The Human Phenotype Ontology
Second CASIMIR Annual Meeting 2008

Sebastian Köhler, Sebastian Bauer, Denise Horn, Dominik Seelow, Stefan Mundlos, Peter N. Robinson

November 30, 2008
Computational Analysis of Human Phenotypes

- (In the eyes of humans) the human phenotype is the richest and most variable and nuanced among all organisms
- This also applies to the human disease phenotype
- A dedicated ontology for human disease features is needed
Mendelian Inheritance in Man

- The single most valuable resource for human genetics\(^a\)
- The first edition in 1966 was 344 pages;
- Online for over a decade

\(^a\)Just my humble opinion!
Online Mendelian Inheritance in Man

▶ It has been difficult to use the information in OMIM for large-scale computational analysis

▶ Synonyms:
  ▶ the descriptions "generalized amyotrophy", "generalized muscle atrophy", "muscular atrophy, generalized", and "muscle atrophy, generalized" are used to described different diseases in OMIM, and may not be easily recognized as synonyms using a purely computational approach.

▶ Limits of hierarchical structures
  ▶ the hierarchical structure does not itself reflect that (for instance) "Hypoplastic philtrum" and "Smooth philtrum" are more closely related to one another than "Hypoplastic nasal septum" (all three of these terms are in the Nose category of OMIM’s clinical synopsis).
Other groups have used text-mining approaches to show that phenotypic similarities are positively correlated with a number of measures of gene function, including relatedness at the level of protein sequence, protein motifs, functional annotation, and direct protein-protein interaction.

Constructing the HPO

<table>
<thead>
<tr>
<th>Name</th>
<th>Terms</th>
<th>Annotations</th>
<th>Terms with 1 annotation</th>
<th>terms with &gt; 1 annotation</th>
<th>Entries with specific annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) CARDIAC</td>
<td>304</td>
<td>1198</td>
<td>211</td>
<td>987</td>
<td>82.39%</td>
</tr>
<tr>
<td>2) JOINTS</td>
<td>164</td>
<td>200</td>
<td>144</td>
<td>56</td>
<td>28.00%</td>
</tr>
<tr>
<td>3) ABDOMEN</td>
<td>657</td>
<td>1925</td>
<td>510</td>
<td>1415</td>
<td>73.51%</td>
</tr>
<tr>
<td>4) BREASTS</td>
<td>43</td>
<td>167</td>
<td>27</td>
<td>140</td>
<td>83.83%</td>
</tr>
<tr>
<td>5) NOSE</td>
<td>280</td>
<td>1194</td>
<td>206</td>
<td>988</td>
<td>82.75%</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>72) LUNGS</td>
<td>225</td>
<td>536</td>
<td>181</td>
<td>355</td>
<td>66.23%</td>
</tr>
</tbody>
</table>

- All descriptions used at least twice (approx 7000) were assigned to an HPO term
- Synonyms were merged
- All assignments based on expert judgment
- Singleton descriptions (~ 30000) were for the most part not assigned to terms. Annotations were made as 'Abnormality of [Category]'
The Human Phenotype Ontology

▶ HPO: ~ 8700 terms, annotations for 4735 diseases
Freely available at the HPO website
human-phenotype-ontology.obo: 8745 terms
phenotype_annotation.omim: 46400 annotations for diseases listed in OMIM
The PhenExplorer is a browser application using Ajax/Google Web Toolkit.
Searching for diseases listed in OMIM
Show list of features associated with the disease
Browsing the HPO

- Visualize features within ontology DAG
<table>
<thead>
<tr>
<th>term</th>
<th>OMIM</th>
<th>genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All</strong></td>
<td></td>
<td><strong>NRAS, PDGFRL, APC, DCC, AXIN2, AURKA, MLH3, PTOPJ, PIK3CA, TLR3, TLR4, BUB1B, PLA2G2A, FLCN, AKT1, EP300, TP53, SMAD7, NPC1, TP53</strong></td>
</tr>
<tr>
<td><strong>Inheritance</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#114500 COLORECTAL CANCER: CRC; COLON CANCER</td>
<td></td>
<td><strong>NRAS, PDGFRL, APC, DCC, AXIN2, AURKA, MLH3, PTOPJ, PIK3CA, TLR3, TLR4, BUB1B, PLA2G2A, FLCN, AKT1, EP300, TP53, SMAD7, NPC1, TP53</strong></td>
</tr>
<tr>
<td>#161450 NASOPHARYNGEAL CARCINOMA: NPCA, NPC; NASOPHARYNGEAL CANCER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--- Autosomal dominant inheritance ---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#114480 BREAST CANCER: BREAST CANCER, FAMILIAL BREAST CANCER, FAMILIAL MALE INCL</td>
<td></td>
<td><strong>RAD51, KRAS, BRIP1, PPM1D, BARD1, CASP8, HMME, RB1CC1, PA28G, SLC22A18, BRCA2, PIK3CA, CHEK2, AKT1, ATM, TP53</strong></td>
</tr>
<tr>
<td>#114550 HEPATOCELLULAR CARCINOMA: HCC; CANCER, HEPATOCELLULAR; LIVER CANCER; LIVER CELLS; CARCINOMA: LCC; HEPATOMA; HEPATOBLASTOMA; INCL</td>
<td></td>
<td><strong>PDGFRL, CASP8, AXIN1, MET, PIK3CA, CTNNB1, TP53</strong></td>
</tr>
<tr>
<td>#115623 LI-FRAUMENI SYNDROME 1: LFS1; SARCOMA FAMILY SYNDROME OF LI AND FRAUMENI; SBLA SYNDROME; LI-FRAUMENI-LIKE SYNDROME; INCL</td>
<td></td>
<td><strong>CDKN2A, TP53</strong></td>
</tr>
</tbody>
</table>

- **Mapping Genes ↔ HPO**
Using the HPO: Calculating phenotypic similarity

- Start with ontology and frequency of annotations to individual terms among a corpus of diseases
Using the HPO: Calculating phenotypic similarity

- Information content of term $t$: $IC(t) = -\log p(t)$
- Term L has information content: $IC(L) = -\log \frac{3}{100} = 3.5066$
Using the HPO: Calculating phenotypic similarity

- General terms have lower information content
- Term C has information content: $\text{IC}(C) = -\log \frac{80}{100} = 0.22314$
Using the HPO: Calculating phenotypic similarity

\[
\text{sim}(t_1, t_2) = \max_{a \in A(t_1, t_2)} - \log p(a),
\]

- Similarity between \(t_1\) and \(t_2\): IC of most specific ancestor
- \(\text{sim}(L, M) = \text{IC}(I) = - \log 7/100 = 2.6593\)
Using the HPO: Calculating phenotypic similarity

\[
\text{sim}(t_1, t_2) = \max_{a \in A(t_1,t_2)} - \log p(a),
\]

- Lower similarity for pairs of terms more distant to one another
- \( \text{sim}(L, K) = \text{IC}(B) = - \log 90/100 = .10536 \)
Using the HPO: Calculating phenotypic similarity

How do we now calculate the similarity between two sets of annotations (two diseases?)

\[
\text{sim}(d_1 \rightarrow d_2) = \text{avg} \left[ \sum_{s \in d_1} \max_{t \in d_2} \text{sim}(s, t) \right]
\]  

(2)
Using the HPO: Calculating phenotypic similarity

\[ \text{sim}(d_1 \rightarrow d_2) = \text{avg} \left[ \sum_{s \in d_1} \max_{t \in d_2} \text{sim}(s, t) \right] \]

\[ = \text{avg} \left[ \text{IC}(B) + \text{IC}(J) \right] = 1.7148 \]
Using the HPO: Calculating phenotypic similarity

\[
sim(d_2 \rightarrow d_1) = \text{avg} \left[ \sum_{t \in d_1} \max_{s \in d_2} \text{sim}(s, t) \right]
\]

\[
= \text{avg} [\text{IC}(A) + \text{IC}(B) + \text{IC}(J)] = 1.7148
\]
Using the HPO: Calculating phenotypic similarity

\[ \text{sim}(d_1, d_2) = \frac{1}{2} \times \text{sim}(d_1 \rightarrow d_2) + \frac{1}{2} \times \text{sim}(d_2 \rightarrow d_1) \] (3)

- The similarity measure is in general not symmetric
- In our example we have
  \[ \text{sim}(d_1, d_2) = 0.5 \times 1.7148 + 0.5 \times 1.7148 \]
Visualization of the Human Phenome using the HPO (Categories from Goh et al., PNAS 2008)
2116 diseases in OMIM annotated to at least 6 HPO terms
Search for pulmonic stenosis and low-set posteriorly rotated ears
Diseases with significant matches according to the similarity measure are listed in order
Medical Genetics Diagnostics: Plans

- Filter by Inheritance
- Show clickable list of features that best discriminate among top possibilities
- Weight features by frequency in each disease
- Add annotation data for chromosomal diseases
  - Chromosomal Variation in Man Online Database
  - ECARUCA
- Add annotation data for microdeletion & CNV diseases
  - DECIPHER
Finding disease genes

- Positional cloning studies identify intervals containing up to hundreds of genes.
Disease-gene families (1)

Genetically heterogeneous disorders: e.g. **Bardet Biedl Syndrome**

- Obesity
- Polydactyly
- Retinopathy

Bardet Biedl Syndrome: Mutations in Ciliary Proteins
Clinically overlapping diseases, e.g. **RAS/MAPK Pathway**

**Syndromes**

- Noonan-Syndrome
- LEOPARD-Syndrome
- CFC-Syndrome
- Neurofibromatose Typ 1

**Disease-gene families (2)**

- CFC-Syndrome: HRAS, Noonan/CFC - KRAS
- Noonan-LEOPARD: SHP-2, Grb2, SOS, Ras-GDP, Ras-GTP
- Neurofibromin: NF1, NFNS
- Ral-GDS Pathway
- PI3 Kinase Pathway
- Transcription of Target Genes
Random Walk with Restart

\[ p^{t+1} = (1 - r)Wp^t + rp^0 \]  \hspace{1cm} (4)

- \( p^{t+1} \): Iteratively wander throughout the interaction network
- \( p^0 \): Initialize the random walker to start with equal probability at each of the known disease genes
- \( W \): Represent protein interactions as an adjacency matrix
- \( r \): Restart from initial configuration with probability \( r \)
- Iterate until convergence
Test scenario: 110 disease-gene families (783 Gene)

- 96 monogenic
- 12 polygenic
- 12 cancer
Example: Stickler Syndrome

Known disease genes: \textit{COL2A1}, \textit{COL9A1}, \textit{COL11A1}, \textit{COL11A2}: All ranked in first place by random walk
GeneWanderer is a computational method to prioritize a set of candidate genes according to their probability of being involved in particular disease or phenotypes. Our approach measures the relative location of each candidate gene in a genomic interval (as defined by linkage analysis or other methods) to genes known to be involved in the phenotypic/disease under investigation in the protein-protein interaction (PPI) network. By using the random walk or diffusion kernel global characteristics of the network are taken into account. The shortest path and direct interaction method use local properties of the PPI network.

For more information:
http://compbio.charite.de/genewanderer
HPO and Disease Gene Prioritization

- HPO can be used to automatically define disease-gene families
- For any gene being sought, choose the $n$ phenotypically nearest disorders, and use the associated genes as a disease-gene family
- Work in preparation
Other Plans

- Define phenotypic networks & relationship to cellular networks for skeletal dysplasias and other areas
- Improve integration with out OBO ontologies esp. PATO, MPATH, MPO
- Refine annotations (frequency, severity data)
- Looking for collaborators: ”Adopt a phenotype”
For more information:
http://www.human-phenotype-ontology.org
also at http://www.obofoundry.org/
Thank you for your attention ....

Computational Biology and Bioinformatics Group:
Sebastian Bauer
Sebastian Köhler
Christian Rödelsperger
Marten Jäger
http://compbio.charite.de

Institute of Medical Genetics, Charité
Denise Horn
Stefan Mundlos

Department of Neuropaediatrics, Charité
Dominik Seelow

Funding: DFG, BMBF