

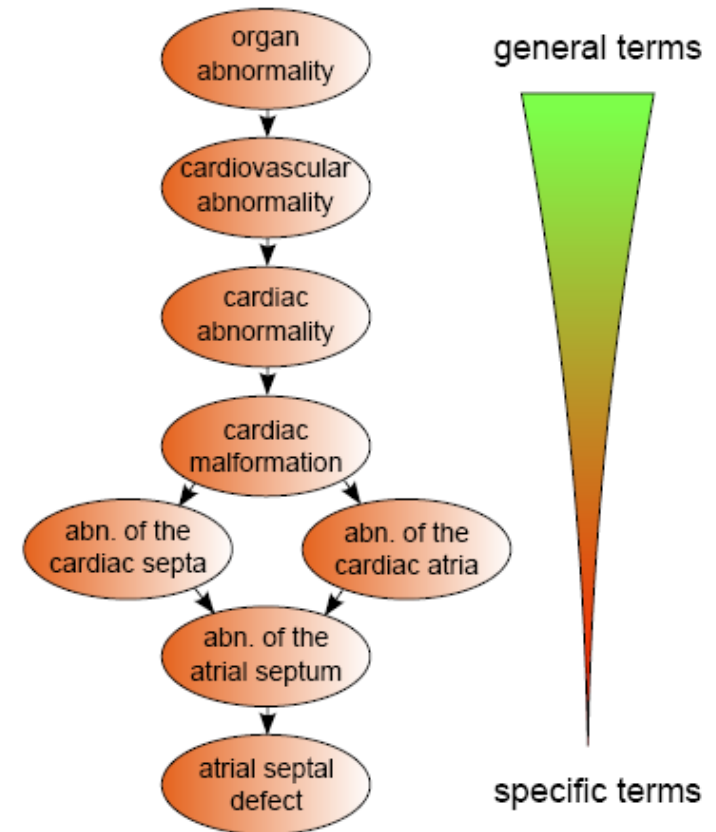
Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies

Signs, Symptoms and Findings:
Towards an Ontology for clinical Phenotypes
Milan, September 2009
Peter Krawitz

The Human Phenotype Ontology (HPO*)

www.human-phenotype-ontology.org

- The HPO provides a standardized vocabulary of phenotypic abnormalities encountered in human genetic syndromes
- Phenotypic features are formally represented as terms of a directed acyclic graph:
 - Terms are related to parent terms by „is a“ relationships, representing subclasses of more general parent terms
 - Multiple parentage allows the representation of different aspects of phenotypic abnormalities



* Robinson P, Köhler S, Bauer S, Seelow D, Horn D, Mundlos: The Human Phenotype Ontology: A Tool for annotating and analyzing human hereditary disease, *Am J Hum Genet.* 2008 Nov

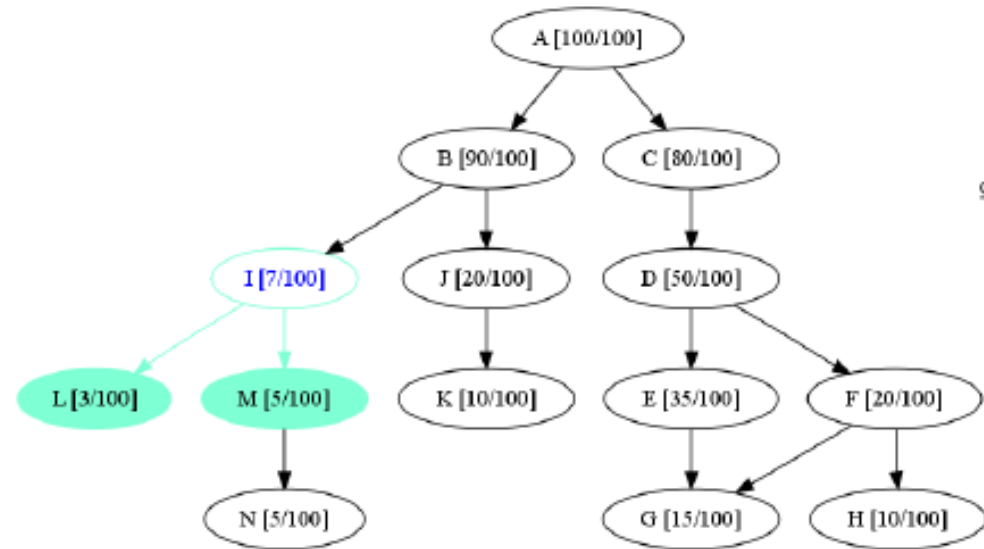
Using the HPO to calculate Phenotypic Similarities

- The importance of a clinical phenotypic finding of the differential diagnosis depends on its specificity
- The specificity of a phenotypic feature, t , is represented by its information content (IC), defined as negative natural logarithm of its frequency of occurrence:

$$IC(t) = -\ln p(t)$$

- The similarity between two terms t_1 and t_2 is defined as the IC of their most specific common ancestor:

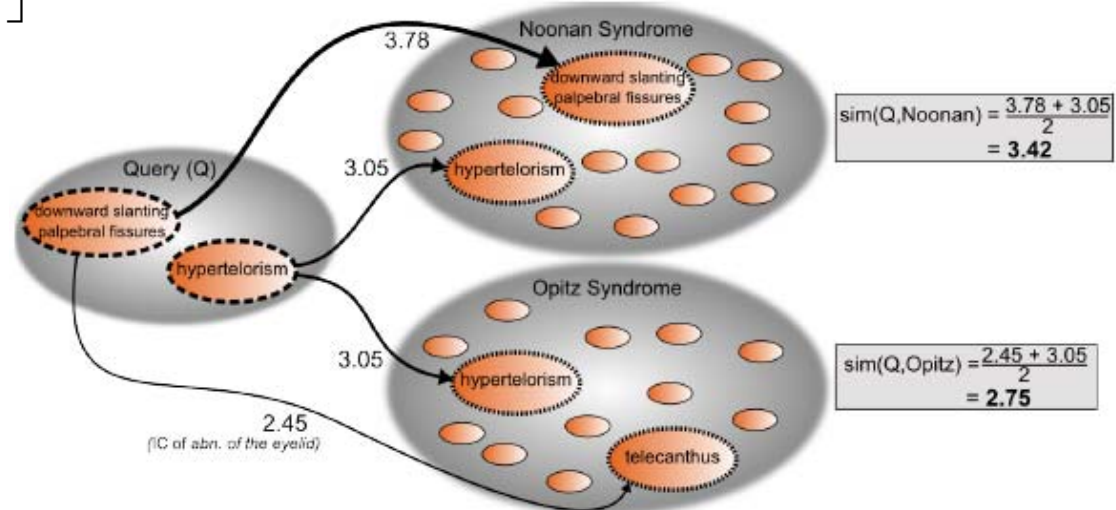
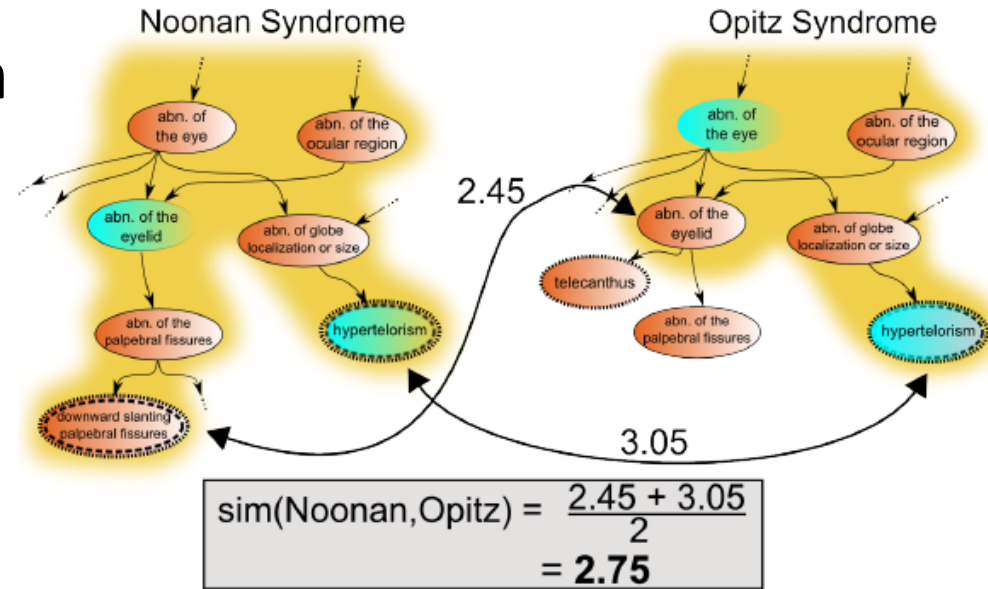
$$sim(t_1, t_2) = \max_{a \in A(t_1, t_2)} (-\ln p(a))$$



Ontological Similarity Search

- Term similarity measures may be used to compute similarity scores between genetic syndromes or phenotypic feature query sets:

$$sim(D_1 \rightarrow D_2) = avg \left[\sum_{t_i \in D_1} \max_{t_j \in D_2} (sim(t_i, t_j)) \right]$$



The Phenomizer*: A Tool for Diagnostics

<http://compbio.charite.de/Phenomizer/Phenomizer.html>

Menu Help

The Phenomizer - Next Generation Diagnostics

Features Diseases Ontology

PULMONIC STENOSIS search reset

HPO Id.	Feature
HP:0004957	PERIPHERAL PULMONARY STENOSIS
HP:0001642	PULMONIC STENOSIS

Patient's Features

HPO	Feature	Modifier
category: ABNORMALITY OF THE EARS (1 Item)		
HP:0000368	LOW-SET, POSTERIORLY ROTATE	observed
category: CARDIOVASCULAR ABNORMALITY (1 Item)		
HP:0001642	PULMONIC STENOSIS	observed

Page 1 of 1

Displaying features 1 - 2 of 2

Clear Mode of inheritance Get diagnosis.

Köhler S, Schulz M, Krawitz P, Bauer S, Dölken S, Ott C, Mundlos C, Horn D, Mundlos S, Robinson PN:

Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies, *Am J Hum Genet.* 2009 Sep.

The Phenomizer*: A Tool for Diagnostics

<http://compbio.charite.de/Phenomizer/Phenomizer.html>

The screenshot displays the Phenomizer web application interface. The main window is titled "The Phenomizer - Next Generation Diagnostics". It features a "Features" tab on the left and a "Diagnosis" tab on the right. The "Features" tab shows a search for "PULMONIC STENOSIS" with results for HPO IDs HP:0004957 (PERIPHERAL PULMONARY STENOSIS) and HP:0001642 (PULMONIC STENOSIS). A red arrow points to a "Specific search" window with a table of HPO IDs and features. The "Diagnosis" tab shows the results of a Resnik algorithm search, listing diseases and associated genes. A red box highlights the top result: Noonan Syndrome 1 (OMIM name) with a p-value of 0.0088 and the gene PTPN11.

p-value	OMIM name	Genes
0.0088	NOONAN SYNDROME 1	PTPN11
0.4608	FRONTOCULAR SYNDROME	
0.4608	HYPERTELORISM AND TETRALOGY OF FALLOT	
0.5854	CARDIOFACIOCUTANEOUS SYNDROME	BRAF
0.5854	COSTELLO SYNDROME	HRAS, KRAS
0.5854	NOONAN-LIKE/MULTIPLE GIANT CELL LESION SYNDROME	PTPN11
0.5854	EMANUEL SYNDROME	
0.5854	SUBAORTIC STENOSIS-SHORT STATURE SYNDROME	
0.5854	PULMONIC STENOSIS AND DEAFNESS	

Köhler S, Schulz M, Krawitz P, Bauer S, Dölken S, Ott C, Mundlos C, Horn D, Mundlos S, Robinson PN: Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies, *Am J Hum Genet.* 2009 Sep.

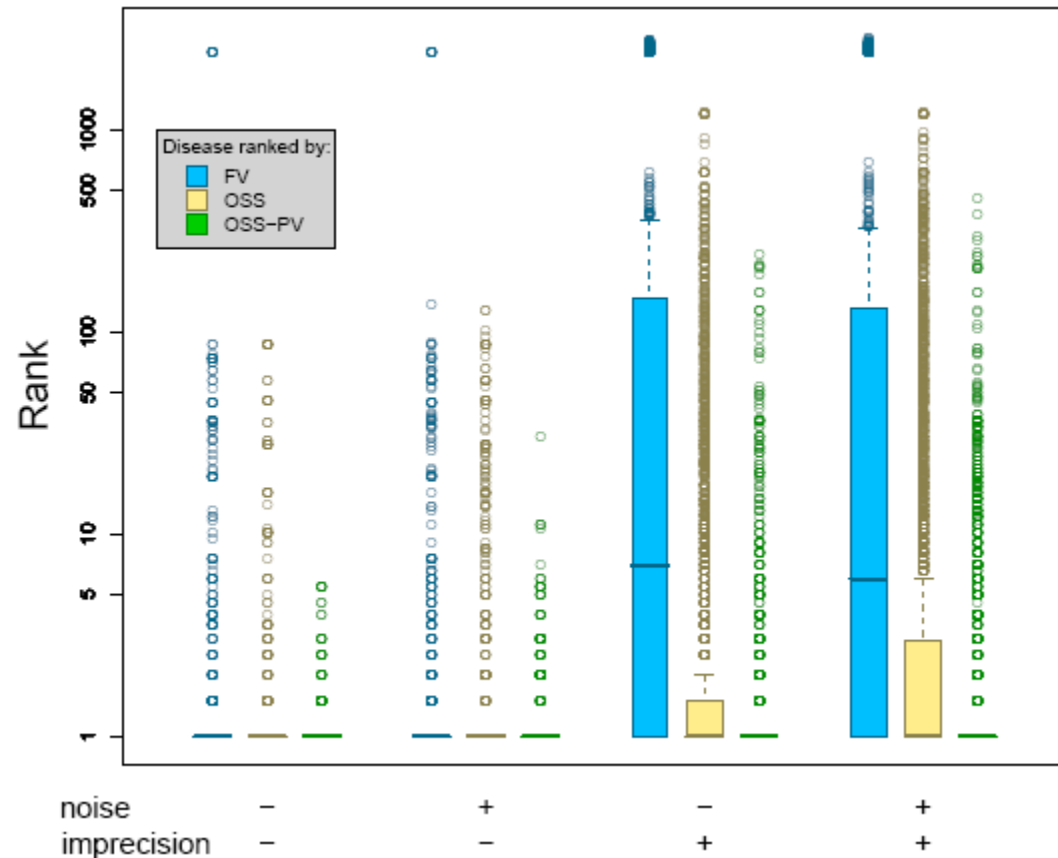
P-value Calculation

- The raw similarity score S depends on the number and specificity of terms both of the query Q and of the diseases D represented in the database
- The distribution of similarity scores can be used to obtain the significance of search results
- The P-value for the null hypothesis that a similarity score of S or greater for a set of query terms Q and a disease D has been observed by chance is defined as:

$$P(s \geq S) = \frac{\text{number of queries such that } \text{sim}(Q, D) \geq S}{\text{total number of possible queries}}$$

Evaluation of *P*-value ranking

- For 44 dysmorphology syndromes different queries of hypothetical patients were simulated based on the phenotypic annotation of the syndrome
- Ranks of the diseases returned by the phenomizer were compared to the original diagnosis
- Three ranking methods were compared:
 - Feature vector comparison
 - Ontological Semantic Similarity (OSS) Score
 - *P*-value of OSS
- Comparisons performed with phenotypic noise and “imprecision”



Acknowledgments

- Computational Biology at Charité

Sebastian Köhler

Christian Rödelsperger

Sebastian Bauer

Martin Jäger

Peter Nick Robinson

- Medizinische Genetik at Charité

Sandra Dölken

Denise Horn

Rici Ott

Stefan Mundlos

...

