Information technology solutions to support translational research on inherited cardiomyopathies

Riccardo Bellazzi
Università di Pavia, Italy

Valentina Favalli, Eloisa Arbustini
IRCCS Fondazione Policlinico S. Matteo, Pavia, Italy

Cristiana Larizza, Matteo Gabetta, Giuseppe Milani, Mauro Bucalo, Francesca Mulas, Angelo Nuzzo,
Università di Pavia, Italy

MIE 2001
OSLO 30/08/2011
Dilated cardiomyopathy

INHERITANCE PROJECT

Project Funded by the European Community's 7th Framework Programme

<table>
<thead>
<tr>
<th>Beneficiary Number</th>
<th>Beneficiary name</th>
<th>Beneficiary short name</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Coordinator</td>
<td>Fondazione IRCCS Policlinico San Matteo</td>
<td>OSM</td>
<td>IT</td>
</tr>
<tr>
<td>2</td>
<td>University College London</td>
<td>UCL</td>
<td>UK</td>
</tr>
<tr>
<td>3</td>
<td>Vasterbottens Lans Landsting</td>
<td>UUH</td>
<td>SW</td>
</tr>
<tr>
<td>4</td>
<td>Universitätsklinikum Heidelberg</td>
<td>UKLHD</td>
<td>DE</td>
</tr>
<tr>
<td>5</td>
<td>Institut National De La Sante et De La Recherche Medicale</td>
<td>INSERM</td>
<td>FR</td>
</tr>
<tr>
<td>6</td>
<td>Health-in-code SL</td>
<td>HIC</td>
<td>ES</td>
</tr>
<tr>
<td>7</td>
<td>The Chancellor, Masters and Scholars of the University of Cambridge</td>
<td>UCAM</td>
<td>UK</td>
</tr>
<tr>
<td>8</td>
<td>Aarhus Universiteitshospital, Skejby</td>
<td>SUH</td>
<td>DK</td>
</tr>
<tr>
<td>9</td>
<td>Academisch Medisch Centrum</td>
<td>AMC</td>
<td>NDL</td>
</tr>
<tr>
<td>10</td>
<td>Università degli Studi di Pavia, Dipt di Informatica e Sistemistica</td>
<td>UNIPV</td>
<td>IT</td>
</tr>
<tr>
<td>11</td>
<td>Università degli Studi di Milano, Dipt. di Scienze Biomolecolari &amp; Biotecnologia</td>
<td>UMIL</td>
<td>IT</td>
</tr>
</tbody>
</table>
DCM: the past and wrong paradigm of post-viral disease

Dilated Cardiomyopathy
30 years of research

Viral etiology

Genetic/familial. Data from family screening studies and serial monitoring of family members
From DCM to …

Clinically oriented genetic investigation

“DCM”

Dystrofinopathies
Laminopathies
Desminopathies
Mitocondriopathies
Epicardinopathies
Actinopathies
Zaspopathies
Desmosonopathies
More than 35 genes may cause DCM
DCM is sometimes accompanied by gene-specific traits → red flags
Grouping patients according to phenotypes:
- DCM
- type of inheritance
- cardiac markers
- extracardiac markers
- any clinical data that may “specify” the subgroups
The IT architecture

- Annotation tools
- Text mining and literature search engines
- Reasoning module
- KB/Red flags

Cardioregister

Data warehouse

I2b2 environment

Web interface

Data analysis plugin

Wiki-based collaborative system
ABSTRACTS
NCBI

DB
ABSTRACTS
NLP

i2b2

NCBI

Genes & Pathologies

Literature Subset

LBD (Literature Based Discovery)

Patient Data

DB
INHERITANCE

Literature Results

DSS (Decision Support System)

Patient Data

MIE 2011
Medical Literature Analysis

DCM Genes, DCM

PubMed

NLP

Paper DB

Biological Concepts (UMLS, Entrez Gene)
Experiment on knowledge retrieval

Abstract corpus

Dilated cardiomiopathy

2005-2010

ABOUT 7000 DOCS
SEARCH AND COMPARE

Abstract corpus

NCBI GENE DB

THROUGH UMLS

RED FLAGS

- ABSTRACTS RELATED TO GENES OF THE GENE LIST
- ABSTRACTS RELATED TO THE RED FLAGS
- GENES AND RED FLAG RELATIONSHIPS

GENE LIST
ENRICHING THE SEARCH

Abstract corpus

THROUGH UMLS

RED FLAGS

GENE-RED FLAG RELATIONSHIPS + POTENTIAL NEW RED FLAGS

GENE PRIORITIZATION

NCBI GENE DB

GENE LIST

MIE 2011
## Phenotype-Genes Map

<table>
<thead>
<tr>
<th>UMLS ID</th>
<th>Phenotype</th>
<th>Gene Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0011053</td>
<td>Deafness</td>
<td>PTPN11</td>
</tr>
<tr>
<td>C0018784</td>
<td>Sensorineural Hearing Loss</td>
<td>EYA4</td>
</tr>
<tr>
<td>C0581883</td>
<td>Complete Hearing Loss</td>
<td>PTPN11</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Case-based-ranking

Assign a score to genes based on similarity of the clinical case with previous, and already known, clinical cases

PTPN11 → DCM, HCM
EYA4 → DCM

Noonan syndrome

Current patient

Patient 1 – gene PTPN11
Patient 2 – gene EYA4...

Case base
Details on case-based ranking

\[ \text{Sim}(P_N, P_x) = w^+ \cdot \text{semDist}(P_N^+, P_x^+) + w^- \cdot \text{semDist}(P_N^-, P_x^-). \]

Red Flags

Hierarchical Representation

Similarity Score

(modification of Melton et al)

\[ \text{semDist}(P_A, P_B) = \frac{\sum_i \min_j (\text{clinDist}(P_{A_i}, P_{B_j}))}{\text{size}(P_A)}. \]
Test of the distance metric

- Simulation based on real complete cases
- Healthy (30)
- DCM (20)
- Lamin mutation (40)
- Dystrophin mutation (10)
State of art and next

- Inheritance: 18 months intermediate report
- Data base up and running and data warehouse populated (http://i2b2inheritance.com/)
- Data from 394 patients and relatives from 229 families
- 1140 cases already collected off-line
- CBR tested on simulated data
- New strategies to find cases “on the net”
- NGS data soon available
Acknowledgements

Cristiana Larizza, Matteo Gabetta, Giuseppe Milani, Mauro Bucalo, Francesca Mulas, Angelo Nuzzo,
Dipartimento di Informatica e Sistemistica,
Università di Pavia, Italy

Valentina Favalli, Eloisa Arbustini
IRCCS Fondazione Policlinico S. Matteo, Pavia, Italy