Mathematical Modelling of High Dimensional Biomedical Data.

Magnus Fontes
International Group for Data Analysis
Institut Pasteur and Lund University
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.

Random job description from Naturejobs:
Bioinformatics is a not welldefined. 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.
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.
WE NEED SOUND MATHEMATICAL AND STATISTICAL MODELLING:

A p-value without explanatory biomedical reasoning for the underlying hypothesis is, ............just a p-value.
### Example: Gene expression data, few measurements (samples) of many variables (genes).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.35041</td>
<td>0.02627</td>
<td>0.12345</td>
<td>0.23456</td>
<td>0.34567</td>
<td>0.45678</td>
<td>0.56789</td>
<td>0.67890</td>
<td>0.78901</td>
<td>0.89012</td>
<td>0.90123</td>
<td>1.23456</td>
<td>1.34567</td>
<td>1.45678</td>
<td>1.56789</td>
<td>1.67890</td>
<td>1.78901</td>
<td>1.89012</td>
<td>1.90123</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>-0.65432</td>
<td>-0.76543</td>
<td>-0.87654</td>
<td>-0.98765</td>
<td>-1.09876</td>
<td>-1.10987</td>
<td>-1.21098</td>
<td>-1.32109</td>
<td>-1.43210</td>
<td>-1.54321</td>
<td>-1.65432</td>
<td>-1.76543</td>
<td>-1.87654</td>
<td>-1.98765</td>
<td>-2.09876</td>
<td>-2.20987</td>
<td>-2.32098</td>
<td>-2.43209</td>
<td>-2.54320</td>
<td>-2.65432</td>
</tr>
<tr>
<td>1000</td>
<td>0.34567</td>
<td>0.45678</td>
<td>0.56789</td>
<td>0.67890</td>
<td>0.78901</td>
<td>0.89012</td>
<td>0.90123</td>
<td>1.01234</td>
<td>1.12345</td>
<td>1.23456</td>
<td>1.34567</td>
<td>1.45678</td>
<td>1.56789</td>
<td>1.67890</td>
<td>1.78901</td>
<td>1.89012</td>
<td>1.90123</td>
<td>2.01234</td>
<td>2.12345</td>
<td>2.23456</td>
</tr>
</tbody>
</table>

**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.

\[ p \times N \text{ data matrices} \]
Measurements of
- \( p \) variables
- \( N \) samples
- +annotations

Typical case \( p \gg N \) and often a need to do integrative analysis.
Biomarkers are defined through underlying distance concepts

A biomarker indicate 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

Similarities between stochastic variables represented as quantiles for top two discriminators

ANOVA Hep B, Hep C, Healthy (p=3e-4 giving 44 analytes)
Synergies with projects on 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 inside the human body. With its $500 million investment, Google is hoping to create a detailed picture of what makes up the human body. The goal is not only to understand the body, but also to use that knowledge to improve health care and extend life expectancy.

Human Immunology Project Consortium

About HIPC
The Human Immune Project Consortium (HIPC) was established in 2013 by the National Institute of Allergy and Infectious Diseases (NIAID). The purpose of HIPC is to coordinate and facilitate research on the immune system. HIPC aims to improve our understanding of the immune system and to develop new treatments for immune-related diseases.
Sharing data analysis

Cloud for bioinformatics-biostatistics