardized methods
- Inconsistency between technologies
Limited data sampling

- 93% called genes
- 850 million reads
- 52% called genes
- ≈ 40 million reads
Limited sample sampling

Genes: > 25 K
Limited sample sampling

SNP: > 2 millions

Patients

< 1000

Patient DNA

Compare differences to discover SNPs associated with diseases

Disease-specific SNPs

Non-patient DNA

Non-disease SNPs

< 1000

Non-patients
Discovery phase

- If top-down approach is used, data can have different in depth depending on the lab producing them.

- **Small biological laboratories:**
  - Data are limited:
    - Time series: 4 points
    - Perturbations: 1 perturbation
    - Number of replicates: 2:3
    - In some cases difficult accessibility to raw data.

- **Public consortiums:**
  - High quality data
  - Large amount of data
  - In some cases difficult accessibility to raw data.
Validation phase

- After having defined some key elements for a specific biological problem various experiments are designed to confirm and motivate the involvement of the key elements in the problem under study.
- Time frame: 2-3 yrs
Agenda

- About the biology domain
- How a biologist designs an experiment
- What a biologist would like to get from a model
- What a biologist could really provide for model design and development
Models in biology

- **What do we want to predict?**
  - Disease appearance
  - Disease development
  - Drug effect

- **What do we want to study?**
  - Effect of perturbations on molecular spaces affecting at least cell space
  - Understanding cell mechanisms:
    - Development
    - Differentiation

- **What do we want to detect?**
  - Key elements involved in disease development
  - Key elements involved in molecular pathways
Biologist’s model idea

- Key element involved in a molecular pathway

Spadaro et al. submitted
Biologist’s model idea

- Models represent a way to define new working hypotheses.

Talattoferrin specific network

Spadaro et al. submitted
Models in biology

- Particularly intriguing are multi-level models:
  - Cell population
  - Molecular networks controlling cell population
Models in biology

Cordero et al. BMC Bioinformatics 2013, 14(Suppl 6):S11
Parameter definition

- The critical issue that require interaction with biologists is parameters definition.

In this case we had a total of 124 reactions.
Agenda

- About the biology domain
- How a biologist designs an experiment
- What a biologist would like to get from a model
- What a biologist could really provide for model design and development
Encode

- 32 Institutes
- 442 Consortium Members
- 1649 Experiments
- 11,972 files
- 15TB of disk space

- 80% of genome participates in at least one biochemical and/or chromatin event in at least 1 cell type
Scientists in the Encyclopedia of DNA Elements Consortium have applied 24 experiment types (across) to more than 150 cell lines (down) to assign functions to as many DNA regions as possible—but the project is still far from complete.

**Experimental Targets**
- **DNA methylation**: regions layered with chemical methyl groups, which regulate gene expression.
- **Open chromatin**: areas in which the DNA and proteins that make up chromatin are accessible to regulatory proteins.
- **RNA binding**: positions where regulatory proteins attach to RNA.
- **RNA sequences**: regions that are transcribed into RNA.
- **ChIP-seq**: technique that reveals where proteins bind to DNA.
- **Modified histones**: histone proteins, which package DNA into chromosomes, modified by chemical marks.
- **Transcription factors**: proteins that bind to DNA and regulate transcription.

**Cell Lines**
- **Tiers 1 and 2**: widely used cell lines that were given priority.
- **Tier 3**: all other cell types.

So far, scientists have examined 13 of about 60 known histone modifications and 120 of about 1,800 transcription factors.

Every shaded box represents at least one genome-wide experiment run on a cell type.

Many more cell types are yet to be interrogated.
Many more cell types are yet to be interrogated.

So far, scientists have examined 13 of about 60 known histone modifications and 120 of about 1,800 transcription factors.
Genomic/transcriptomic data

Model perturbation

New working hypothesis

parameters

model

biologists

interactome

pathways

So far, scientists have examined 13 of about 60 known histone modifications and 120 of about 1,800 transcription factors. Many more cell types are yet to be interrogated.
Conclusions

- Mathematical models offer new perspective to biologists.
- With the present amount of data models can give new incites in providing new working hypotheses.
- Strong interaction between biologists and mathematician is needed to
  - correctly define data for parameters definition
  - highlight methodological limits
Thank you!
raffaele.calogero@unito.it

Matteo Carrara
Chiara Fornari

Immunology Lab

Federica Cavallo
Elena Quaglino
Irene Merighi
Maddalena Arigoni
Stefania Lanzardo
Laura Conti
A Pathway-Based Approach using Latent Variable Structural Equation Modeling: An Application to 1000 Genomes Exon Sequencing Data

by

Nora L. Nock, Ph.D.
**Candidate pathway analysis**

- Preselect pathways of interest
  - Hypothesis driven → targeted biological scope
- Acquire pathway annotations
  - Consider database source, curation and coverage
- Test for pathway associations
  - Enrichment (probably self-contained)
  - Network analysis or clustering
  - Other methods
- Biological interpretation
  - Replication in independent data sets

**Genome-wide pathway analysis**

- Assess all pathways (no preselection)
  - Data driven → extensive biological scope
- Acquire pathway annotations
  - Consider database source, curation and coverage
- Test for pathway associations
  - Enrichment (probably competitive)
  - Network analysis or clustering
  - Other methods

**Methodological issues**

- Input data (genotype, P-value)
  - Imputation
  - Map data elements to genes
  - Score gene significance
  - LD (linkage disequilibrium)
  - Gene and pathway size
  - Pathway overlap
  - Multiple comparisons

**Input: 1 association signal per gene**

**Latent Gene Construct Approach**