ΗΛΕΚΤΡΟΝΙΚΑ «ΕΡΓΑΛΕΙΑ»
ΕΓΚΡΙΤΗΣ ΑΝΑΛΥΣΗΣ ΤΗΣ
ΒΙΒΛΙΟΓΡΑΦΙΑΣ

Καθηγητής Παναγιώτης Βλάμος

Πρόεδρος Τμήματος Πληροφορικής
Διευθυντής Εργαστηρίου Βιοπληροφορικής
& Ανθρώπινης Ηλεκτροφυσιολογίας
Ιονίου Πανεπιστημίου
Contemporary scene

- fundamental shift of biomedical research domain towards integrative and translational methodologies

- needs related to the collection, management, integration, analysis and dissemination of large-scale, heterogeneous biomedical data sets.

- well-established and broadly adopted theoretical and practical frameworks intended to address such needs are still largely developmental
Challenges

- the management of multi-dimensional and heterogeneous data sets
- the size and complexity of data sets that must be managed and analyzed are growing at an extremely rapid rate
- the data management practices currently used in most research settings are both labor intensive and rely upon technologies that have not be designed to handle such multi-dimensional data
- the translational science community demands the creation and delivery of information management platforms capable of adapting to and supporting heterogeneous workflows and data sources
The rise of precision medicine and the availability of high-throughput molecular analyses in the context of clinical care have increased the need for adequate tools for translational researchers to manage and explore these data.
Translation from basic science to human studies

Translation of new data into the clinic and health decision making
Key Definitions

Knowledge Acquisition (KA)
- Quantitative Evaluation
- Qualitative Evaluation
- Targeted Metrics

Verification and Validation

Knowledge Representation (KR)
- Interviews
- SME Analyses
- Computational Tools

System Implementation or Refinement
- Software Engineering Methods
- Knowledge-based Technologies

- Ontologies
- Terminologies
- Logical Models
Translational Research Data is Big Data

Volume, Variety, Velocity
Immediate demands

informatics methods that connect molecular entities to clinical entities

able to integrate large data sets of clinical information with omics data
should provide

(i) the storage and integration of clinical and omics data;

(ii) an analysis framework, enabling scientists to explore their data and generate hypotheses; and

(iii) additional information cross-referenced from external databases
Translational research platforms: Overview

✦ BRISK: Biology-Related Information Storage Kit (2011)
✦ iDASH: Integrating data for analysis, anonymization and sharing (2011)
✦ tranSMART (2010)

Cancer

✦ G-DOC Georgetown Database of Cancer (2012)
✦ iCOD: Integrated Clinical Omics Database (2010)
✦ caTRIP (2006)
Translational research platforms: Overview

- BRISK: Biology-Related Information Storage Kit (2011)
- iDASH: Integrating data for analysis, anonymization and sharing (2011)
- tranSMART (2010)

Cancer

- G-DOC Georgetown Database of Cancer (2012)
- caTRIP (2006)
<table>
<thead>
<tr>
<th>Platform</th>
<th>tranSMART</th>
<th>cBioPortal</th>
<th>BRISK</th>
<th>iDASH</th>
<th>iCOD</th>
<th>G-DOC</th>
<th>caTRIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community</td>
<td>Memorial</td>
<td>Sloan-Kettering Cancer Center, NY, USA</td>
<td>University of British Columbia, Canada</td>
<td>iDASH team, USA</td>
<td>Tokyo Medical University, Japan</td>
<td>Georgetown University, DC, USA</td>
<td>Institution/project initiators</td>
</tr>
<tr>
<td>Institution/project initiators</td>
<td>Initially Johnson &amp; Johnson, USA</td>
<td>Initially Johnson &amp; Johnson, USA</td>
<td>Initially NIH and industry grants</td>
<td>Initially Public and private NIH</td>
<td>Initially Public HSS</td>
<td>Initially NIH</td>
<td></td>
</tr>
<tr>
<td>Software availability</td>
<td>Open source</td>
<td>Open source</td>
<td>Open source</td>
<td>Open source</td>
<td>Not distributed</td>
<td>Not distributed</td>
<td>Not distributed</td>
</tr>
</tbody>
</table>

*Canuel, V., et al, Briefings in Bioinformatics, 2014;16:280–290*
<table>
<thead>
<tr>
<th>Platform</th>
<th>tranSMART</th>
<th>cBioPortal</th>
<th>BRISK</th>
<th>iDASH</th>
<th>iCOD</th>
<th>G-DOC</th>
<th>caTRIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demographics</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Outcomes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Biological results</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Images</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Unknown</td>
</tr>
<tr>
<td>Structured clinical research data</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>‘omics’ data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mRNA expression</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Unknown</td>
</tr>
<tr>
<td>miRNA expression</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Unknown</td>
</tr>
<tr>
<td>SNPs</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Unknown</td>
</tr>
<tr>
<td>Copy number variations</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Unknown</td>
<td>No</td>
<td>Yes</td>
<td>Unknown</td>
</tr>
<tr>
<td>DNA methylation</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Unknown</td>
<td>No</td>
<td>Yes</td>
<td>Unknown</td>
</tr>
<tr>
<td>Platform</td>
<td>tranSMA</td>
<td>cBioPortal</td>
<td>BRISK</td>
<td>iDASH</td>
<td>iCOD</td>
<td>G-DOC</td>
<td>caTRIP</td>
</tr>
<tr>
<td>---------------------</td>
<td>---------</td>
<td>------------</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td>Clinical data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demographics</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Outcomes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Biological results</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Images</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Unknown</td>
</tr>
<tr>
<td>Structured clinical</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>research data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘omics’ data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mRNA expression</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Unknown</td>
</tr>
<tr>
<td>miRNA expression</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Unknown</td>
</tr>
<tr>
<td>SNPs</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Unknown</td>
</tr>
<tr>
<td>Copy number variations</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Unknown</td>
<td>No</td>
<td>Yes</td>
<td>Unknown</td>
</tr>
</tbody>
</table>
Desiderata for translational research platforms

Privacy
Interoperability and standards
Heterogeneity of granularity of the data models
Deployment and maintenance
The ultimate goal of translational medicine is enabling personalized care. Nowadays efforts are made to populate translational research platforms with patient data to fuel discovery. Allowing real-time data-driven decision algorithms to leverage translation research results in the context of clinical care should be a short-term objective.
Description
Translational research system for the caBIG project (cancer Biomedical Informatics Grid™), initiated by the U.S. National Cancer Institute.

Goal
data-integration project with a goal to develop an open-source information network across the United States for secure data exchange on cancer research.

Aims
to create a platform that allows physicians to find patients with similar characteristics, analyze their clinical outcomes and find information about treatments that were administered with success across the caBIG data network.
Oncologists could access a rich live data network that could provide strong statistically significant facts in mere minutes.

This could benefit clinicians to query data from a cohort of preexisting patients to help guide treatment of another patient, rather than relying on single cancer centers or limited facts published in the literature.

caBIG project was cancelled in 2011, few cancer centers were using caBIG's clinical data management tools or its cloud computing infrastructure.

Also caBIG applications cost millions of dollars to build.

The majority of the life sciences research tools developed for caBIG have had limited use or impact, according to the report.
tranSMART (2010) - Foundation members

Gold Members (7)
- ORACLE
- Pfizer
- Roche
- Sanofi
- Takeda
- ConvergeHEALTH by Deloitte
- University of Michigan

Silver Members (15)
- PHILIPS
- Institute for Systems Biology
- assurex health
- DEXSTR
- PerkinElmer
- ONE MIND
- BioSci Consulting
- thehyve
- THOMSON REUTERS
- BT
- Rancho BioSciences
- IO Informatics
- ITTM
- Imperial College London
- Data Science Institute

Affiliate Members (3)
- elevada
- HARVARD Medical School
- idbs
Cohort comparison

- Analysis goals
  - Compare survival and chromosomal abnormalities across disease subtypes
- Available data
  - Clinical data: E.g. staging, age, gender, treatment arm
  - Non-omics data: MSI/MSS (microsatellite instability), mutation data
  - Genomics: Comparative genomic hybridization (arrayCGH)
- Methods
  - Histograms, boxplots, t-test, Kaplan-Meier plot, Cox regression, arrayCGH group test
Results: Comparison of cohorts

Comparison of overall survival in subjects with MSI vs MSS

Comparison of chromosomal alterations between cohorts

Contributed by Gerrit Meijer,
Dutch Translational Research IT (CTMM-TraIT)
Collaboration, Analysis, Customization

- Goals
  - Create central data storage for a preclinical oncology group at a large pharmaceutical company
  - Incorporate a custom R script that processes tumor volume and weight data
- Available data
  - In vivo studies using animal model
  - Animal model metadata
Results: Collaboration, Analysis, Customization

Custom R script incorporated

Animal model metadata stored in tranSMART

Contributed by Tania Khasanova, Rancho BioSciences
Patient Stratification

• Analysis goals
  - Association between demographic characteristics and Bortezomib response?
  - Association between disease subtype and drug response?
  - Any gene expression data predictive of drug response?

• Available data
  - Clinical study results from Bortezomib multiple myeloma study
    - Treated, untreated, responder, nonresponder patients
    - Demographic information
    - Disease subtype
    - Gene expression data

• Methods
  - Chi-squared
  - K-means clustering
  - Kaplan-Meier analysis
Results: Patient Stratification

Age distribution of responders and non-responders is similar.

No significant difference in response rate for different types of disease tested.

Expression of NDUFB1 correlates with OS in intermediate time frames, but not through end of study.

Contributed by Kristen Sweet, Thomson Reuters
Specific objectives were to identify:

- Common biomarker changes across Parkinson and Alzheimer disease
- Common pathway changes across Parkinson and Alzheimer disease
- The normal distribution of imaging and fluid biomarkers across controls
- Novel hypotheses, research findings or conclusions about these neurodegenerative diseases.
What is a “Datathon”?

A datathon is an intense 3-day workshop that challenges researchers to turn data and information into knowledge.

• The datathon format is modeled after hackathons, which are focused on software development. Datathons use research questions and datasets to advance knowledge, not to develop applications.

• At a datathon, participants work in teams to frame research questions, create and implement a research design, mobilize data resources and present their findings in front of a panel of judges.
Open Big Data: Preparing for the Datathon

Alzheimer's Disease Neuroimaging Initiative (ADNI)

Parkinson's Progression Markers Initiative (PPMI)

The Cancer Genome Atlas (TCGA)
Open Big Data: ADNI

The Alzheimer’s Disease Neuroimaging Initiative

(ADNI) unites researchers with study data as they work to define the progression of Alzheimer’s disease. ADNI researchers collect, validate and utilize data. Data including Alzheimer’s disease patients, mild cognitive impairment subjects and elderly controls, are available from this site.
Open Big Data: PPMI

Since 2002, the Michael J. Fox Foundation has been an essential driver of PD biomarker development efforts, funding nearly $90 million in biomarker research... MJFF has enabled researchers to identify several promising biomarker candidates. . . MJFF has invested $60 million to develop a solution that would accelerate verification of PD biomarkers. PPMI is the resulting “collaboration” between researchers, industry, government and study participants.
Open Big Data: TCGA

The Cancer Genome Atlas (TCGA), a collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), aims to generate comprehensive, multi-dimensional maps of the key genomic changes... initiated in 2006... showed that a national network of research and technology teams... could pool the results of their efforts... that making the data freely available would enable researchers make important discoveries.
Open Big Data: The Reality

Alzheimer's Disease Neuroimaging Initiative (ADNI) Data Use Agreement

I will not further disclose these data beyond the uses outlined in this agreement and my data use application and understand that redistribution of data in any manner is Prohibited.

Parkinson's Progression Markers Initiative (PPMI) Data Use Agreement

I will not further disclose these data beyond the uses outlined in this agreement and my data use application.

I will do my best to ensure that Investigators who utilize PPMI data use appropriate administrative, physical and technical safeguards to prevent use or disclosure of the data other than as provided for by this Agreement and will promptly report any use or disclosure of data that does not comply with the guidelines established by this Agreement.

The Cancer Genome Atlas (TCGA) Data Use Agreement

The Requester and Approved Users agree to retain control over the data and further agree not to distribute data obtained through this Data Access Request to any entity or individual not covered in the submitted Data Access.
Key Outcomes:

Very interesting preliminary findings

– 4 SNP’s that predict Parkinson’s progression BETTER than Age

– 3 Biomarkers that cross Alzheimers and Parkinsons

– Sets of biomarkers that predict disease stage


– tranSMART Foundation, MJFF and LONI are working together to enable the distribution and access of ADNI, PPMI, LRRK2 and BioFIND datasets in tranSMART.
Key Outcomes:

• Bringing data scientists, neuroscientists and biostatisticians together to leverage an integrated cross-neurodegenerative disease dataset in tranSMART validated the utility for the platform, and the value of these data when integrated.

• The Datathon produced an innovative approach using machine learning, new biomarker findings, and scientifically thorough pathway analyses.

• It successfully demonstrated that data sharing and the Datathon approach in particular can expedite discovery, as well as offer new fruitful ways to explore datasets. These new approaches are available for the research community to adopt.
Key Outcomes:

- The tranSMART Platform can support many different large and distributed datasets as well as meet varied end-user needs simultaneously.
- The platform's ability to organize these datasets in useful fashions enabled participants to generate new findings and approaches for future research.
- The Platform proved to be a powerful tool in expediting research in a cost efficient way.
- The value of bringing together experts from different disciplines and organizations to work in newly formed teams on integrated datasets deployed in the tranSMART platform was evident.
- The Datathon also provided excellent technical and scientific feedback on how to improve the tranSMART platform for use by the research community in real scientific projects.
G-DOC Georgetown Database of Cancer (2012)

G- Doc Plus uses advanced computational tools to handle a variety of biomedical BIG DATA. It has 3 entry points for the user based on their interest.

1. Precision Medicine workflow
2. Translation Research workflow
3. Population Genetics workflow
Summary of G-DOC Plus data management and analysis features

Data Sources
- EBI Array Express
- NCBI GEO

Supported data types
- Copy number
- Clinical
- WGS
- Cell lines
- Transcriptomics
- microRNA
- Metabolomics

Bioinformatics pipeline
- Molecular data and pre-processing QC
- Clinical mapping & standardization
- Data curation & annotation

Additional links
- NCBI Entrez Gene
- Cytoscape
- GeneCards
- mirBase

Core functionalities
- KM plot
- Pathway enrichment (Reactome)
- Group comparison
- Heatmap viewer
- Explore WGS data
- Principal Component Analysis

Local data center
- Analysis server
- Oracle database

Amazon cloud
- Amazon S3
- EC2 compute instances
- DB on Instance

Collaborators
Researchers
Clinicians
Publications using G-DOC


The long-term vision

• is to extend translational bioinformatics platforms to stay current with emerging omics technologies and analysis methods to continue supporting novel hypothesis generation, analysis and validation for integrative biomedical research.

• integrating several aspects of the disease and exposing various data elements, such as outpatient lab workup, pathology, radiology, current treatments, molecular signatures and expected outcomes over a web interface, these platforms will continue to strengthen precision medicine research.
The long-term vision

• Using former or existing platforms’ best practices
• Designing architectures that can survive in time
• Evaluating low cost solutions
• Learning from huge existing databases in other fields of science
• Thinking about metadata
• Believing in future high performance computing
HelpDesk for existing translational platforms:

- PhD Candidate: Αντιγόνη Αβραμούλη
- PhD Candidate: Σωτήρης Φωτίου
Focusing on one platform risks missing an obvious signal!!!