



# Mining Phenotypes for Protein Function Prediction

EBI Industry Workshop 'Disease Ontologies and Information' – 19 June 2008

Philip Groth, Ulf Leser, Hans-Dieter Pohlenz, Bertram Weiss

# Phenotypes

---

- What is a phenotype?
  - Visible characteristic of an organism
  - Description of a disease
  - Response to a drug
  - Characterization of mutants
  - Results of RNAi / gene knock-out
  - Expression levels of genes
  - ...
- Describing phenotypes
  - Text, keywords, abstracts, community-specific vocabulary
  - The future: structured data, cross-species phenotype ontology
- systematic measurement of phenotypes has gone high-throughput (mutant screens, RNAi screens)





# Genotype-Phenotype Relationship: General Idea

---



- Molecular basis for similar phenotypes ?
  - Different molecular defects (e.g. mutation, truncation) in the very same gene
  - Defects in genes of the same biological process, module, pathway
- Hypothesis: If genes have similar phenotypes they are part of the same biological process, module or pathway.
- Workplan
  - get phenotypes
  - cluster phenotypes based on similarity (text-clustering)
  - evaluate biological coherence of resulting phenotype<->genotype cluster

Groth P, Weiss B, Pohlenz HD, Leser U.  
*Mining phenotypes for gene function prediction.*  
BMC Bioinformatics (2008) 9: 136.

# Get Phenotypes -> 'Phenodocs'

---

411,102 textual phenotype descriptions

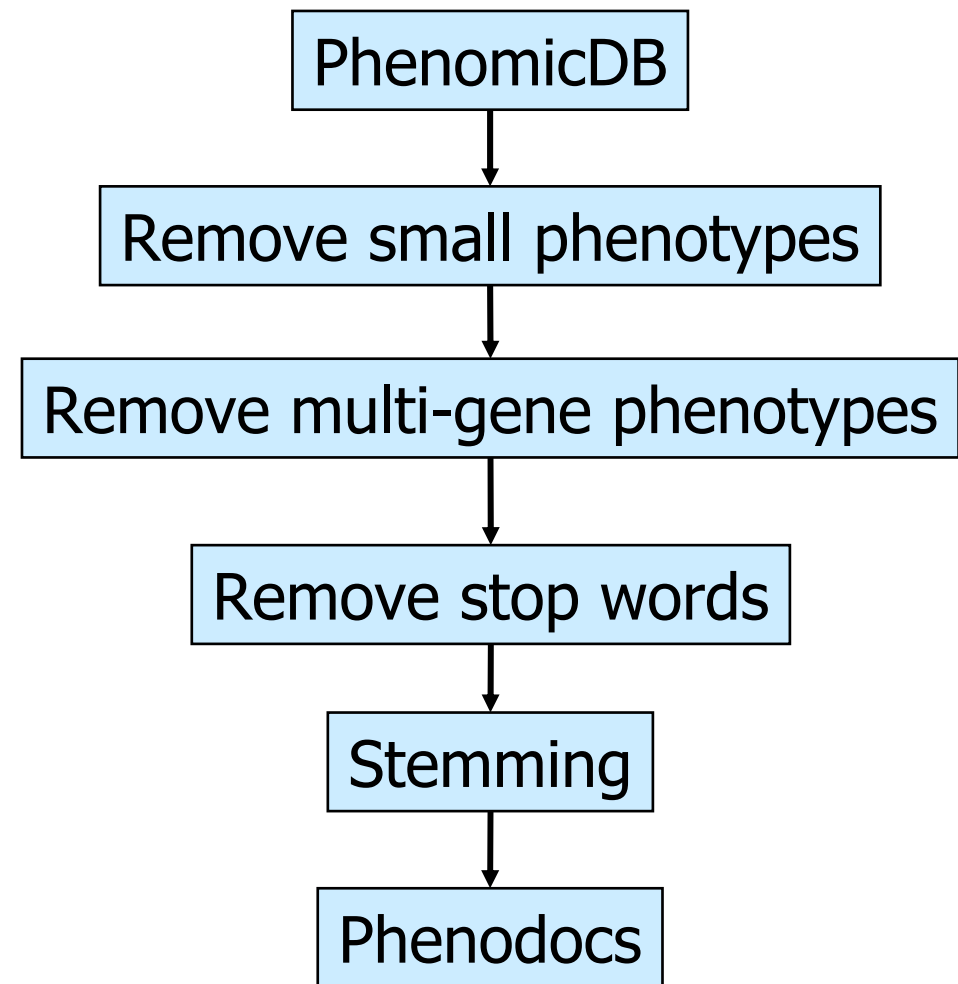
<250 words

phenotypes associated to more than one gene (~500)

e.g. and, this, or, the ....

mitochondrial, mitochondrium

39,610 phenodocs  
relating to 15,426 genes



# Cluster Phenotypes -> phenodoc similarity

---

- Every phenodoc is converted into a vector
  - Dimensions: All different words in all phenodocs
  - Order of words (and hence grammar) is lost
  - Value of a dimension is binary (word in/not in doc) or TF\*IDF

**Wilson's disease**, an autosomal recessive disorder, is characterized by the excessive accumulation of copper in the liver. **WND** gene, which encodes a putative **copper transporting** P-type ATPase, is defective in the patients. To investigate the /in vivo/ function of **WND**

...



(0,0,0.8,0,0,0,1,0,0,0,0...,0.78,0.4,0,0,...0)

**Wilson disease** is an autosomal recessive copper transport disorder resulting from defective biliary excretion of **copper** and subsequent hepatic copper accumulation and liver failure if not treated. The disease is caused by mutations in the /**ATP7B**/ (/WND/) gene, which is expressed predominantly in the liver and encodes a copper-transporting

...

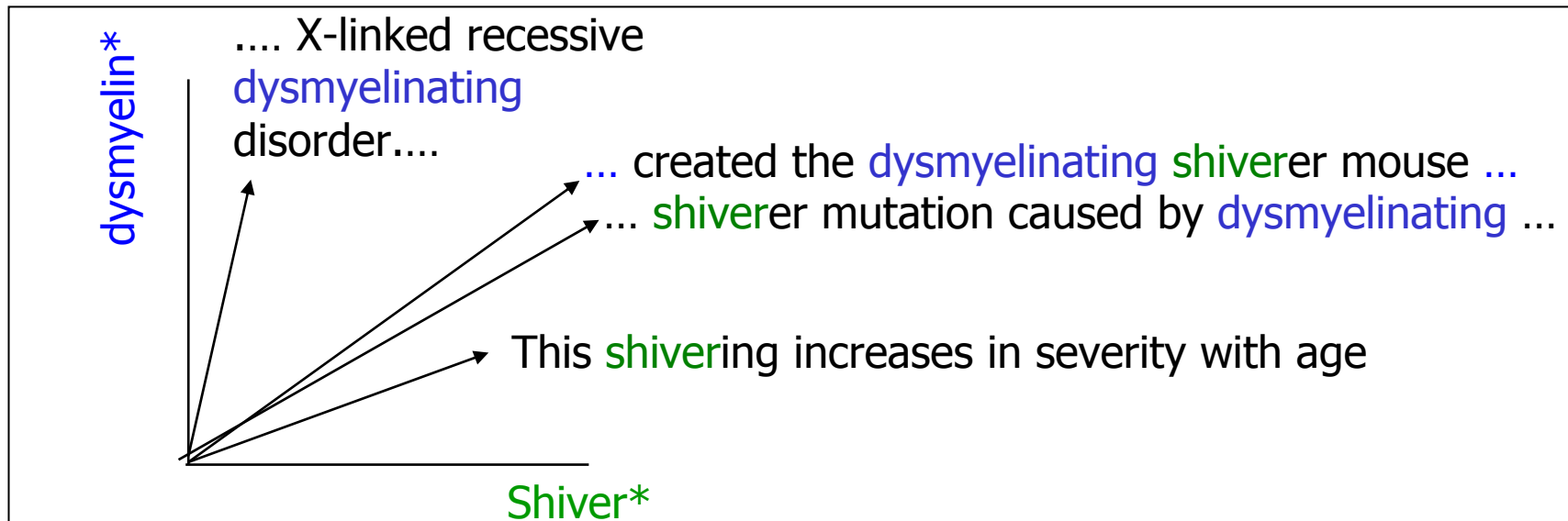


(0.4,0,0,0,1,0,1,0,0,0,0...,0.25,0.38,0,0,...0)

# Cluster Phenotypes -> phenodoc similarity

- Cosine distance of vectors in a multi-dimensional feature space

$$\text{sim}(\vec{x}, \vec{y}) = \cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| \cdot |\vec{y}|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \cdot \sqrt{\sum_{i=1}^n y_i^2}}$$



- K-Means to cluster phenodocs.
- Number of clusters must be predefined (250 ... 3000 clusters)

# Biological coherence of phenotype clusters

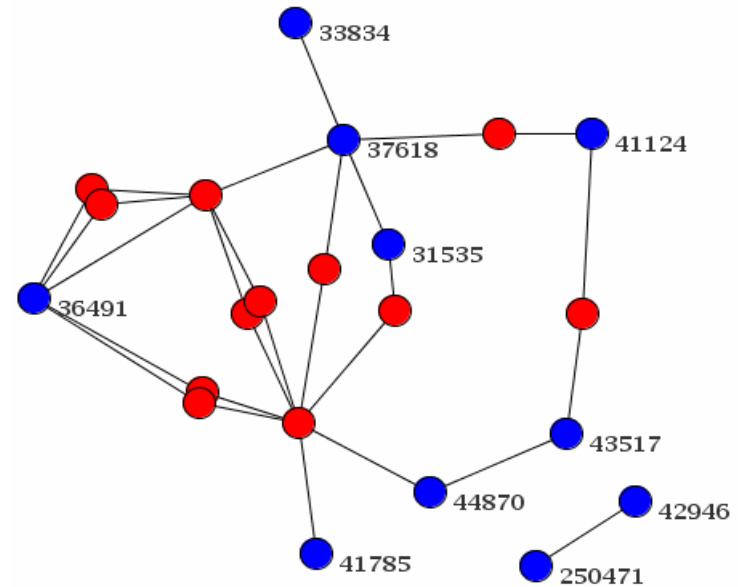
---

- Check if similar phenotypes of a cluster have more biological coherence than random cluster of equal size
  - Protein-protein interactions (interconnectedness)
  - GO annotations of associated genes (GO similarity)
  - Phenocopies tend to cluster (manual curation) ?
- Summary of results:
  - Phenotype clusters have more protein-protein interactions than random clusters
  - Phenotype clusters share more common or similar GO terms than random clusters
  - Genes of *phenocopies* tend to co-occur in clusters

# Biological Coherence: more protein-protein interactions than random

---

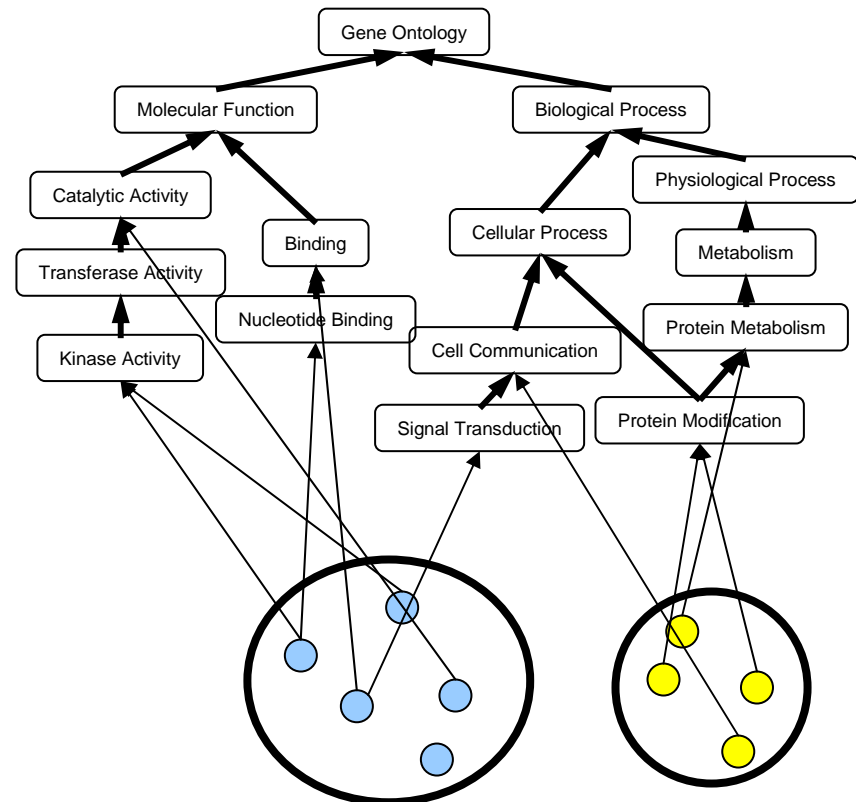
- PPI from BIOGRID database (edges in graph)
- In  $\sim 200$  clusters,  $>30\%$  of genes interact with each other ( $P \leq 0.003$ )
- Result: Genes in phenoclusters interact with each other more often than expected by chance



*Interaction network of genes from a phenotype cluster.  
Red proteins have no phenotypes in PhenomicDB but interact with cluster members*

# Biological Coherence: more GO Terms than random

- Comparison of GO annotations of genes in phenoclusters
- ~200 clusters with score  $>0.4$  ( $p \leq 0.002$ )
- Results: Genes in phenoclusters have a much higher coherence in functional annotation than expected by chance

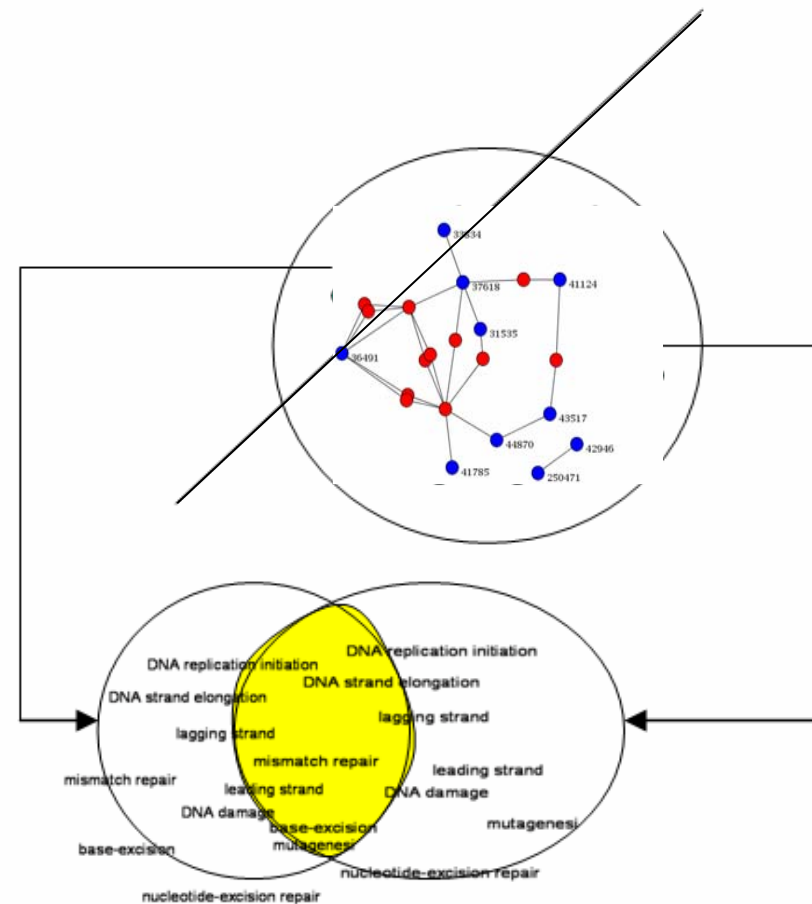


# Function Prediction

---

- Can increased functional coherence of clusters be exploited for **function prediction**?

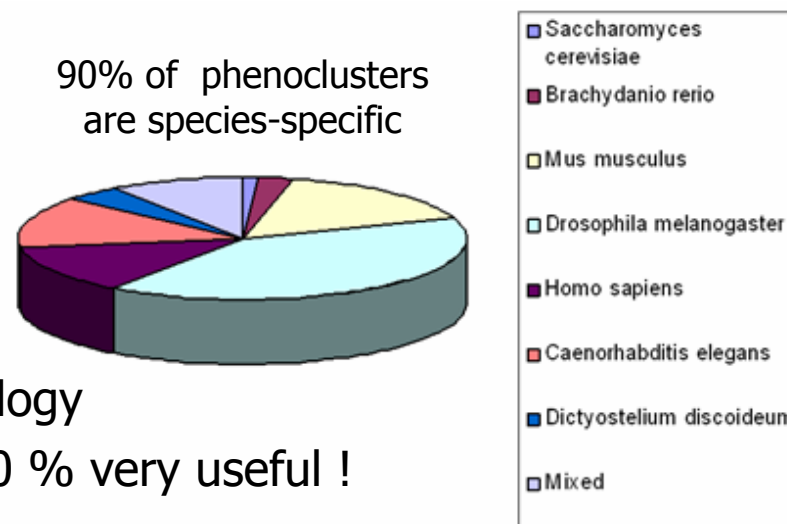
- Complementary approach
- Comparable to other function prediction methods
- **Precision 71% and recall 27% (k=1000)**



# Conclusions

---

- Similarity of phenotype descriptions is indicative for
  - high degree of protein-protein interaction within a cluster
  - homogeneity of functional annotation
- Phenotype clustering can be used for function prediction
  - Adding to sequence, structure, motifs, PPI networks, orthology, ...
- Open issues
  - Still many genes with no or incomplete phenotype
  - Handling of the many, many short phenotype descriptions ?
  - Phenotype descriptions have little structural granularity allowing only for blunt methods like text-clustering
  - Lack of cross-species phenotype ontologies. Community-specific terminology
  - ~ 80% of clusters are not useful yet, 20 % very useful !



# Acknowledgements



Bayer HealthCare  
Bayer Schering Pharma



- Bayer Schering Pharma AG
  - **Philip Groth**
  - Hans-Dieter Pohlentz
- Humboldt-University
  - Ulf Leser



- <http://www.phenomicdb.de/>

- Groth P, Weiss B, Pohlentz HD, Leser U.  
*Mining phenotypes for gene function prediction.*  
BMC Bioinformatics (2008) **9**: 136.
- Groth P, Pavlova N, Kaley I, Tonov S, Georgiev G, Pohlentz HD, Weiss B.  
*PhenomicDB: a new cross-species genotype/phenotype resource.*  
Nucleic Acids Res. 2007 **35**(Database issue):D696-9
- Groth P & Weiss B  
*Phenotypes: a neglected resource in biomedical research?*  
Current Bioinformatics (2006) **1**(3), 347–358