Mathematical Modelling of High Dimensional Biomedical Data.

Magnus Fontes
International Group for Data Analysis
Institut Pasteur
The new centers, PIBnet, and high tech network

Center for Translational Research

Pasteur International Biobank (PIBnet)

Center for Bioinformatics & Biostatistics

Center for Global Health

Center for Technological Innovation (CInTeR)

High-technology nodes

Institut Pasteur Paris
Bioinformatics is needed!

- Expertise in NGS data analysis
- Expertise in high volume data visualization.
- Strong data analysis experience using R/Bioconductor.
- Strong Linux bash scripting experience.
- Strong perl programming experience, and familiar with Python, and JAVA.
- Experience in amplicon based NGS data analysis, and in the analysis of high-throughput sequencing data (preferably Solexa/Illumina and ionTorrent sequence data).
- Familiarity with Galaxy or other pipeline building system.
- Must speak English plus either French or German, any other language would also be an advantage.
- be a team player, and work effectively with the sales and service teams for key customer accounts.
- passion for bio-medical technology and an interest in solving complex challenges using machine learning.
- instil confidence and can build trust with customers.
- set-up and optimise bioinformatics pipelines to meet customers needs.
Bioinformatics is a not well-defined field.

From Wikipedia (2014-11-17): **Bioinformatics** is an interdisciplinary scientific field that develops methods and software tools for storing, retrieving, organizing and analyzing biological data. ...bioinformatics combines computer science, statistics, mathematics and engineering to study biological data and processes.
The IP International Network for Data Analysis-INDA
The Institute Pasteur INDA project

- The Institute Pasteur International Network (IPIN) constitutes a unique asset of biomedical knowledge and unique biomedical data.
- To boost international collaboration and coordination we are currently reinforcing and coordinating data analysis, biomedical informatics and biostatistics within the IPIN.
- We are planning an international PhD school in Bioinformatics as a vehicle for broad collaborations within the network.
- We are establishing high profile collaborative international projects.
- The network is open for international partners.
Sharing data analysis

Cloud for bioinformatics-biostatistics
Project example: Understanding Healthy Human Variation

Google's New Moonshot Project: the Human Body
Baseline Study to Try to Create Picture From the Project's Findings

Google has embarked on what may be its most ambitious and difficult science project ever: a deep dive into the human body. WSJ's Alastair Barr joins Lunch Break with Tara Hessen with the details. 

Human Immunology Project Consortium

About HIPC
The Human Immunology Project Consortium (HIPC) program was established in 2013 by the NIAID Division of Allergy, Immunology, and Transplantation as part of the overall NIAID focus on human immunology. The purpose of HIPC is to capitalize on recent advances in immune profiling methods in order to create a novel public resource that characterizes diverse states of the human immune system following infection, prior to and following vaccination against an infectious disease, or prior to and following treatment with an immunoregulatory agent that targets a known innate immune response(s).
Only around 1% of the genomes in our bodies belong to "our" human genome.
Welcome to the Data Analysis and Coordination Center (DACC) for the National Institutes of Health (NIH) Common Fund supported Human Microbiome Project (HMP). This site is the central repository for all HMP data. The aim of the HMP is to characterize microbial communities found at multiple human body sites and to look for correlations between changes in the microbiome and human health. More information can be found in the menus above and on the NIH Common Fund site.

Areas of Interest

Human Microbial Sampling

16S rRNA and whole metagenome sequencing of samples collected from 300 healthy human participants, to characterize complexity of microbial communities at individual body sites and to provide insights into functions performed by the human microbiome...

DACC Member Organizations

Related Sites

Institut Pasteur
The Healthy Human Global Project: Defining Health to Fight Disease

**THE HEALTHY HUMAN GLOBAL PROJECT**

**$18 MILLION**

**DEFINING HEALTH TO FIGHT DISEASES**

**BRINGING CUTTING-EDGE DIAGNOSTIC TOOLS TO THE WORLD**

Over four billion people in the world do not have access to quality molecular diagnostics, a pre-requisite for the identification of personalized treatment and immunization plans. The expectation is that with the development of new point-of-care technologies, it will be possible to extend diagnostic facilities to all people, including those living in remote regions of resource-limited countries. Realization of this vision requires the establishment of reference values for what constitutes a healthy state of being. As the biological set-points are established by a combination of genetic and environmental determinants, it will be necessary to establish reference values that are guided by ancestry and local habits (e.g., dietary practices, commercial microbial communities).

**THE HEALTHY HUMAN GLOBAL PROJECT RATIONALE**

While the question of human variability continues to be a focal point of scientific research, medical practices and public health policies typically take a "one size fits all" approach to disease management and drug development. Efforts to restore the "personalized" in medical care are current challenges, and it is the driving force of the proposed project. To define the parameters (genetic, epigenetic, or environmental) that constitute a "healthy" immune system, scientists at Institut Pasteur created the Milieu Interieur research project. This project has characterized the immune response variability in 1000 healthy individuals of European descent living in France, and analyzed the observed heterogeneity in terms of their genetic and environmental background. In order to have an in-depth and global understanding of what a healthy immune system is, it is crucial to extend the Milieu Interieur study to other healthy cohorts in the world and to define the impact of diverse ethnic and environmental backgrounds. In this context, the Pasteur Center for Global Health offers a unique opportunity to perform this exceptional study in populations living in low, middle, and high-income countries, spread over three continents (Africa, Asia, Latin America). A cohort of 400 healthy individuals each will be recruited at 5 Institut Pasteur sites, with standardized immune stimulation studies, whole-genome genotyping and metagenome profiling. Additionally, we will collect depth serological information about past exposures and current latent infections, demographies, lifestyle and nutritional data, all serving to enrich the proposed data repositories and help define the impact of environment on a healthy response to infection. With these data, we expect to forge a direct path toward personalized medicine, serve the patient with optimized therapeutic strategies, and clarify the optimal path of vaccine development.

**PROJECT MILESTONES (2015-2020)**

- Recruitment of 5 cohorts of 400 healthy individuals each.
- Establishment of a backlog consisting of DNA, plasma, sample sets, and novel panels.
- Characterization of cellular, protein, and metagenomic phenotypes.
- Association of immune phenotypes with donor genetic diversity.
- Cross comparison of immunological and genetic diversity across different ethnic and geographic sites.
- Establishment of algorithms for personalized reference values defining a "healthy state of being".

**$18 MILLION BUDGET**

The requested support will be utilized to build capacity at the designated Institut Pasteur sites, enabling them to perform standardized use of diagnostic tools, perform phenotype analysis and associate high level computational data mining. Additionally, we will reinforce existing central platforms to generate massively large databases that can be validated locally. The proposed budget breakdown is:

- $11 million will be allocated to human resources (high-level scientists, junior scientists, PhDs).
- $1.5 million is needed to fund the core equipment and computing infrastructure.
- $5.5 million will cover the running costs of the project (immuno-monitoring, genetics, genotyping, collections).

**KEY SCIENTISTS**

Matthew L. Albert, MD-PH.D. is head of the Denovil-Gall Immunology Laboratory, Institut Pasteur. His current positions also include Director of the Department of Immunology, and acting Director of the Center of Translational Research. He is a world leader in Human Immunology and has led efforts in the discovery of biomarkers useful in stratification of patients with chronic infections.

Magnus Forst, MD is a visiting professor at Institut Pasteur in Paris where he is responsible for building and coordinating the International Network in Molecular and Data Analysis. He is a professor of Mathematics at Lund University and the past president of the European Consortium for Mathematics in Industry (ECMI).

Amadou Fontane, MD-PhD is the head of the Emerging Diseases Epidemiology Laboratory at Institut Pasteur. In 2014, he was appointed as Director of the newly created Pasteur Center for Global Health Research and Education. A medical epidemiologist specialized in infectious and tropical diseases with strong public health experience, he leads several international and nationally recognized consortia. He is also involved in teaching as co-Director and founder of the Pasteur-Cam School of Public Health, and coordinator of a master's program in public health with a focus on Infectious Diseases.
Unifying immunology with informatics and multiscale biology

Brian A. Kidd, Lauren A. Peters, Erik E. Schadt & Joel T. Dudley

The immune system is a highly complex and dynamic system. Historically, the most common scientific and clinical practice has been to evaluate its individual components. This kind of approach cannot always expose the interconnecting pathways that control immune-system responses and does not reveal how the immune system works across multiple biological systems and scales. High-throughput technologies can be used to measure thousands of parameters of the immune system at a genome-wide scale. These system-wide surveys yield massive amounts of quantitative data that provide a means to monitor and probe immune-system function. New integrative analyses can help synthesize and transform these data into valuable biological insight. Here we review some of the computational analysis tools for high-dimensional data and how they can be applied to immunology.
Data driven biomedical research has led to a rapid diffusion of informatics and statistics into the analysis of biomedical Big Data. This analysis, however, is still often an *ad hoc* statistical post-analysis.

We need to include mathematical modelling as an already present force in collaborative formulations of the biomedical research questions, with the derived models being used to influence the biomedical workflow.

Biomedical researchers need to assimilate basic mathematical and statistical concepts and reasoning in their mind-set.
Why Most Published Research Findings Are False

John P.A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller, when effect sizes are smaller, when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Factors that influence this problem and some corollaries thereto.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one of a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is R/(R + 1). The probability of a study finding a true relationship reflects the power 1 – β (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2 × 2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2

It can be proven that most claimed research findings are false.

should be interpreted based only on p-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. “Negative” research is also very useful.

Dangers with Exploratory Analysis

WE NEED SOUND MATHEMATICAL AND STATISTICAL MODELLING:

A p-value without explanatory biomedical reasoning for the underlying hypothesis is, ..........just a p-value.
How to Make More Published Research True

John P. A. Ioannidis

The achievements of scientific research are amazing. Science has grown from the occupation of a few dilettanti into a vibrant global industry with more than 15,000,000 people authoring more than 25,000,000 scientific papers in 1996-2011 alone [1]. However, true and readily applicable major discoveries are far fewer. Many new proposed associations and/or effects are false or grossly exaggerated [2,3], and translation of knowledge into useful applications is often slow and potentially inefficient [4]. Given the abundance of data, research on research (i.e., meta-research) can derive empirical estimates of the prevalence of risk factors for high false-positive rates (underpowered studies; small effect sizes; low pre-study odds; flexibility in designs, definitions, outcomes, analyses; biases and conflicts of interest; bandwagon pattern; and lack of collaboration) [3].

Currently, an estimated 85% of research findings are false or exaggerated, and an estimated 85% of research resources are wasted. To make more published research true, practices that have improved credibility and efficiency in specific fields may be transplanted to others which would benefit from them—possibilities include the adoption of large-scale collaborative research; replication culture; registration; sharing; reproducibility practices; better statistical methods; standardization of definitions and analyses; more appropriate (usually more stringent) statistical thresholds; and improvement in study design standards, peer review, reporting and dissemination of research, and training of the scientific workforce.

Selection of interventions to improve research practices requires rigorous examination and experimental testing whenever feasible.

Optimal interventions need to understand and harness the motives of various stakeholders who operate in scientific research and who differ on the extent to which they are interested in promoting publishable, fundable, translatable, or profitable results.

Modifications need to be made in the reward system for science, affecting the exchange rates for currencies (e.g., publications and grants) and purchased academic goods (e.g., promotion and other academic or administrative power).
Classical statistics and the Central Limit Theorem

Classical statistics has mostly been occupied with confirmatory statistical analysis (i.e. hypothesis testing) involving many measurements of a few selected macroscopic phenotypic traits that often can be assumed to be normally distributed or transformed to be normally distributed, e.g. wing length or color.


FlyBase: improvements to the bibliography. Nucleic Acids Res. 41(D1):D751-D757
A classic phenomenon; The Central Limit Theorem

The normal distribution

\[ x \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

**Figure:** Plot of a normal distribution with \( \mu = 12 \) and \( \sigma = 3 \).

Approximating a binomial distribution with \( n = 48 \) and \( p = 1/4 \).
Example: Gene expression data, few measurements (samples) of many variables (genes).

Data from Gene Expression Omnibus ( GEO): GDS 3835
Gene expression of 7365 different genes of 4*12 samples of
Drosophila Melanogaster, Drosophila Sechellia, Drosophila Simulans and the hybrid
Drosophila Sechellia x Simulans
See article: Artieri and Singh, BMC Biology 2010, 8:26
High dimensional datasets. The case of more variables than samples.

$p$ times $N$ data matrices
Measurements of
- $p$ variables
- $N$ samples
- +annotations

Typical case $p \gg N$ and often a need to do integrative analysis.
Curse of Dimensionality:
Sampling and approximation of a function in high dimension scales badly. Multiple Hypotheses Testing.

Blessing of Dimensionality:
Concentration of Measure Phenomena of which the Central Limit Theorem only is one particular manifestation
Measurements of a macroscopical phenotypic trait obey the concentration of measure phenomenon: *If many independent stochastic variables influence the measurement of a phenotypical trait, this trait will be highly concentrated around its median.*

Assuming that the underlying stochastic variables influence the measurements in a linear fashion, the Central Limit Theorem in addition tells us *that the measurements of the trait in addition will be approximately normally distributed.*
Biomarkers are defined through underlying distance concepts

A biomarker indicates similarity or dissimilarity between different conditions. There are always notions of distances involved.

Visualizations connected with Similarities or Distances

Hierarchical Clustering together with a Heatmap

ANOVA Hep B, Hep C, Healthy (p=3e-4 giving 44 analytes)

Similarities between stochastic variables represented as quantiles for top two discriminators
Approximate geodesic distances reveal biologically relevant structures in microarray data

J Nilsson, T Fioretos, M Höglund and M Fontes
Bioinformatics 20 (6), 2004, pg: 874-880

Some recent publications connected with distances and the concentration of measure phenomena

- C. Soneson and M. Fontes *A framework for list representation, enabling list stabilization through incorporation of gene exchangeabilities.* Biostatistics (2012) 13 (1)
- K. Johnsson, C. Soneson and M. Fontes *Low Bias Local Intrinsic Dimension Estimation from Expected Simplex Skewness* Available at early access IEEE Transactions on pattern analysis and machine intelligence