600 JAHRE
UNIVERSITÄT LEIPZIG

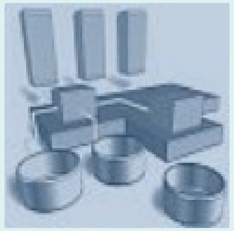
Quality of Functional Annotations in Life Science Data Sources

Anika Groß

IZBI UNIVERSITÄT LEIPZIG

Interdisciplinary Centre for Bioinformatics
http://www.izbi.uni-leipzig.de

Database Group Leipzig
http://dbs.uni-leipzig.de

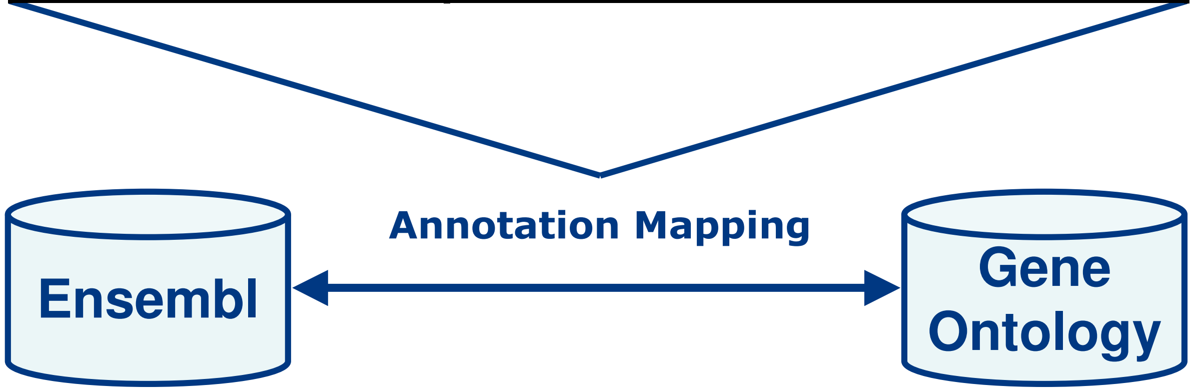


Leipzig, 25th November 2009

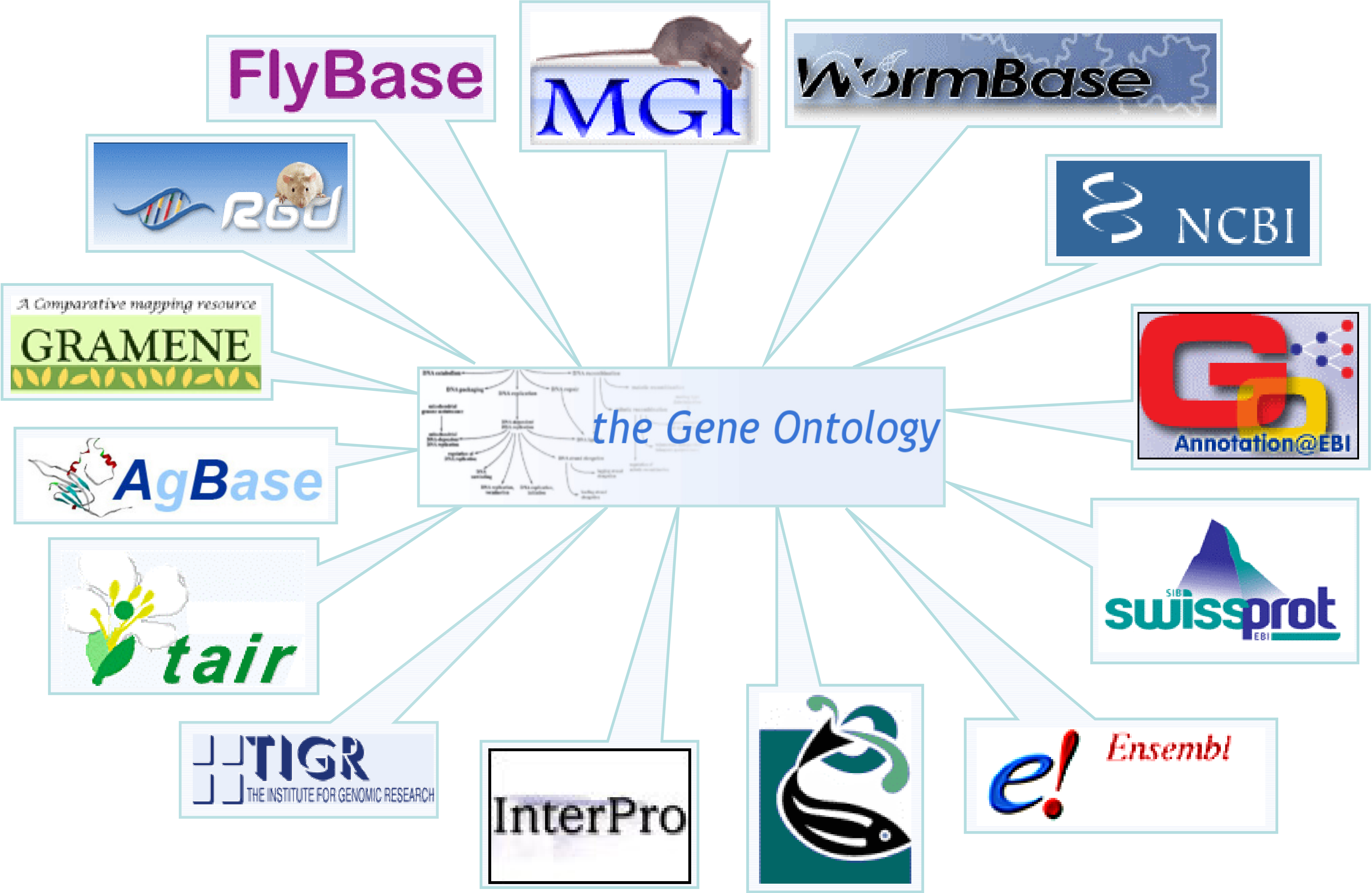
Functional Annotations in Life Sciences

- Increasing use of ontologies in life sciences, mainly ontology-based annotations
- **Functional annotations**
  Semantic and uniform descriptions of properties of biological objects, e.g., a protein is involved in a specific biological process

Annotation	
Ensembl ID	**Gene Ontology Concept ID**
ENSP00000344151	GO:0015808 (L-alanine transport)
ENSP00000230480	GO:0005615 (extracellular space)
ENSP00000352999	GO:0006915 (apoptosis)

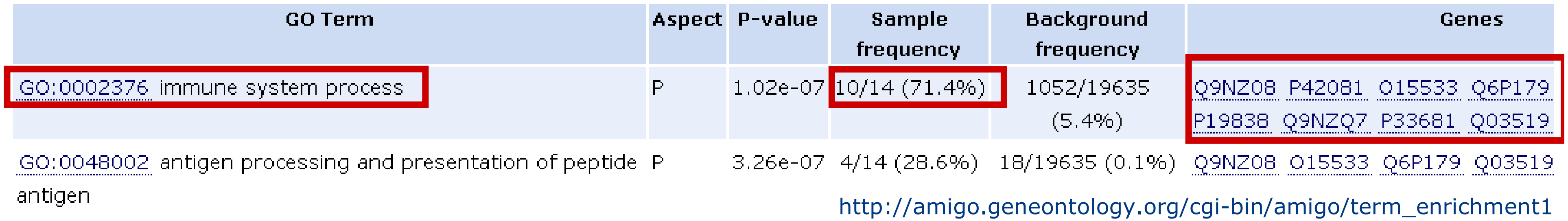


Usage of Gene Ontology (GO)



Application of GO Annotations

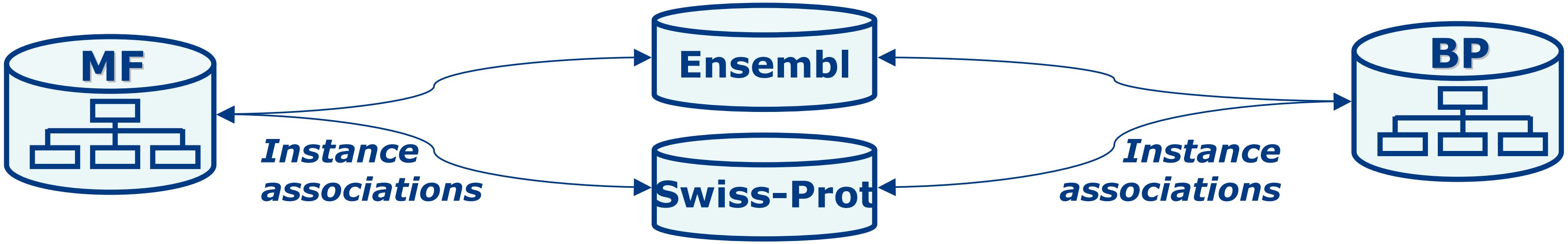
- Functional profiling of large data sets (e.g., gene expression microarrays) to find significantly overrepresented GO terms
  - FUNC*, Term Enrichment Tool (Amigo), …



GO Term	Aspect	P-value	Sample frequency	Background frequency	Genes
GO:0002376 immune system process	P	1.02e-07	10/14 (71.4%)	1052/19635 (5.4%)	Q9NZ08 P42081 O15533 Q6P179 P19838 Q9NZQ7 P33681 Q03519
GO:0048002 antigen processing and presentation of peptide antigen	P	3.26e-07	4/14 (28.6%)	18/19635 (0.1%)	Q9NZ08 O15533 Q6P179 Q03519

http://amigo.geneontology.org/cgi-bin/amigo/term_enrichment1

- Instanced-based ontology matching



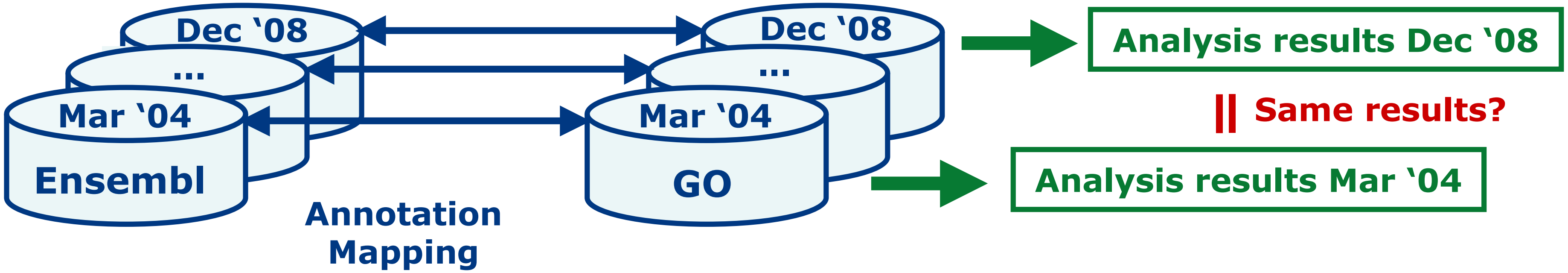
* Prüfer, K. et al: FUNC: a package for detecting significant associations between gene sets and ontological annotations, BMC Bioinformatics, 2007

Motivation

- Computed results of such applications depend on the **quality** of the underlying **functional annotations**

→ (Garbage In/Garbage Out principle)

- Domain knowledge changes
  - New findings, addition and revision of knowledge
  - Result: modification of data sources (evolution)



Example – Changing Annotations

Annotation		Provenance				
Ensembl ID	**Gene Ontology Concept ID**	v_{48}	v_{49}	v_{50}	v_{51}	v_{52}
ENSP00000344151	GO:0015808 (L-alanine transport)	🟩	🟩	🟩	🟩	🟩
ENSP00000230480	GO:0005615 (extracellular space)	🟨	🟨	🟩	🟨	🟥
ENSP00000352999	GO:0006915 (apoptosis)	🟩	-	-	-	🟩

🟩 experimentally verified 🟨 author statement 🟥 automatically annotated

Dec 2007 – Dec 2008

- Evolution of annotations
  - varying provenance
  - absence/presence of annotations
- Major changes in annotation mappings may substantially influence or even invalidate earlier findings

Quality of Annotations

Quality criteria

- Correctness
- Completeness
- **Provenance**
- **Stability**
- …

Which annotations fit best for my application?

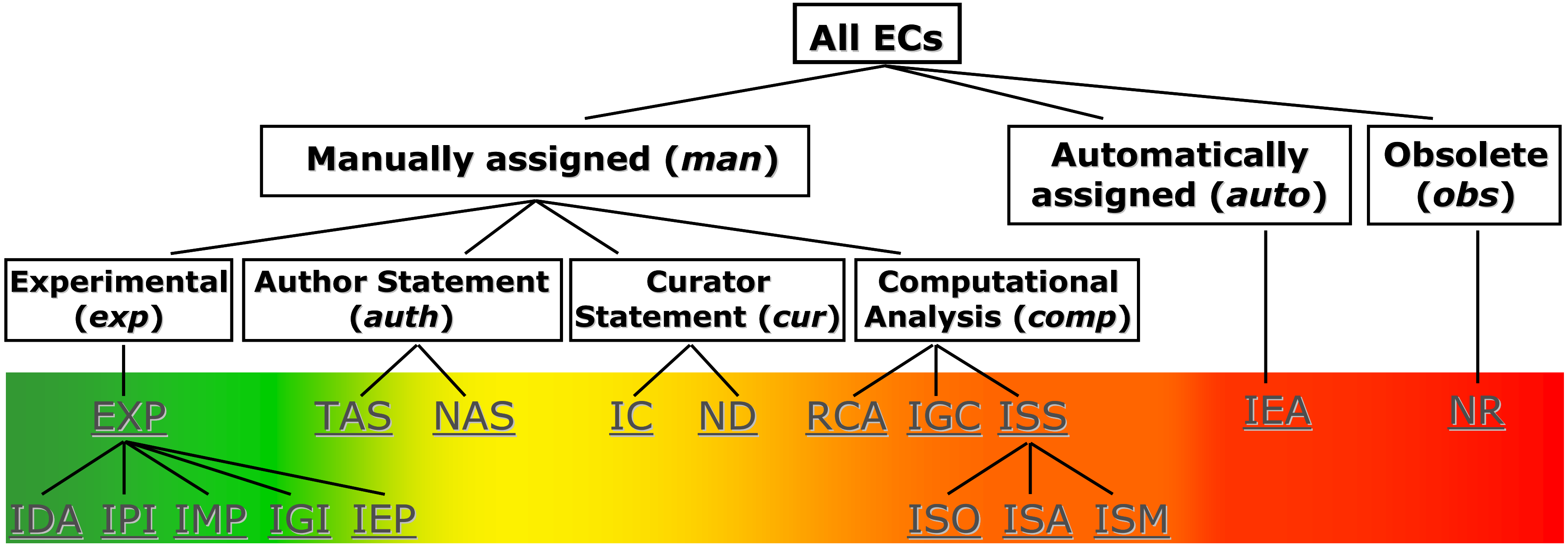
How many high-quality annotations are available in a source?

How was the annotation created?

How reliable is the annotation?

Provenance of Functional Annotations

- Annotations can be generated by different creation methods → have different provenance
- Evidence Code (EC) * = indicates how the annotation to a particular term has been derived, e.g., by which type of experiment or analysis


All ECs
Manually assigned (*man*)
Automatically assigned (*auto*)
Obsolete (*obs*)
Experimental (*exp*)
Author Statement (*auth*)
Curator Statement (*cur*)
Computational Analysis (*comp*)
EXP
TAS
NAS
IC
ND
RCA
IGC
ISS
IEA
NR
IDA
IPI
IMP
IGI
IEP
ISO
ISA
ISM

* http://www.geneontology.org/GO.evidence

- ➢ Gives information how biologically founded or reliable an annotation is

First Step: Comparative Analysis

- Analysis of annotation evolution *
  - Trend chart
  - Provenance Changes
  - …

- Two large life science sources (Mar 2004 – Dec 2008)
- GO Annotations for human proteins

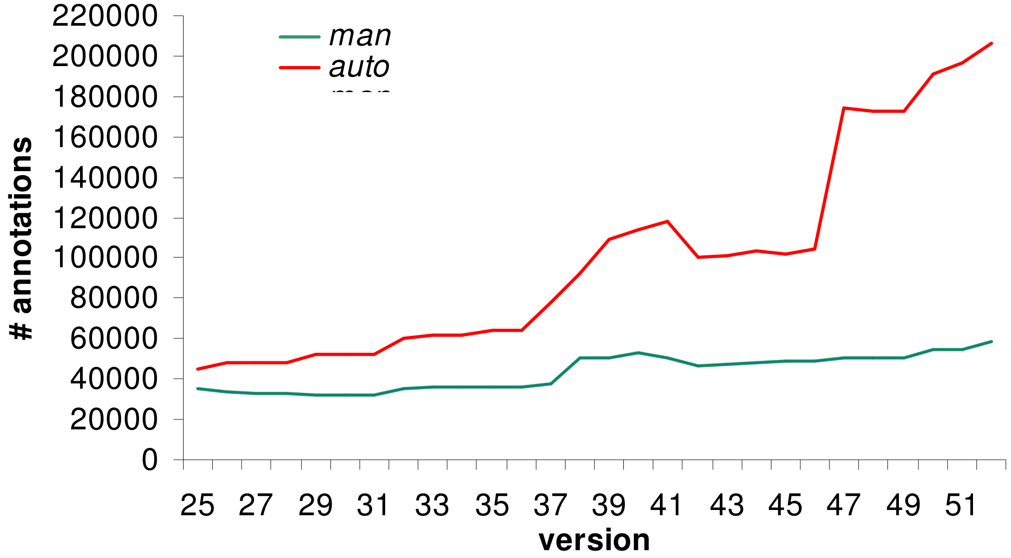
Ensembl $v_{31}–v_{52}$

Swiss-Prot $v_{47}–v_{56}$

* Groß, A.; Hartung, M.; Kirsten, T.; Rahm, E.: Estimating the Quality of Ontology-based Annotations by Considering Evolutionary Changes, Proc. DILS, 2009

Analysis Results

Trend chart



- Manually assigned vs. automatically assigned
- 78% (22%) of 265,000 *auto* (*man*) assigned annotations
- growth$_{auto}$ 4.6
- v_{40} – v_{42} considerable number of deletions

Provenance changes

Annotations that changed **from** one provenance type **to** another

from / to	*exp*	*auth*	*cur*	*comp*	*auto*	*obs*	Sum	
exp	896	413	11	1,259	2,966	3	*5,548*	13%
auth	1592	798	73	1,038	11,901	23	*15,425*	35%
cur	21	27	0	16	182	0	*246*	1%
comp	1,280	1,206	26	0	3,101	0	*5,613*	13%
auto	3,311	10,169	228	2,329	0	116	*16,153*	37%
obs	79	391	9	12	725	0	*1,216*	3%
Sum	*7,179* 16%	*13,004* 29%	*347* 1%	*4,654* 11%	*18,875* 43%	*142* 0%	*44,201*	

- EC changes predominantly between auth and auto (in both directions)
- No obvious trend for the rest
- Due to vast amount of auto annotations

Second Step: Assessing Annotation Quality

Idea: Assessing the quality of annotations based on their history and occurred changes (stability)

Aim: Filtering annotations w.r.t. different quality criteria

Stability Measures

- Existence stability

a_{age} age of annotation (in #versions)
$a_{present}$ presence within a_{age}

$$stab_{exis}(a) = a_{present} / a_{age}$$

- Quality stability

$a_{changed}$ # provenance changes
$a_{unchanged}$ # unchanged provenance

$$stab_{qual}(a) = a_{unchanged} / (a_{unchanged}+a_{changed})$$

v_0	v_1	v_2	v_3	v_4	a_{age}	$stab_{exis}$	$stab_{qual}$	$stab_{comb}$
q_1	q_1	q_1	q_1	(i_1,c_1,q_1)	5	5/5=1	4/(4+0)=1	1
q_1			q_1	(i_2,c_2,q_1)	5	3/5=0.6	2/(2+0)=1	0.6
	q_2	q_2	q_1	(i_3,c_3,q_3)	4	4/4=1	1/(1+2)=0.33	0.33

=min ($stab_{exis}$, $stab_{qual}$)

Ensembl Annotations Classified by Stability

	$\lvert\text{stab}_{\text{exis}}\rvert$	$\lvert\text{stab}_{\text{qual}}\rvert$	$\lvert\text{stab}_{\text{comb}}\rvert$
exp	21,659	20,486	20,122
	650	1,880	2,187
auth	29,157	26,862	26,067
	1,033	3,116	4,123
cur	462	399	393
	15	78	84
comp	3,127	2,409	2,317
	205	1,078	1,015
auto	183,127	201,968	179,490
	23,210	4,369	26,847
sum	**237,532**	**252,124**	**228,389**
	25,113	**10,521**	**34,256**

stable	$\text{stab} \geq 0.9$
unstable	$\text{stab} < 0.9$

High share of temporal absence

13% unstable, mainly *auto* (80%) and some *auth* (12%)

Use – Putting different criteria together

Protein ID	GO Concept ID	Provenance	Age in Years	stab_{exis}	stab_{qual}	stab_{comb}
ENSP00000344151	GO:0015808 (L-alanine transport)	*exp*	3	1	1	1
ENSP00000230480	GO:0005615 (extracellular space)	*auto*	2.5	1	**0.462**	0.462
ENSP00000352999	GO:0006915 (apoptosis)	*exp*	3	**0.824**	1	0.824

- Different criteria to assess the quality of annotations w.r.t. provenance, stability, …
- Users/Applications can filter less/more reliable annotations (e.g. stable, old, manually assigned)

How was the annotation created?

How reliable is the annotation?

Use – Putting different criteria together

- Stable, old, manually assigned:

  In Ensembl about 30,000 (11%)

How many high-quality annotations are available in a source?

- Criteria selection is highly dependent on application!
- Annotation instability is not necessarily a negative aspect
- Alternative interpretation

  novel or unstable annotations (in Ensembl 96,000; 37%) are of special research interest / significant new biological findings

Which annotations fit best for my application?

Conclusion and Future Work

- Generic approach to estimate the quality of ontology-based, functional annotations by taking their evolution history and provenance into account
- Evaluation in two large life science sources
  - Instabilities for *auth* or *auto* annotations
- Different quality criteria: provenance, stability, age to classify annotations
  - Users/applications can filter annotations
- Investigate other quality aspects
- Explore the impact of unstable annotations on dependent applications (e.g., FUNC, instance-based ontology matching)

Thank you for your attention!

http://dbs.uni-leipzig.de

http://www.izbi.de

cellular_component
macromolecular complex
intracellular
membrane
organelle
synapse
cell part
protein complex
synapse part
cell fraction
membrane-enclosed lumen
ribonucleoprotein complex
cell projection
proteasome complex (sensu Eukaryota)
membrane attack complex
organelle part
cytoplasm
membrane fraction
membrane-bound organelle
integral to membrane
intracellular organelle
nuclear lumen
extracellular region
plasma membrane
cytosol
contractile fiber
intracellular non-membrane-bound organelle
ribosome
integral to plasma membrane
organelle membrane
proteinaceous extracellular matrix
chromatin
organelle envelope
nucleolus
intercellular junction
membrane coat
extracellular space
chromosome
cytoplasmic vesicle
intracellular membrane-bound organelle
large ribosomal subunit
endoplasmic reticulum membrane
vacuole
small ribosomal subunit
cytoskeleton
proton-transporting two-sector ATPase complex
endoplasmic reticulum
vesicle membrane
cytoplasmic membrane-bound vesicle
cytosolic ribosome (sensu Eukaryota)
Golgi apparatus
mitochondrion
proton-transporting ATP synthase complex
actin cytoskeleton
microtubule
nuclear envelope
mitochondrial envelope
cytosolic large ribosomal subunit (sensu Eukaryota)
microtubule cytoskeleton
cytoplasmic vesicle membrane
coated vesicle
nucleus
mitochondrial membrane
microtubule associated complex
coated vesicle membrane
clathrin-coated vesicle
mitochondrial matrix
cytosolic small ribosomal subunit (sensu Eukaryota)
mitochondrial inner membrane
nucleoplasm
mitochondrial respiratory chain
DNA-directed RNA polymerase II, holoenzyme
transcription factor complex