Semantic resources for biomedical data integration: phenotypes
Presentation

Dietrich Rebholz-Schuhmann
University of Zurich

Faculdade de Medicina
da Universidade do Porto
# Outline

1. Motivation
2. Phenotypes
3. Analysis
4. The future
Needs for phenotype data

A phenotype is . . . :

- Characterization of model organisms: to measure phenotype modifications in genetic experiments
- Judgement of medical patients: to assess the medical status of the patient
- Translational medicine: to judge a patient (or a drug) against the findings from biomedical research
- Biomedical data integration: to exchange data from the bench to the bedside, core research vs. lab results vs. anamnesis of a patient
What is a phenotype?

A phenotype is . . . :

- an observable trait concerning the representation of a body or organism (by measurement, by human judgment)
- induced by the genotype and the environment (e.g. increased size of muscle)
- giving clues to the understanding of the functioning of
  - the genotype or
  - the treatment,

if we resolve the dependencies

- may look quite different (human vs. mouse) although the genetic background is very similar (about 96%).
Motivation Phenotypes Analysis The future

Phenotype ‘aging’?

Aging is . . .:

- (Wikipedia) . . . the accumulation of changes in a person over time
- Not a disease, but changes induced by diseases.
- Not a process, but changes could be induced by any processes.
- Changes are neither positive nor negative.
- It is a status as the result of changes.
- Is it a phenotype?
  The signs and symptoms of aging are phenotypes, i.e. changes to the expression of the functioning of the body.
Types of phenotypes

Typical examples of phenotypes – in ontologies

- Increased / decreased size of an organ, cells, tissues, e.g. megalosplenie.
- Increase / decrease of functional physiological and metabolic processes, e.g. increase or decrease of the heart rate, the blood pressure, the blood glucose tolerance.
- Existing / Missing structural components, e.g. abnormal connections between vessels in the lungs or in the brain.
- Molecular abnormalities, e.g. missing enzymatic activities in genetic metabolic disorders.
Aging related phenotypes

- Reduced physiological functions:
  e.g. tissue perfusion, tissue regeneration, lower O2 perfusion of the brain

- Reduced metabolic functions:
  e.g. prediabetic conditions, increased blood sugar due to reduced insulin activity

- Changes to the body structure:
  e.g. lost elements (hair, teeth, kidney), broken bones
Sources of phenotypes

Description

Angelman syndrome is a neurodevelopmental disorder characterized by mental retardation, balance disorder, typical abnormal behaviors, and severe limitations in speech. It is caused by absence of a maternal contribution to the imprinted region on chromosome 15. Willi syndrome (PWS; 176270) is a clinically distinct disorder resulting from a duplication in the 15q11-q13 region. In addition, the chromosome 15q11-q13 duplication overlapping clinical features.

Phenotypes as free form text from the patient records: requires feature extraction, morphological / syntactical normalization, mapping to semantic resources.
Sources of phenotypes

Phenotypes from public terminologies (Snomed): mainly collection of terms, limited structure for semantic support.
Sources of phenotypes

Phenotypes as a tree structure:
complex, well represented, difficult to build
Comprehensiveness impaired due to time-consuming development.
Phenotypes of different kinds

1) narrative description

2) structured vocabularies

3) ontological resources: pre-composition/post-composition

Phenotypes as (1) free-form text (patient recorded), (2) predefined terminology, and (3) formal ontology.
Sources of phenotypes

Example phenotypes

- **HPO** hearing loss
- **MP** hearing loss
- **HPO** Progressive childhood hearing loss
  - **MP** abnormal hearing physiology – abnormal ear physiology – hearing/vestibular/ear phenotype
- **HPO** Optic nerve hypoplasia
  - **MP** abnormal optic nerve morphology

Exploitation of phenotypes in research

- Comparison of model organisms with humans
  ⇒ prediction/analysis of phenotypes induced by genes
- Influence of SNPs and genes on phenotypes
  ⇒ prediction of adverse side effects of drugs
- Analysis of complex phenotypes
  ⇒ modelling and analysis of genuine disease effects vs. treatment side effects
- Altogether: better interpretation of the patients condition
Analysis and normalisation of the patient record (similar to other texts)

- The patient record contains the information on the patient.
- Terminologies are available to normalize the record, but the level of normalisation is not sufficient.
- With text mining methods and statistical methods we can normalise the content.
- Using standardized interfaces adds to the degree of normalisation.
- We can compare and the patient record against other data resources, e.g. the scientific literature.
From the patient record to the patient profile

- Analysis of the patient record
- Generation of the profile: use terminology, statistical distribution, vector representation
- Normalisation of the profile using concept URIs instead of terms

URI = Uniform Resource Identifier
Sumo is a competitive full-contact sport where a wrestler (rikishi) attempts to force another wrestler out of a circular ring or to touch the ground with anything other than the soles of the feet. [...] The winner of a sumo bout is either: the first wrestler to force his opponent to step out of the ring, or the first wrestler to force his opponent to touch the ground with any part of his body other than the bottom of his feet. [...] Matches often last only a few seconds, as usually one wrestler is quickly ousted from the circle or thrown to the ground.
**Relevante Termini**

*Sumo* is a competitive full-contact sport where a wrestler (rikishi) attempts to force another wrestler out of a circular ring or to touch the ground with anything other than the soles of the feet. [...] The winner of a *sumo* bout is either: the first wrestler to force his opponent to step out of the ring, or the first wrestler to force his opponent to touch the ground with any part of his body other than the bottom of his feet. [...] Matches often last only a few seconds, as usually one wrestler is quickly ousted from the circle or thrown to the ground.
Sumo is a competitive full-contact sport where a wrestler (rikishi) attempts to force another wrestler out of a circular ring or to touch the ground with anything other than the soles of the feet. [...] The winner of a sumo bout is either: the first wrestler to force his opponent to step out of the ring, or the first wrestler to force his opponent to touch the ground with any part of his body other than the bottom of his feet. [...] Matches often last only a few seconds, as usually one wrestler is quickly ousted from the circle or thrown to the ground.
Relevante Termini

Sumo is a competitive full-contact sport where a wrestler (rikishi) attempts to force another wrestler out of a circular ring or to touch the ground with anything other than the soles of the feet. [...] The winner of a sumo bout is either: the first wrestler to force his opponent to step out of the ring, or the first wrestler to force his opponent to touch the ground with any part of his body other than the bottom of his feet. [...] Matches often last only a few seconds, as usually one wrestler is quickly ousted from the circle or thrown to the ground.
Sumo is a competitive full-contact sport where a wrestler (rikishi) attempts to force another wrestler out of a circular ring or to touch the ground with anything other than the soles of the feet. [...] The winner of a sumo bout is either: the first wrestler to force his opponent to step out of the ring, or the first wrestler to force his opponent to touch the ground with any part of his body other than the bottom of his feet. [...] Matches often last only a few seconds, as usually one wrestler is quickly ousted from the circle or thrown to the ground.
Relevante Termini

Sumo is a competitive full-contact sport where a wrestler (rikishi) attempts to force another wrestler out of a circular ring or to touch the ground with anything other than the soles of the feet. [...] The winner of a sumo bout is either: the first wrestler to force his opponent to step out of the ring, or the first wrestler to force his opponent to touch the ground with any part of his body other than the bottom of his feet. [...] Matches often last only a few seconds, as usually one wrestler is quickly ousted from the circle or thrown to the ground.
Sumo is a competitive full-contact sport where a wrestler (rikishi) attempts to force another wrestler out of a circular ring or to touch the ground with anything other than the soles of the feet. [...] The winner of a sumo bout is either: the first wrestler to force his opponent to step out of the ring, or the first wrestler to force his opponent to touch the ground with any part of his body other than the bottom of his feet. [...] Matches often last only a few seconds, as usually one wrestler is quickly ousted from the circle or thrown to the ground.
Statistical Analysis

<table>
<thead>
<tr>
<th>Number</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>wrestler</td>
</tr>
<tr>
<td>3</td>
<td>force, ground</td>
</tr>
<tr>
<td>2</td>
<td>feet, first, opponent, ring, sumo, touch,</td>
</tr>
<tr>
<td>1</td>
<td>competitive, full-contact, sport, rikishi, attempts, ...</td>
</tr>
</tbody>
</table>
Example for Normalisation: SUMO

- SUMO is a **martial art**.
- SUMO (Suggested Upper Merged Ontology) is an **ontology**.
- SUMO (Small Ubiquitin-like Modifier) is a **protein**.
- SUMO (Surgery and Molecular Oncology) is a **department** of the Dundee University.
- But also:
  - Software Update Monitor, Mozilla Support
  - Sufficiently Uniform Memory Organization
  - Spacecraft for the Unmanned Modification of Orbits
  - Stanford University Mathematical Organization
Example from the literature

A number of data resources are available to do the normalisation: UMLS, ICD, Snomed-CT, UniProt, ChEBI, ...
Infrastructure for textual analysis
Semantic categorization of gene/protein naming conventions

Identified from the annotation guidelines for genes/proteins . . .
A gene/protein may represent

- a recognized macromolecule (e.g., “hemoglobin”, “prolactin”)
- the function of the gene or protein (e.g., “methyl-transferase”)
- (part of) a structure of a protein (e.g., “Cytochrome c oxidase subunit 2”)
- (part of) a process (e.g., “Mitochondrial fission process protein 1”)
- an action on a target (e.g., “DNA gyrase inhibitor”)
- a phenotype (e.g., “protein hunchback”)
- a chemical or physical properties (e.g., “37.8 kD protein”)
- a gene/protein that is similar to another known gene/protein (e.g., “Myc homolog protein”)
Remaining problems in the semantic normalisation

- *mice* versus *MicE protein*
- *left breast cancer* is not a concept:
  - *breast cancer* and *left breast* do exist as concepts
  - Now, is *left breast cancer* different from *right breast cancer*
- *Retinoblastoma*: a disease and a gene, which initiates the cancer type.
- *Streptococcus pneumoniae* (is a species) but not *Streptococcus pneumonia* (is a disease)
Use scenarios

- Cross evaluation of phenotypes
- Linking genes to diseases
  - Identification of candidate genes for Diabetes mellitus
  - Identification of candidate genes based on GO terms
  - Identification of candidate genes based on phenotype comparisons (mouse / human)
  - Identification of drugs for diseases (PharmGKB, PhenomeNet)
  - Identification of candidate genes for diseases, cross-species comparisons (PhenomeNet)
- Repurposing of drugs
### SESL project: Genes linked to Diabetes mellitus type II (DmT2)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Public &amp; proprietary data</th>
<th>[%]</th>
<th>Public data only</th>
<th>[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArrayExpress</td>
<td>182,840</td>
<td>0.5%</td>
<td>182,840</td>
<td>0.7%</td>
</tr>
<tr>
<td>EFO</td>
<td>49,026</td>
<td>0.1%</td>
<td>49,026</td>
<td>0.2%</td>
</tr>
<tr>
<td>UMLS, homebrew</td>
<td>6,906,735</td>
<td>18.8%</td>
<td>6,906,735</td>
<td>26.5%</td>
</tr>
<tr>
<td>Disease Ontology</td>
<td>1,863,664</td>
<td>5.1%</td>
<td>1,863,664</td>
<td>7.2%</td>
</tr>
<tr>
<td>Gene Ontology</td>
<td>495,595</td>
<td>1.3%</td>
<td>495,595</td>
<td>1.9%</td>
</tr>
<tr>
<td>UniProt filtered for Human</td>
<td>12,552,239</td>
<td>34.1%</td>
<td>12,552,239</td>
<td>48.2%</td>
</tr>
<tr>
<td>Triples on Meta-Data from FT Lit.</td>
<td>3,485,212</td>
<td>9.5%</td>
<td>1,949,293</td>
<td>7.5%</td>
</tr>
<tr>
<td>Triples with gene annot. From FT Lit.</td>
<td>2,373,584</td>
<td>6.5%</td>
<td>300,773</td>
<td>1.2%</td>
</tr>
<tr>
<td>Triples with disease annot. From FT Lit.</td>
<td>4,983,788</td>
<td>13.6%</td>
<td>662,824</td>
<td>2.5%</td>
</tr>
<tr>
<td>Triples with Go annot. From FT Lit.</td>
<td>3,870,834</td>
<td>10.5%</td>
<td>1,099,410</td>
<td>4.2%</td>
</tr>
<tr>
<td><strong>Total number of triples</strong></td>
<td><strong>36,763,517</strong></td>
<td><strong>40.0%</strong></td>
<td><strong>26,062,399</strong></td>
<td><strong>15.4%</strong></td>
</tr>
<tr>
<td><strong>Total number of public triples</strong></td>
<td><strong>14,713,418</strong></td>
<td><strong>40.0%</strong></td>
<td><strong>4,012,300</strong></td>
<td><strong>15.4%</strong></td>
</tr>
</tbody>
</table>
SESL project: Genes linked to Diabetes mellitus type II (DmT2)
### SESL project: Genes linked to Diabetes mellitus type II (DmT2)

<table>
<thead>
<tr>
<th>Phenotypes</th>
<th>OIMIM (+)</th>
<th>OIMIM (-)</th>
<th>OMIM (+)</th>
<th>OMIM (-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniProt (+)</td>
<td>UniProt (-)</td>
<td>UniProt (+)</td>
<td>UniProt (-)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Review (+)</th>
<th>GXA (+)</th>
<th>ABCC8, CAPN10, HNF1A, HNF1B (TCF2), HNF4A, INSR, NeuroD1, PPARG, TCF7L2</th>
<th>WFS1</th>
<th>IRS1, PDX1</th>
<th>HHEX, JAZF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review (+)</td>
<td>GXA (-)</td>
<td>GCK, KCNJ11</td>
<td>IGFBP2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Review (-)</td>
<td>GXA (+)</td>
<td>MAPK8IP1, PAX4</td>
<td>LIPC, PTPN1</td>
<td>GBP28 (ADIPOQ), PPP1R3A</td>
<td></td>
</tr>
<tr>
<td>Review (-)</td>
<td>GXA (-)</td>
<td>SLC2A4</td>
<td>IL6, RETN</td>
<td>INS</td>
<td></td>
</tr>
</tbody>
</table>
SESL project: Genes linked to Diabetes mellitus type II (DmT2)
Automatic analysis of the patient record for medical diagnostics

- The content is not well standardised.
- The data may be available in electronic form.
- We have efficient means to process the patient record and to draw conclusions.
- We can exploit multi-linguality and can identify hidden connections and can draw conclusions (semi-)automatically.
Comparison of patient and disease profiles

Disease profile (left): Generated from a database (OMIM).
Patient profile (right): Generated from the patient record.
Using similarity measures for the comparison (e.g. Cosinus)
Identification of 1,154 potential candidate genes

About 63 % could be verified from the scientific literature.
Reasoning over ontologies
Reasoning over ontologies

Hochndorf, R., ... Rebholz-Schuhmann, D. (2011) Bioinformatics. 2011 Feb 21

Anatomy ontologies:
- FMA, MA, WA, ZFA, FA
- GO-CC, ...
- (> 100,000 classes)

Quality ontology:
- PATO
- (> 2,000 classes)

Process and function:
- Gene Ontology, ...
- (> 25,000 classes)

Phenotype ontologies:
- HPO, MP, WBPhenotype, FBcv, APO, ...
- (> 20,000 classes)
Repurposing of drugs
Model for genes, diseases
Model for genes, diseases
Model for genes, diseases
Model for genes, diseases
Scenario 1: Diagnostics support

**Aim** Propose relevant diagnostics decisions

**Approach**
- Comparison of the patient profile against the profiles from the diseases (see above)
- Identification of relevant parameters, e.g. dysfunctions of organs (liver, kidney)
- Consider genetic parameters

**Gain**
- Decision support to the medical doctor.
- Integration of public data sources with patient information: easy access.
Motivation Phenotypes Analysis The future

Szenario 2: Risks linked to drugs

**Aim** Identification of adverse side effects in the patient profile.

**Approach**
- Determine the drug profile: pos./neg. regulation of metabolism X, Y; of physiological parameters A, B; of phenotypes O, P.
- Comparison of the drug profile against the laboratory results and the patient record.
- Statistical methods for the identification of synergistic effects.
- Comparison between doctor’s notes in the EHR against the predictions.

**Gain** Identification / reduction of unwanted adverse side effects of drugs.
Szenario 3: System medicine and Omim 2.0

**Aim**
Integration of the patient data in system medicine / system biology (translational medicine).

**Approach**
- Phenotype profile for all relevant diseases.
- Analysis and normalisation of public data resources: Omim, literature, databases in molecular biology, PharmGKB, ChEMBL.
- Identification of regulatory processes: positive regulation in process X increases parameter Y in phenotype Z
- Judgement on synergistic / antagonistic effects: identification if risks, e.g. over-regulation.

**Gain**
Comparison of the patient data against models from system biology, system medicin, personalized medicine, pharmacogenomics.
First Future: artificial intelligence

- IBM Watson: A cluster computer analyzes facts.
- Has been the champion in “Jeopardy” (US).
- Could predict successfully the right diagnosis on existing patients.
- The biomedical knowledge domain provides lot of ontologies and electronic data bases.
Second future: Big data

- 23andme: Genomics analysis for everybody.
- Exom sequencing for about 5,000 USD, a complete SNP set for 99 USD.
- 10,000 people will be sequenced, maybe 100,000 soon.
- Peta-Bytes of data in biomedical data bases.
- Which domain knowledge is not available yet, nor used yet?
Third future: social networks, Patients-like-me

- PatientsLikeMe: Exchange of patient-related data.
- Exchange of phenotype information.
Semantic resources are still the key element for the normalisation of data resources.

It is possible (easy) to process large data volumes in real time and on an everyday basis.

The semantic integration requires textual data in the first place, but can be applied to numerical data as well.

Profile comparisons, statistical methods and automatic reasoning are all equally important.

The integration of the patient record with data from molecular biology and genetics data resources becomes increasingly important.
Fragen?