Statistics for Omics

3 July 2014
Dipartimento di Matematica, Politecnico di Milano
Seminar room, 6th floor

09:00. Francesca Chiaromonte - The Pennsylvania State University
Exploiting structure to reduce and integrate high dimensional, under sampled "omics" data

09:50. Giorgio Melloni - Istituto Italiano di Tecnologia
DOTS-Finder: a comprehensive tool for assessing driver genes in cancer genomes

10:40. *Break*

11:10. Marzia Cremona and Alice Parodi - Politecnico di Milano
Peak shape cluster analysis reveals novel biological insights

12:00. Kateryna Makova - The Pennsylvania State University
Maternal Age Effect and Severe Germline Bottleneck in the Inheritance of Human Mitochondrial DNA

12.50. Marco Masseroli and Stefano Ceri - Politecnico di Milano
Genometric query system as research enabler to discover genome properties

---

Francesca Chiaromonte
Professor of Statistics and Public Health Sciences
Director, Institute for Genome Sciences (Huck)
The Pennsylvania State University

Exploiting structure to reduce and integrate high dimensional, under sampled "omics" data

Abstract:

Dimension reduction techniques especially formulated for regression problems can be very useful in the analysis of high dimensional "omics" data. These techniques produce a small number of composite predictors that can then be used to construct effective models. Compared to other application settings however, high dimensional "omics" data are often characterized by (1) marked heterogeneity and structure, with both samples and features differing in terms of origin and/or information available on their nature or function, and (2) under-sampling, with the number of available samples rather small relative to the number of measured features. This data thus requires a dimension reduction approach that can simultaneously exploit known structure in samples and features and withstand under-sampling. Once formulated, such an approach will allows us to integrate data collected through multiple studies and diverse high-throughput platforms into unified and parsimonious regression models. We are currently in the process of defining theory and methods for this approach, building upon a number of recent developments in the field of Sufficient Dimension Reduction. This talk illustrates some of our preliminary results though analyses of simulated and actual data.

Giorgio Melloni
Istituto Italiano di Tecnologia

DOTS-Finder: a comprehensive tool for assessing driver genes in cancer genomes

Abstract:

A key challenge in the analysis of cancer genomes is the identification of driver genes from the vast number of mutations present in a cohort of patients. DOTS-Finder, is a new tool which allows the detection of driver genes through the sequential application of functional and frequentist approaches, and is specifically tailored to the analysis of few tumor samples. We have identified driver genes in the genomic data of 34 tumor types derived from existing exploratory projects such as The Cancer Genome Atlas and from studies investigating the usefulness of genomic information in the clinical settings.

Marzia Cremona and Alice Parodi
MOX, Dipartimento di Matematica, Politecnico di Milano

Peak shape cluster analysis reveals novel biological insights

Abstract:

In recent years many techniques have been developed to study genetic and epigenetic processes. We focus on a particular Next Generation Sequencing method called ChIP-Seq (Chromatin Immuno Precipitation Sequencing), that permits to investigate protein-DNA interactions. At present, in the relevant literature, the analysis of ChIP-Seq data is mainly restricted to the detection of enriched regions (peaks) in the genome, considering signal intensity. Motivated by the fact that these peaks can show very different shapes, we propose an innovative approach that takes into consideration also the shape of such peaks.

We introduce some indices to summarize the shape and we use multivariate clustering techniques in order to detect statistically significant differ ences in peak shape, with the idea that it can be associated with a functional role and a biological meaning.

Moreover we suggest a functional data analysis approach, considering the proper shape of data for the definition of a suitable metric. This analysis lets us define a new classification to validate the previous results and to introduce a more general and applicable method.

Kateryna Makova
Pentz Professor of Biology
Director, Center for Medical Genomics
The Pennsylvania State University

Maternal Age Effect and Severe Germline Bottleneck in the Inheritance of Human Mitochondrial DNA

Abstract:

The manifestation of mtDNA diseases depends on the frequency of heteroplasmy (the presence of several alleles in an individual), yet its transmission across generations cannot be readily predicted due to the lack of data on the size of mtDNA bottleneck during oogenesis. For deleterious heteroplasmies, a severe bottleneck may abruptly transform a benign (low) frequency in a mother into a disease-causing (high) frequency in her child. Here we present a high-resolution study of heteroplasmy transmission conducted on blood and buccal mtDNA of 39 healthy mother-child pairs

of European ancestry (a total of 156 samples, each sequenced at ~20,000x/site). On average, each individual carried one heteroplasmy, and one in eight individuals carried a disease-causing heteroplasmy, with minor allele frequency ≥1%. We observed frequent drastic heteroplasmy frequency shifts between generations and estimated the size of the bottleneck at only ~29-35 mtDNA molecules. Strikingly, we found a positive association between the number of heteroplasmies in a child and maternal age at fertilization, likely attributable to oocyte aging. Accounting for heteroplasmies, we estimate mtDNA germline mutation rate to be 1.3x10-8 mutations/site/year – lower than in previous pedigree studies but in agreement with phylogenetic studies, thus solving a long-standing controversy and informing the use of mtDNA in dating evolutionary events. This study takes advantage of droplet digital PCR (ddPCR) to validate heteroplasmies and confirms a de novo mutation. These results have profound implications for predicting the transmission of disease-causing mtDNA variants and illuminate mitochondrial genome evolutionary dynamics.

Marco Masseroli and Stefano Ceri
Dipartimento di Elettronica Informazione e Bioingegneria, Politecnico di Milano

Genometric query system as research enabler to discover genome properties

Abstract:

Next Generation Sequencing (NGS) is changing biological research and will change medical practice by providing fast and inexpensive DNA readings of individual genomes. Massive availability of individual genomic and epigenomic data is opening a potential for data sharing, querying and analysis that may soon become the biggest and most important "big data" problem of mankind.

NGS data are up to now managed in physical formats and standards which are influenced by the data production processing of sequencing machines, but do not carry explicit information about high-level properties of the human genome. As a consequence, asking high-level queries on DNA-related information is not adequately supported.

In this exciting framework, we propose a new paradigm for raising the level of abstraction in NGS data management and analysis. We introduce a genomic data model (GDM), which encodes NGS experiment results and available genomic annotations in a format that takes into account the organization of the genome into regions, and a genometric query language (GMQL), which uses such regions as the main data abstractions, and computes high-level operations to extract regions of interest. Their use enables the easy extraction of (epi)genomic features from multiple heterogeneous data and their characterization in different biological conditions, whose analysis can unveil new biological phenomena that cannot be observed at the small experimental scale supported by current bioinformatics languages and standards.