Data integration and analysis for personalized medicine

Georgios Paliouras
Institute of Informatics and Telecommunications
NCSR “Demokritos”
Many drugs don’t work for everyone

<table>
<thead>
<tr>
<th>Drug Category</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-depressants (SSRIs)</td>
<td>38%</td>
</tr>
<tr>
<td>Asthma Drugs</td>
<td>40%</td>
</tr>
<tr>
<td>Diabetes Drugs</td>
<td>43%</td>
</tr>
<tr>
<td>Arthritis Drugs</td>
<td>50%</td>
</tr>
<tr>
<td>Alzheimer’s Drugs</td>
<td>70%</td>
</tr>
<tr>
<td>Cancer Drugs</td>
<td>75%</td>
</tr>
</tbody>
</table>

Source: Spear et al., Clinical Trends in Molecular Medicine, 2001.
Many drugs don’t work for everyone

Source: E.M. Rogers, dnascribe.com
Personalized/Precision Medicine

What is it?

“Identifying which approaches will be effective for which patients based on genetic, environmental, and lifestyle factors.”

NRC, NIH, US

Why Now?
Advances in biology and bioinformatics that provide understanding of the molecular basis of disease.
Personalized/Precision Medicine

What can it do?

- Optimal therapies for individuals
- Avoid adverse drug reactions
- Reduce treatment costs
- Early detection of disease
- Better prognosis of disease progression
- Facilitate pro-active preventive medicine
Example 1: Cancer

Driver Mutations in non-small-cell lung cancers

Source: Hiley et al., The Lancet, 2016.
Example 2: Rare Disease

Mutations in Duchenne Muscular Dystrophy

Source: cureduchenne.org

Types of Alzheimer’s:
- Early onset (<65)
- Late onset (>65)
- Familial (inherited)

Big Data for Personalized Medicine

- Technology helps understanding disease mechanisms
- Understanding leads to cure
- Discovery depends on the analysis of big data

➡️ Big data analytics is essential for personalized medicine
Biomedical Data Landscape

Patient
- History
- Social
- Feedback/Reports

Open
- Publications
- Databases
- Ontologies

Clinical
- Medical records
- Images
- Lab Tests

Financial
- Claims/Payments
- Prescriptions
- Prices

Research
- Clinical Trials
- Registries
- Omics data
Integration of heterogeneous big data

• Epidemiological data analysis insufficient for personalized decisions
• Need to understand disease mechanisms
• Combine breadth (across a population) with depth (e.g. personal genome) in the analysis
• Big data analysis can address both breadth and depth
The project iASiS

- H2020 Personalized Medicine: Big Data for Public Health Policies
- Start: Apr 1, 2017
- End: Mar 31, 2020
- Budget: € 4.3M
The iASiS approach

Big Data → Analysis → Knowledge Graph → Analysis → Decision

- Image Analysis
- EHR Analysis
- Open Data Analysis
- Genomic Analysis
- Tumor Signature
- Response to Treatment
- Adverse Reactions
Alzheimer's Disease Pilot Data

- EHRs in English
- Brain Images
- Genomic Data

+ Pharmacological knowledge extracted from publicly available datasets
+ Biomedical ontologies and taxonomies
  - terminology standardization
  - semantically describing the EHRs
Example: Survival and progression rates in AD

Which factors are important for progression rate and survival?

**Input**

Patient’s parameters:
- Disease stage (MMSE)
- Cognitive enhancing drugs
- Age, Sex
- Comorbidities
- Lifestyle
- Family history

**Output**

- Stratification of patients
- Survival and progression curves for different factors
- Associations with supporting literature

**Analysis**

- Longitudinal analysis of clinical records of all available patients
- Analysis of open data (e.g. related publications)

**Use models to assist in prognosis**

---

online

offline
Biomedical literature

[Graph showing the number of Medline articles per year from 1950 to 2014, with data source: dan.corlan.net]
Information extraction from publications

- **PubMed Text**: “...histamine H2-receptor blockers, or calcium-channel blockers seems to prevent Alzheimer's disease.”
  - Ontological entity recognition (UMLS):
    - “Calcium Channel Blockers” (CUI C0006684)
    - “Alzheimer's disease” (CUI C0002395)
  - Relation extraction: “Calcium Channel Blockers” PREVENTS “Alzheimer's disease”
Information extraction from publications

- Concept - concept relations: **SemRep** (Rindflesch et al. 2003)
- Article - concept relations
  - Concept occurrence in article text
  - MeSH topics of articles (from manual annotations in PubMed)

<table>
<thead>
<tr>
<th>Disease</th>
<th>Articles</th>
<th>Concepts</th>
<th>Relationships</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>108,458</td>
<td>75,985</td>
<td>11,631,891</td>
</tr>
<tr>
<td>LC</td>
<td>141,712</td>
<td>92,846</td>
<td>14,778,391</td>
</tr>
</tbody>
</table>
Refining publication semantic indexing

• **Fine-grained semantic indexing** of biomedical literature
  - Some MeSH topics are coarse-grained corresponding to multiple concepts
  - E.g. MeSH “Alzheimer Disease” (AD) maps to a set of distinct concepts including: Early Onset AD, Late Onset AD, Familial AD and Presenile Dementia.

![Diagram showing dataset development, model development, classifier model, weakly supervised training dataset, and logistic regression model with macro F1 scores](image)
Enrichment with open databases

- "Dexverapamil" (CUI C1098510) is a "Calcium Channel Blockers"

Diagram:
- **Calcium Channel Blockers**
- **Dexverapamil**
- **Alzheimer's disease**

- Dexverapamil IS A Calcium Channel Blockers
- Calcium Channel Blockers PREVENTS Alzheimer's disease
- Dexverapamil MAY PREVENT Alzheimer's disease
Enrichment with open databases

- **Semantic indexing** of entities from open resources
  - Mapping entities to UMLS concepts: UMLS Terminology Server (UTS)
  - e.g. MeSH “Apolipoprotein E4” (D053327) -> UMLS “Apolipoprotein E4” (C0052201)

<table>
<thead>
<tr>
<th>Resource</th>
<th>Concepts</th>
<th>Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease Ontology</td>
<td>5,307</td>
<td>5,129</td>
</tr>
<tr>
<td>Gene Ontology</td>
<td>64,751</td>
<td>125,629</td>
</tr>
<tr>
<td>MeSH Thesaurus</td>
<td>55,400</td>
<td>123,287</td>
</tr>
<tr>
<td>DrugBank</td>
<td>581,055</td>
<td>1,628,077</td>
</tr>
</tbody>
</table>
Semantic indexing of open data

Calcium Channel Blockers IS A Dexverapamil

PREVENTS

Alzheimer's disease

MAY PREVENT

DISEASE ONTOLOGY
Omics data analysis

Liquid biopsy: circulating tumor DNA

Identification of biomarkers

Correlation with treatment outcome

Source: Joehanes et al., Genome Biology, 2017.

Identification of potential biomarkers

Trans-eQTLs: *trans*-acting expression quantitative trait loci

Observable level:
- Genomic variants (SNPs)

Mechanistic level:
- RBP gene
- RNA-binding protein
- Target mRNAs +/-
- Downstream effects

Source: Joehanes et al., Genome Biology, 2017.
Omics data analysis

- **trans-eQTLs** likely indicating protein–RNA interactions (**omixCore** interaction predictions*)

<table>
<thead>
<tr>
<th>Upstream gene (trans-eQTL locus)</th>
<th>Downstream gene (differentially expressed transcript)</th>
<th>Downstream transcript Ensembl identifier</th>
<th>Transcript Ensembl Biotype</th>
<th>omixCore protein-RNA interaction prediction score</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBRG4</td>
<td>CACNA1C</td>
<td>ENST00000406454.7</td>
<td>protein_coding</td>
<td>0.965</td>
</tr>
<tr>
<td>BAZ2B</td>
<td>MACF1</td>
<td>ENST00000361689.6</td>
<td>protein_coding</td>
<td>0.953</td>
</tr>
<tr>
<td>TBRG4</td>
<td>CACNA1C</td>
<td>ENST0000399634.5</td>
<td>protein_coding</td>
<td>0.941</td>
</tr>
<tr>
<td>RBM6</td>
<td>SNX8</td>
<td>ENST0000222990.7</td>
<td>protein_coding</td>
<td>0.935</td>
</tr>
<tr>
<td>KHNYN</td>
<td>KLC1</td>
<td>ENST0000348520.10</td>
<td>protein_coding</td>
<td>0.904</td>
</tr>
<tr>
<td>RBM5</td>
<td>SNX8</td>
<td>ENST0000222990.7</td>
<td>protein_coding</td>
<td>0.898</td>
</tr>
<tr>
<td>BAZ2B</td>
<td>MGAM</td>
<td>ENST0000475668.6</td>
<td>protein_coding</td>
<td>0.845</td>
</tr>
<tr>
<td>BAZ2B</td>
<td>CNNM2</td>
<td>ENST0000369878.8</td>
<td>protein_coding</td>
<td>0.793</td>
</tr>
<tr>
<td>DDX17</td>
<td>CRTC1</td>
<td>ENST0000338797.10</td>
<td>protein_coding</td>
<td>0.79</td>
</tr>
<tr>
<td>TBRG4</td>
<td>CACNA1C</td>
<td>ENST0000399617.5</td>
<td>protein_coding</td>
<td>0.764</td>
</tr>
<tr>
<td>BAZ2B</td>
<td>NCDN</td>
<td>ENST0000373253.7</td>
<td>protein_coding</td>
<td>0.762</td>
</tr>
<tr>
<td>MRPS18B</td>
<td>C4A</td>
<td>ENST0000428956.6</td>
<td>protein_coding</td>
<td>0.732</td>
</tr>
<tr>
<td>KIAA0391</td>
<td>PSD</td>
<td>ENST0000020673.5</td>
<td>protein_coding</td>
<td>0.724</td>
</tr>
<tr>
<td>DDX54</td>
<td>LINC00963</td>
<td>ENST0000419300.3</td>
<td>lincRNA</td>
<td>0.698</td>
</tr>
<tr>
<td>PSMA6</td>
<td>PSD</td>
<td>ENST0000020673.5</td>
<td>protein_coding</td>
<td>0.691</td>
</tr>
<tr>
<td>VARS2</td>
<td>C4A</td>
<td>ENST0000428956.6</td>
<td>protein_coding</td>
<td>0.68</td>
</tr>
</tbody>
</table>

*Armaos et al. Bioinformatics 2017
Lang et al. NAR 2018
Electronic health records: lung cancer

+ 11,000 New Cancer Patients since 2009
+ 180,000 Consultations
+ 3,600,000 values
+ 77,000 Ambulatory Treatments
+ 3,600,000 values
+ 9,000 Inpatient admissions
+ 660,000 values
+ 9.200 ER episodes
+ 250,000 values
+ 3,400,000 Clinical notes
+ 13,800,000 values
+ Free Text
+ 600,000 Reports
+ 2,400,000 values
+ Free Text
+ 500,000 Forms
+ 9,600,000 values
+ Free Text
+ 1,500,000 Test requests
+ 6,000,000 values
+ 18,000,000 Lab values
+ 500,000 Forms
+ Free Text
+ 1,500,000 Test requests
+ 6,000,000 values
+ 18,000,000 Lab values
Longitudinal database for patients
Electronic health records: lung cancer

- 750 patients:
  - 199,953 clinical records
- Improved NLP:
  - Event detection, UMLS, Drugs, Negation detection
- Analysis of EGFR mutated:
  - 105 patients
Image analysis: Alzheimer’s Disease

White/Gray/CSF tissue probability volumes

Feature extraction
- Supervoxel segmentation
- Feature calculation at supervoxel level
- PCA dimensionality reduction

SVM training

Cognitive Normal (CN) vs. Alzheimer’s Disease (AD)
Image analysis: Alzheimer’s Disease

OASIS-1 dataset: tissue segmentation probabilities

<table>
<thead>
<tr>
<th>Tissue Type</th>
<th>SVM – linear kernel</th>
<th>SVM – polynomial kernel</th>
<th>SVM – RBF kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Accuracy</td>
<td>F-score</td>
<td></td>
</tr>
<tr>
<td>Gray matter</td>
<td>0.6918</td>
<td>0.7214</td>
<td></td>
</tr>
<tr>
<td>White matter</td>
<td>0.5995</td>
<td>0.6266</td>
<td></td>
</tr>
<tr>
<td>Cerebrospinal fluid</td>
<td>0.6503</td>
<td>0.6874</td>
<td></td>
</tr>
<tr>
<td>Aggregate</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Linguistic Risk Assessment: Alzheimer’s Disease

Patients text analysis: CogAware

- Automatic Risk Assessment of Alzheimer’s Patients
- Detection of early language characteristics of AD that are either cross-linguistic or language-specific
- Samples obtained by employing the Cookie Theft Description Task

1 part of the Boston Diagnostic Aphasia Examination and CAMCOG)
Cross Linguistic Analysis

• **Aim of the study:** Identify AD-induced language characteristics (cross-linguistic or language-specific)

<table>
<thead>
<tr>
<th>Language Data (Speech Transcripts AD:NC)</th>
<th>GREEK</th>
<th>DEMENTIA</th>
<th>OPTIMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>GREEK</td>
<td>100 (50:50)</td>
<td>555 (309:246)</td>
<td>428 (180:248)</td>
</tr>
</tbody>
</table>
Classification Results

Accuracy

<table>
<thead>
<tr>
<th>Category</th>
<th>Crosslingual</th>
<th>Extended Crosslingual</th>
<th>Best Model (BoW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demetia</td>
<td>0.57</td>
<td>0.71</td>
<td>0.79</td>
</tr>
<tr>
<td>Optima</td>
<td>0.73</td>
<td>0.72</td>
<td>0.92</td>
</tr>
<tr>
<td>Greek</td>
<td>0.6</td>
<td>0.74</td>
<td>0.77</td>
</tr>
</tbody>
</table>
The iASiS approach

Big Data → Image Analysis → Knowledge Graph → Tumor Signature

EHR Analysis → Response to Treatment

Open Data Analysis → Adverse Reactions

Genomic Analysis → Decision
Classes include: Drugs, Publications, Mutations, Biomarkers, Genes, Side Effects, Proteins, Enzymes, Transporters
The iASiS Knowledge graph

Knowledge Graph describes and integrates:

- Genes and mutations from COSMIC
- Publications (PubMed) and annotations
- Protein and RNA Interactions
- Drugs from DrugBank
- Drug and Target interactions from STITCH
- Drug and Side Effects from SIDER
- Drug-Drug Interactions
- Clinical data and images (lung cancer patients)

Knowledge Graph is linked to:

- DBpedia, Bio2RDF, DrugBank
High-level analysis

Discovery of Drug-Target Interactions

Discovery of Drug Side-Effects

Similar Drugs

Discovery of Drug-Drug Interactions

Similar Drugs

Similar Side Effects

Similar Targets

Cluster #1

Discovery of communities of Similar Patients

Patient 1

Patient 6

Patient 4

Patient 2

Patient 7

Patient 5

Patient 8

Patient 9

Patient 10

Patient 3
Other Uses of Big Data in Medicine

12 Pilots
- Population Health & Chronic Disease Management
- Oncology: Prostate, Lung, Breast
- Industrializing Healthcare Services

- Comorbidities
- Kidney Disease
- Diabetes
- COPD/Asthma
- Heart Failure
- Hyper-Acute Workflows: Sepsis, Stroke
- Asset Management
- Radiology Workflows
Even more data ...

Patient
- History
- Social
- Feedback/Reports

Clinical
- Medical records
- Images
- Lab Tests

Financial
- Claims/Payments
- Prescriptions
- Prices

Open
- Publications
- Databases
- Ontologies

Research
- Clinical Trials
- Registries
- Omics data
Ethical and safe use of data

- Data Ownership
- Explicit Consent
- Data Access & Portability
- Right to be Forgotten
- Intended Use
- Data Minimisation
- Secure Storage
- Data Accuracy
Food for thought

• Technology helps understanding disease mechanisms
• Understanding leads to cure
• Discovery depends on the analysis of big data

• Not sharing or not using data costs lives
• 80% of rare diseases are genetic. Data sharing is imperative.
• Beyond genetics: environmental & social factors for precision medicine
• Modeling disease over time: personalized trajectories
• Beyond drug discovery: data analytics for personalized prevention