

# **Instance-based matching of large life science ontologies**

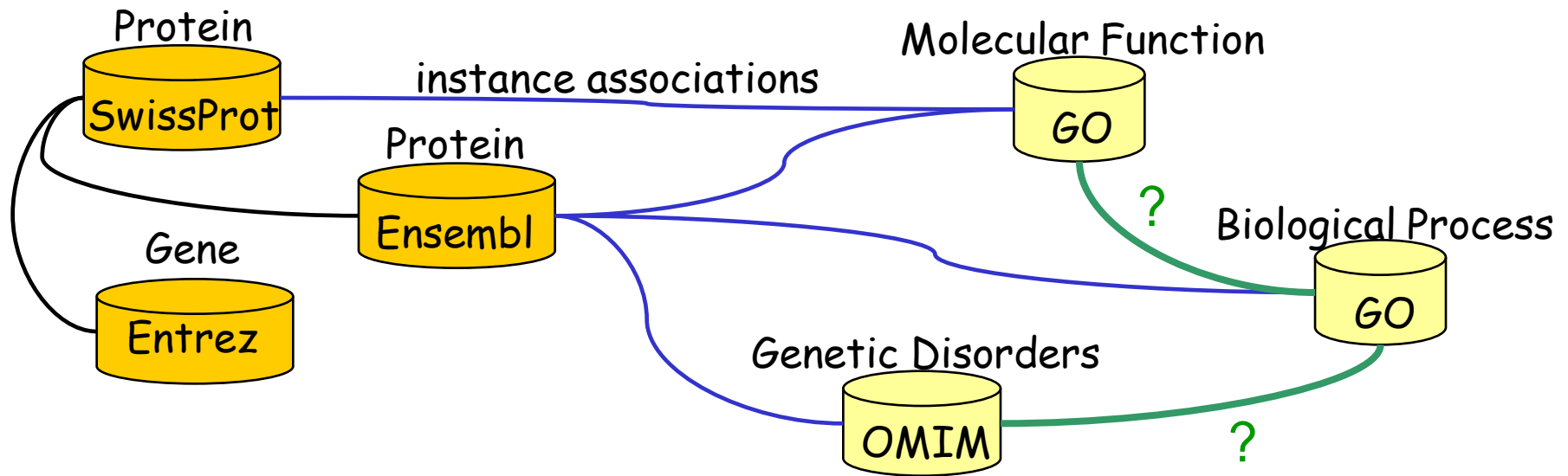
**T. Kirsten, A. Thor, E. Rahm**

**University of Leipzig, Germany**

***www.izbi.de, dbs.uni-leipzig.de***

# Motivation

- Increasing number of connected sources and ontologies



- **Ontology matching**

- Goal: Find semantically related concepts
- Output: Set of correspondences (ontology mapping)
- Use: Validation (curation) and recommendation of instance associations

# Metadata-based Match Approaches

---

- Metadata: Concept names, descriptions, ontology structure, ...
- Match mainly focused on syntax and structure
- Limited use of domain knowledge
- Highly similar names with opposite semantics, e.g., ion vs. anion, organic vs. inorganic

		$Sim_{2-Gram}$
ion transporter	- anion transport	: 0.77
ion transporter activity	- ion transport	: 0.66

# Outline

---

- Motivation
- Instance-based Match Approach
- Similarity and Evaluation Metrics
- Selected Match Results
- Conclusions

# Our Instance-based Match Approach

---

- Approach
  - Use of available domain-specific knowledge
  - Exploitation of instance associations to create ontology mappings
- Key idea: "Two concepts are related if they share a significant number of associated objects"
- Flexible and extensible approach
  - Instance associations of pre-selected sources
  - Different metrics to determine the instance-based similarity
  - Combination of different ontology mappings

# Instance-based Matching

## Molecular Function (MF)

- ...
- GO:0005215  
Transporter activity
- ...
- GO:0015075  
Ion transporter activity**
- ...
- GO:0008504  
Anion transporter activity
- ...
- GO:0008514  
Organic anion transporter activity
- GO:0015103  
Inorganic anion transporter activity

Correspondence  
creation using  
shared & directly  
associated instances

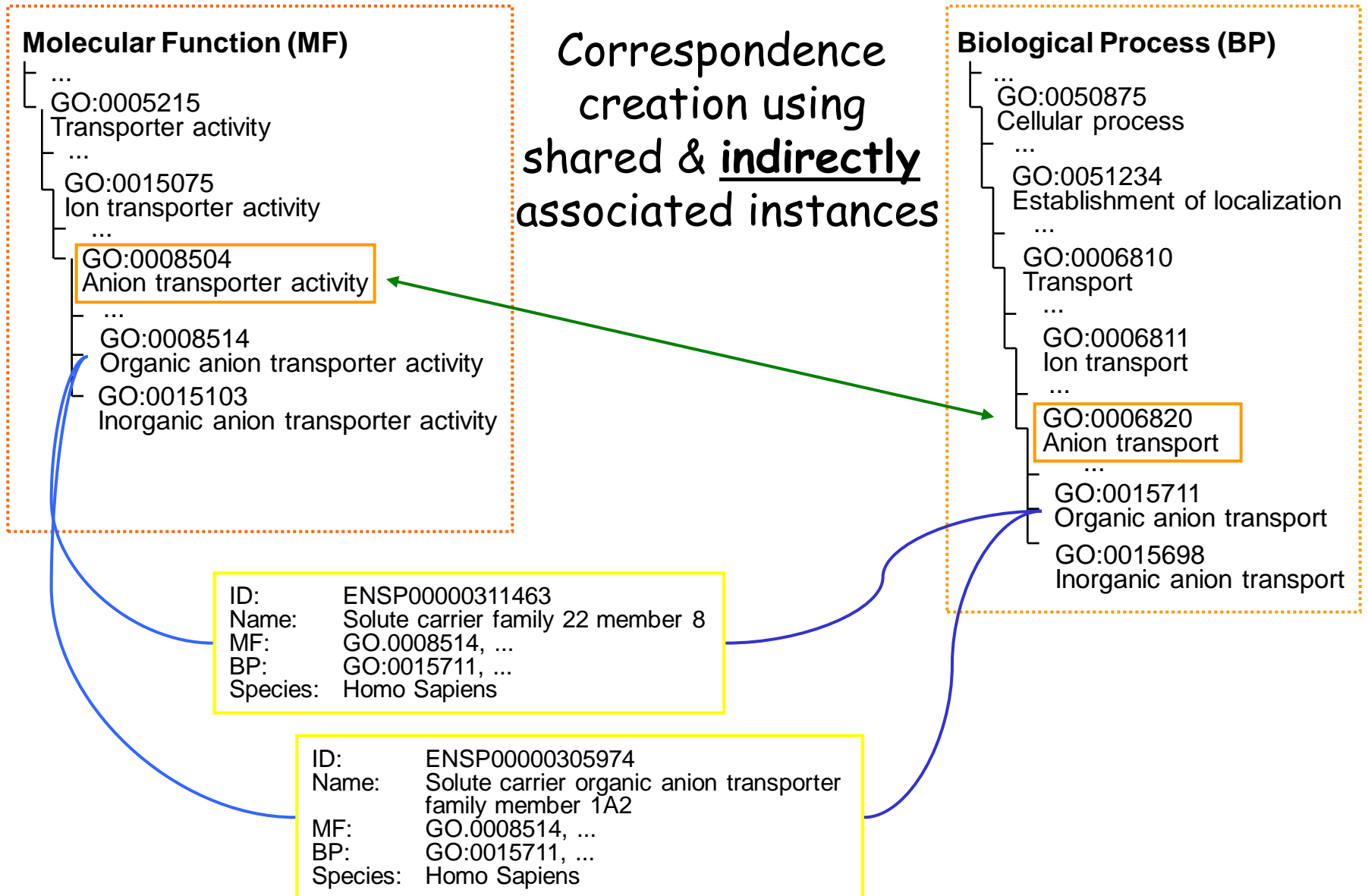
## Biological Process (BP)

- ...
- GO:0050875  
Cellular process
- ...
- GO:0051234  
Establishment of localization
- ...
- GO:0006810  
Transport
- ...
- GO:0006811  
Ion transport**
- ...
- GO:0006820  
Anion transport
- ...
- GO:0015711  
Organic anion transport
- GO:0015698  
Inorganic anion transport

ID: ENSP00000355930  
Name: Solute carrier family 22 member 1 isoform a  
MF: GO.0015075, ...  
BP: GO:0006811, ...  
Species: Homo Sapiens

ID: ENSP00000325240  
Name: LIM and SHB domain protein 1  
MF: GO.0015075, ...  
BP: GO:0006811, ...  
Species: Homo Sapiens

# Instance-based Matching cont.



# Similarity Metrics

- Baseline similarity  $\text{Sim}_{\text{Base}}$

$$\text{Sim}_{\text{Base}}(c_1, c_2) = \begin{cases} 1 & , \text{ if } N_{c_1 c_2} > 0 \\ 0 & , \text{ if } N_{c_1 c_2} = 0 \end{cases}$$

- Dice similarity  $\text{Sim}_{\text{Dice}}$

$$\text{Sim}_{\text{Dice}}(c_1, c_2) = \frac{2 \cdot N_{c_1 c_2}}{N_{c_1} + N_{c_2}}$$

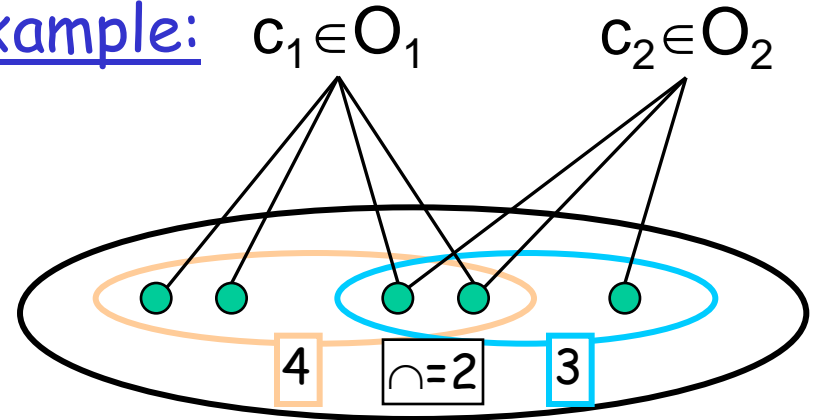
- Minimum similarity  $\text{Sim}_{\text{Min}}$

$$\text{Sim}_{\text{Min}}(c_1, c_2) = \frac{N_{c_1 c_2}}{\min(N_{c_1}, N_{c_2})}$$

- Kappa similarity  $\text{Sim}_{\text{Kappa}}$

Please, see the paper for details

Example:



$$\begin{aligned} \text{Sim}_{\text{Base}} &= 1 \\ \text{Sim}_{\text{Dice}} &= 2 \cdot 2 / (4 + 3) = 0.57 \\ \text{Sim}_{\text{Min}} &= 2 / 3 = 0.67 \end{aligned}$$

$$\begin{aligned} 0 \leq \text{Sim}_{\text{Dice}} \leq \text{Sim}_{\text{Min}} \leq \text{Sim}_{\text{Base}} \leq 1 \\ 0 \leq \text{Sim}_{\text{Kappa}} \leq \text{Sim}_{\text{Base}} \leq 1 \end{aligned}$$

# Evaluation Metrics

- Computation of precision & recall needs a perfect mapping
  - Laborious for large ontologies
  - Might not be well-defined
- Metric *Match Coverage* to approximate "recall"
  - Idea: Measuring the fraction of matched concepts

$$\text{MatchCoverage}_{o_1} = \frac{|C_{o_1\text{-Match}}|}{|C_{o_1}|} \in [0 \dots 1] \quad \text{Combined InstMatchCoverage} = \frac{|C_{o_1\text{-Match}}| + |C_{o_2\text{-Match}}|}{|C_{o_1\text{-Inst}}| + |C_{o_2\text{-Inst}}|} \in [0 \dots 1]$$

- Metric *Match Ratio* to approximate "precision"
  - Idea: Measuring the average number of match counterparts per matched concept

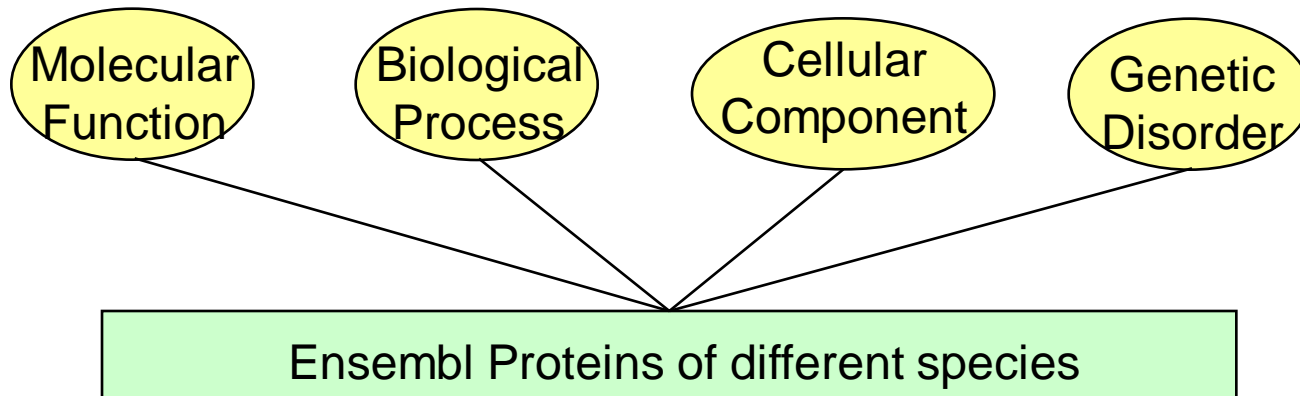
$$\text{MatchRatio}_{o_1} = \frac{|Corr_{o_1-o_2}|}{|C_{o_1\text{-Match}}|} \geq 1 \quad \text{Combined MatchRatio} = \frac{2 \cdot |Corr_{o_1-o_2}|}{|C_{o_1\text{-Match}}| + |C_{o_2\text{-Match}}|} \geq 1$$

- Goal: High Match Coverage with low Match Ratios

# Match Scenario

---

- Ontologies
  - Subontologies of GeneOntology: Mol. function, biol. processes and cell. components
  - Genetic disorders of OMIM
- Instances: Ensembl proteins of different species, i.e., homo sapiens, mus musculus, rattus norvegicus



# Exhaustive Match Study

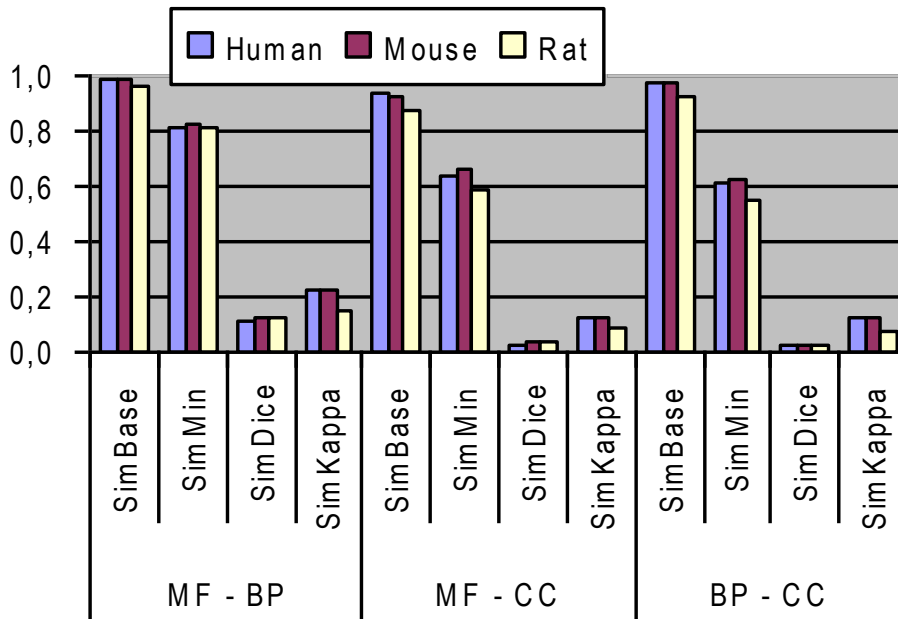
---

- Instance-based matching
  - Direct protein associations of human, mouse, rat
  - Study of match combinations: Union, intersection
  - Generation and utilization of indirect associations
- Metadata-based matching
  - First & simple approach: Utilization of concept names
  - Trigram string similarity; different thresholds
- Comparison of instance- and metadata-based match results

# Match Results: Direct Associations

- Sim<sub>Base</sub>: High Coverage (99%), moderate to high Match Ratios
- Sim<sub>Dice</sub> & Sim<sub>Kappa</sub>: Very restrictive (Coverage < 20%) but low Match Ratios
- Sim<sub>Min</sub>: High Coverage (60%-80%) with high number of covered concepts but significantly lower Match Ratios than Sim<sub>Base</sub>

Combined Instance Coverage



Match Ratios per ontology

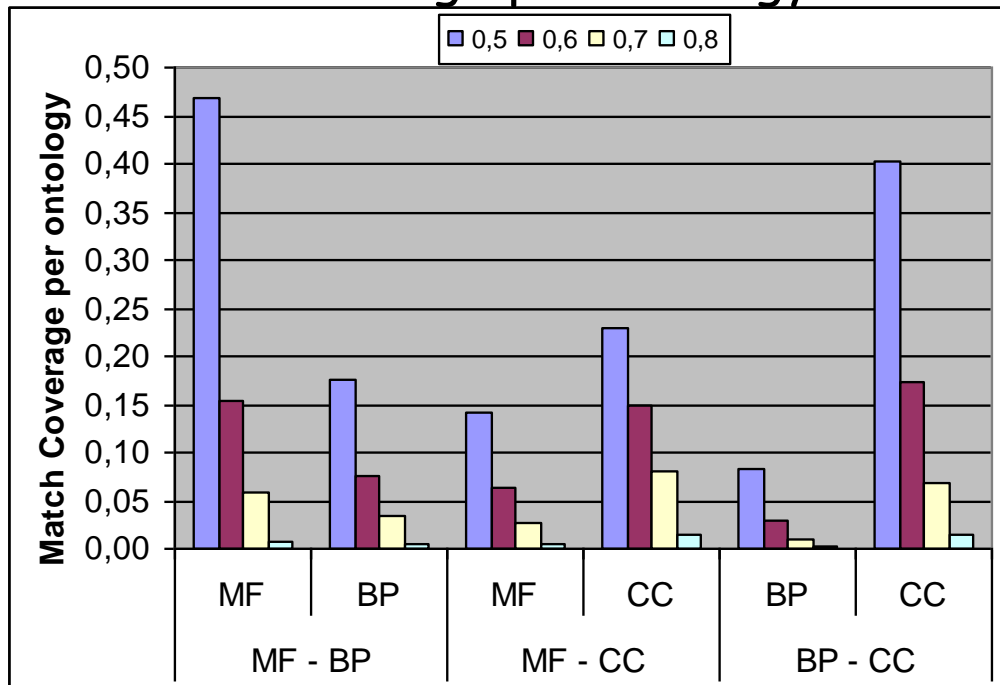
	MF - BP		MF - CC		BP - CC	
	MF	BP	MF	CC	BP	CC
Base	20.4	17.0	7.6	28.6	9.8	46.3
Min	4.4	4.0	2.2	7.8	2.4	8.6
Dice	1.3	1.2	1.0	1.3	1.0	1.3
Kappa	2.0	2.0	1.9	2.7	1.7	2.6

(Match Ratios for Homo Sapiens)

# Match Results: Metadata-based Matching

- Growing Coverage and Match Ratios for lower thresholds
- No correspondences with a similarity  $\geq 0.9$
- Moderate to low Match Ratios
- Inclusion of false positives for low thresholds, e.g. 0.5

Match Coverage per ontology



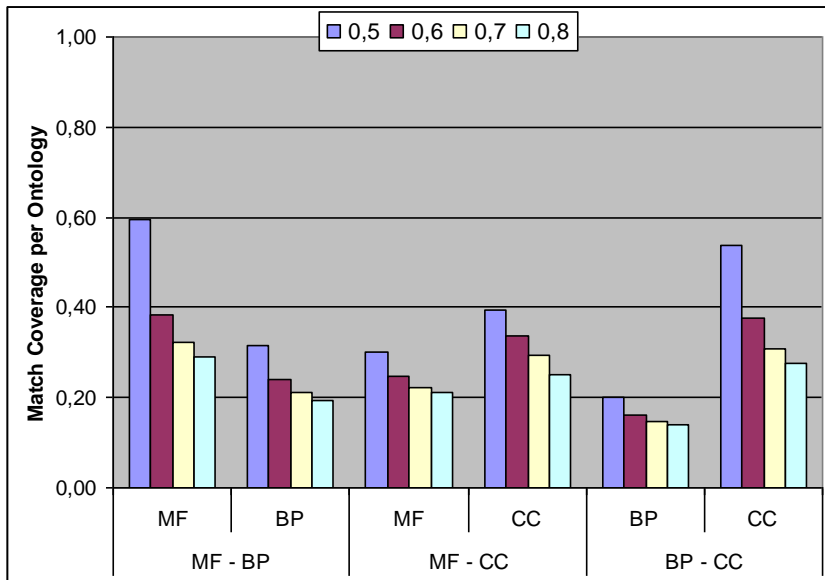
Match Ratios per ontology

	MF - BP		MF - CC		BP - CC	
	MF	BP	MF	CC	BP	CC
0.5	4.4	6.9	2.5	6.3	2.5	3.4
0.6	2.4	2.9	2.7	4.6	1.7	2.0
0.7	1.4	1.4	1.1	1.5	1.4	1.4
0.8	1.1	1.1	1.1	1.2	1.1	1.2

# Match Results: Match Combinations

- Combinations between instance- ( $Sim_{Min}$ ) and metadata-based match approach
- Union: Increased coverage, higher influence of  $Sim_{Min}$  for increased thresholds of the metadata-based matcher
- Intersection: Low Match Coverage (<1%) and Match Ratios

Match Coverage per ontology for unified mappings



Match Ratios per ontology (threshold 0.7)

	MF - BP		MF - CC		BP - CC	
	MF	BP	MF	CC	BP	CC
$\cup$	4.1	3.7	2.2	6.7	2.4	7.6
$\cap$	1.0	1.0	1.0	1.0	1.0	1.3

( $Sim_{Min} = 1.0$ , Homo Sapiens)

# Summarized Match Results

---

- Similarity metrics
  - $Sim_{Base}$ : Maximal # correspondences, high Match Coverage but also high Match Ratios
  - $Sim_{Dice}$  &  $Sim_{Kappa}$ : Very restrictive, Match Ratio near 1
  - $Sim_{Min}$ : High Match Coverage, moderate Match Ratios
- Indirect protein associations
  - Increased # associated proteins (factor 3)
  - Improved Match Coverage but also high Match Ratios
- Metadata-based matching
  - Low # correspondences even for threshold  $\geq 0.7$
  - Inclusion of false positives by applying lower thresholds
  - Low overlap between instance- and metadata-based mappings
  - More sophisticated metadata matcher could help

# Conclusions & Future Work

---

- Instance-based matching to map large ontologies in life sciences
  - Flexible and extensible match approach
  - Exhaustive match evaluation as a first step
    - Different similarity metrics for mapping generation
    - Comparison with a simple metadata-based approach
- Future work:
  - More experiments: Source- and species-specific comparisons and combinations
  - Mapping validation by user feedback

## Further Information

---

<http://dbs.uni-leipzig.de>

<http://www.izbi.de>