Research group on Statistics in the Biosciences

September 12, 2014
Many issues in the health, medical and biological sciences are addressed by collecting and exploring relevant data. The development and application of techniques to better understand such data is the fundamental concern of our group.
People

- G. Adimari (Dept. of Statistical Sciences, Padova)
- M. Chiogna (Dept. of Statistical Sciences, Padova)
- V. Djordjilovic (Dept. of Statistical Sciences, Padova)
- L. Maragoni (Dept. of Statistical Sciences, Padova)
- M.S. Massa (Dept. of Statistics, Oxford, UK)
- D. Risso (Dept of Statistics, University of California, Berkeley)
- C. Romualdi (Dept. of Biology, Padova)
- F. Rotolo (Gustave Roussy, France)
- G. Sales (Dept. of Biology, Padova)

Some people collaborating with us: S. Dudoit (University of California, Berkeley), M. Bottai (Karolinska Institute, Sweden), M. Studený (Academy of Sciences, Czech Republic)
Research themes

- Analysis of genes: microarray data analysis, RNA-seq data analysis, quantile inference
- Evaluation of diagnostic tests
- Survival analysis
With the advent of new technologies (microarray technology, deep sequencing, etc) scientists can now study tens of thousands of genes at once. Unfortunately, they are often daunted and confused by the complexity of data analyses.

Such complexity has provided also statisticians with a wealth of novel problems: in fact, also simple exploratory data analysis can be a challenging task.
Flow of a typical microarray experiment

1. **Biological Question**
2. **Experimental Design** → **Microarray Experiment**
3. **Pre-processing Low-Level Analysis**
   - **Image Quantification** → **Background Adjustment** → **Normalization** → **Summarization**
4. **Expressed Values**
5. **High-level Analysis**
   - **Estimation** → **Testing** → **Clustering** → **...** → **Discrimination**
6. **Differential Expressed Genes**
7. **Biological Verification and Interpretation**
   - **Annotation** → **Enrichment Analysis** → **GSEA** → **...** → **Ingenuity**

Statistics in the Biosciences
Analysis of genes - Here, we work on:

- **pre-processing**: removal of technical artifacts

- **high-level analysis**: modelling of genes expression

- **biological significance analysis**: pathway analysis
Testing differential expression between two groups can translate into testing statistical hypothesis such as:

\[ H_0 : \xi_A(\tau) = \xi_B(\tau) \]

with tools which resemble a Student's t-test.

- Data transformation;
- Testing different aspects of the distribution;
- Multiple testing issues.
Biological Significance Analysis

Statistics in the Biosciences
The evaluation of the ability of a diagnostic or a screening test to separate diseased from non-diseased subjects is a crucial issue in modern medicine. In fact, before a test can be applied in a clinical setting, rigorous statistical assessment of its performance in discriminating the diseased state from the non-diseased state is required.
The ROC curve is the most commonly used statistical tool for describing the discriminatory accuracy of a diagnostic test.

True Positive Rate = \( \Pr(Y > \text{threshold} | \text{diseased}) \)

False Positive Rate = \( \Pr(Y > \text{threshold} | \text{non diseased}) \)
Summary measures

1. Area under the ROC curve (AUC)
2. Partial area under the ROC curve (pAUC)

Ideally, inference on ROC curves and derived measures relies on a random sample from the target population composed on healthy and diseased subjects.
Diagnostic tests - Here, we work on:

- **nonparametric inference**: empirical likelihood methods for making inference on accuracy of diagnostic tests

- **semiparametric inference**
Survival analysis

This is the field of statistics providing tools for handling duration data, i.e., continuous and positive numerical variables measuring the time from an origin event until the occurrence of an event of interest.

(Refer to Prof Pierce and Mazzuco courses)
OPEN PROBLEMS
Data modelling. The best way to model data arising from the most recent technologies is yet to be highlighted.

Inference. We claim that quantile-based inference could be an adequate tool for pursuing inference on such data, as it enjoys some useful features, like making few distributional assumptions and having the potential to investigate differences over the whole distribution of the data (Refer to L. Maragoni final presentation).

Pathway analysis. The point is, given a pathway, how to act on the genes to reach certain goals.

- Given a known intervention, is there evidence that change in the network of the kind expected actually occurred, and, if so, what can be said of the nature and magnitude of the change?
- Which genes are important in mediating association between two endpoints of a path?
Evaluation of diagnostic tests

Often in practice, not all (none of the) subjects undergo the definitive assessment of disease since the verification procedure is expensive, invasive or both. This results in some (all) individuals having missing information on their disease status.

We aim to develop methods to make ROC analysis in presence of verification bias with fully non-parametric strategies.