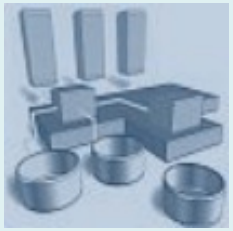
600 JAHRE
UNIVERSITÄT LEIPZIG

Estimating the Quality of Ontology-Based Annotations by Considering Evolutionary Changes

Anika Gross, Michael Hartung, Toralf Kirsten, Erhard Rahm

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

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



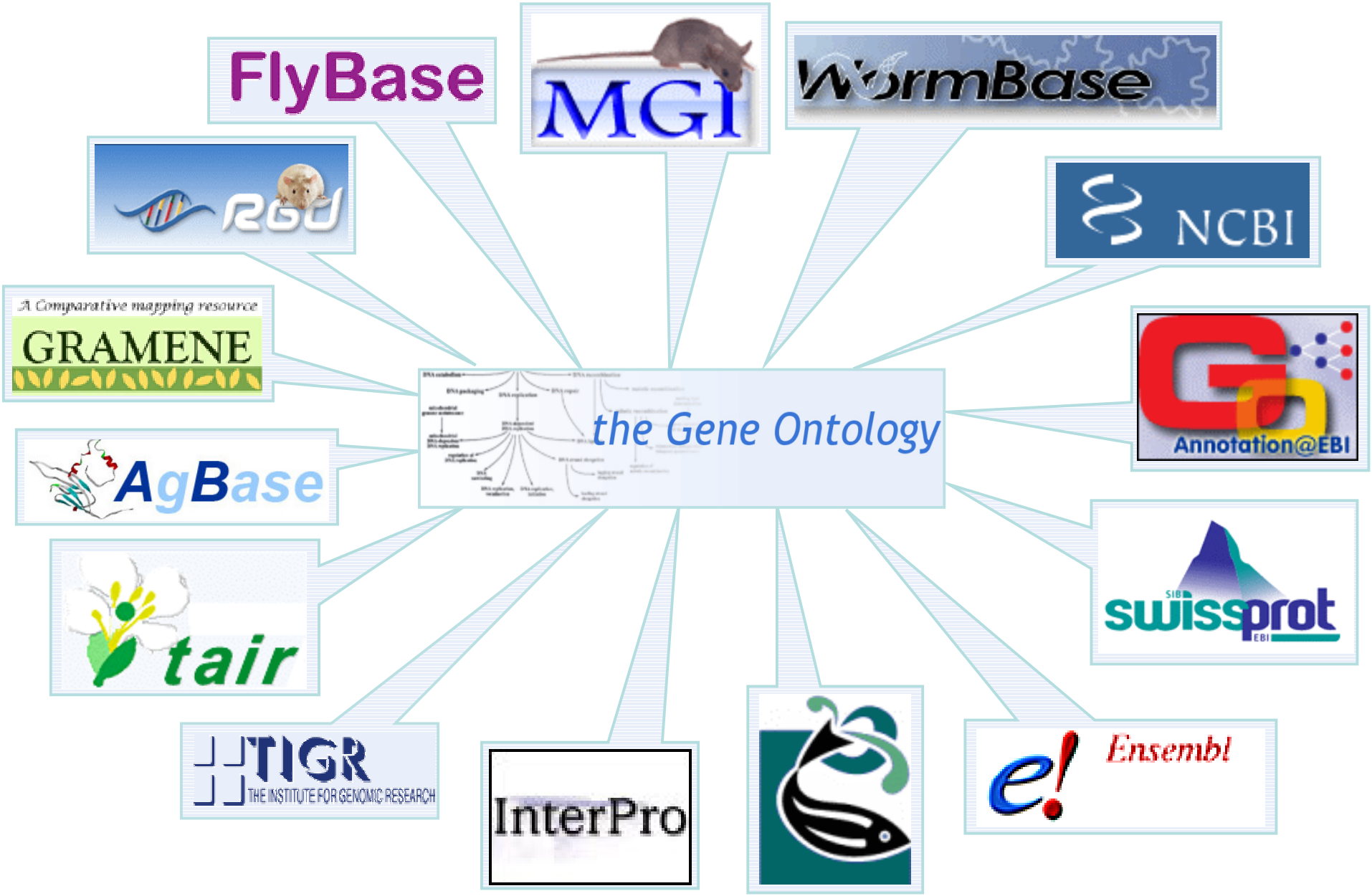
Manchester, 21st July 2009

Annotations in Life Sciences

- Increasing use of ontologies in life sciences, mainly ontology-based annotations
- **Annotations** Semantic descriptions of properties of biological objects, e.g., a protein is associated to a specific biological process

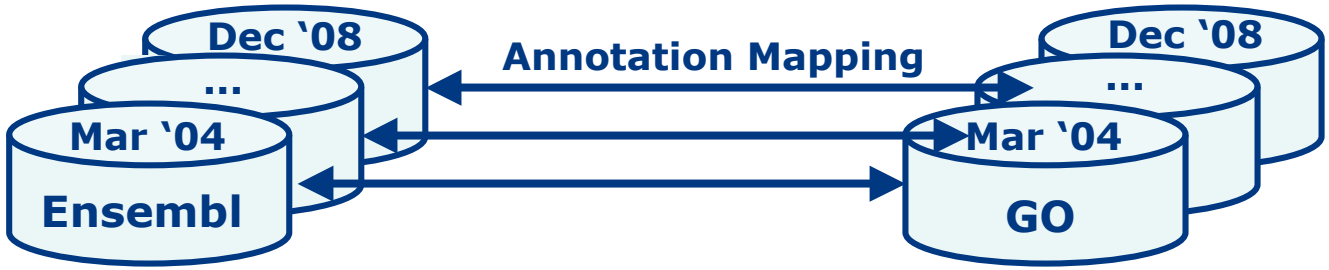
Annotation	
Ensembl ID	**Gene Ontology Concept ID**
ENSP00000344151	GO:0015808 (L-alanine transport)
ENSP00000230480	GO:0005615 (extracellular space)
ENSP00000352999	GO:0006915 (apoptosis)

Usage of Gene Ontology (GO)



Motivation

- Domain knowledge changes
  - New findings, addition and revision of knowledge
  - Result: evolution of data sources
- First analysis at DILS 2008*
  - Focused on evolution in ontologies and protein sources
- What about 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

- Different stability of annotations due to different evolution and provenance changes

* Hartung, M.; Kirsten, T.; Rahm, E.: Analyzing the Evolution of Life Science Ontologies and Mappings, Proc. of 5th DILS, 2008

Application of GO Annotations

- Functional profiling of large data sets (e.g. gene expression microarrays) to find significantly shared GO terms

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

- Unstable input annotations

↓

- Impact on application results (Garbage In/Garbage Out principle)

Quality of Annotations

Possible criteria

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

Contributions

- General approach to analyze annotation mappings considering evolutionary changes
- Evolution-aware annotation model, change operations
- Evolution-based quality measures to identify reliable annotations (stability, provenance)
- Comparative evaluation in two large life science sources (Ensembl, Swiss-Prot)

Overview

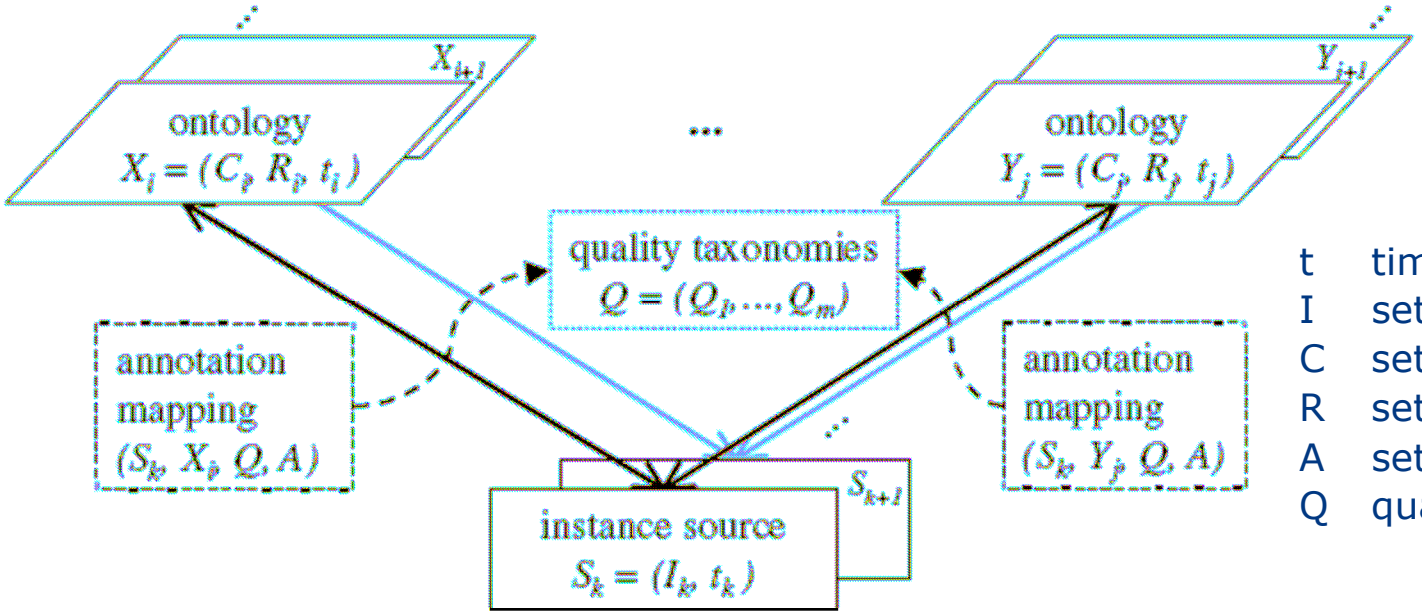
- *Motivation*
- *Annotations and quality*

- Annotation model, change operations
  - Quantitative evaluation of annotation evolution

- Estimating annotation quality
  - Stability measures
  - Evaluation results

- Conclusion

Annotation Model

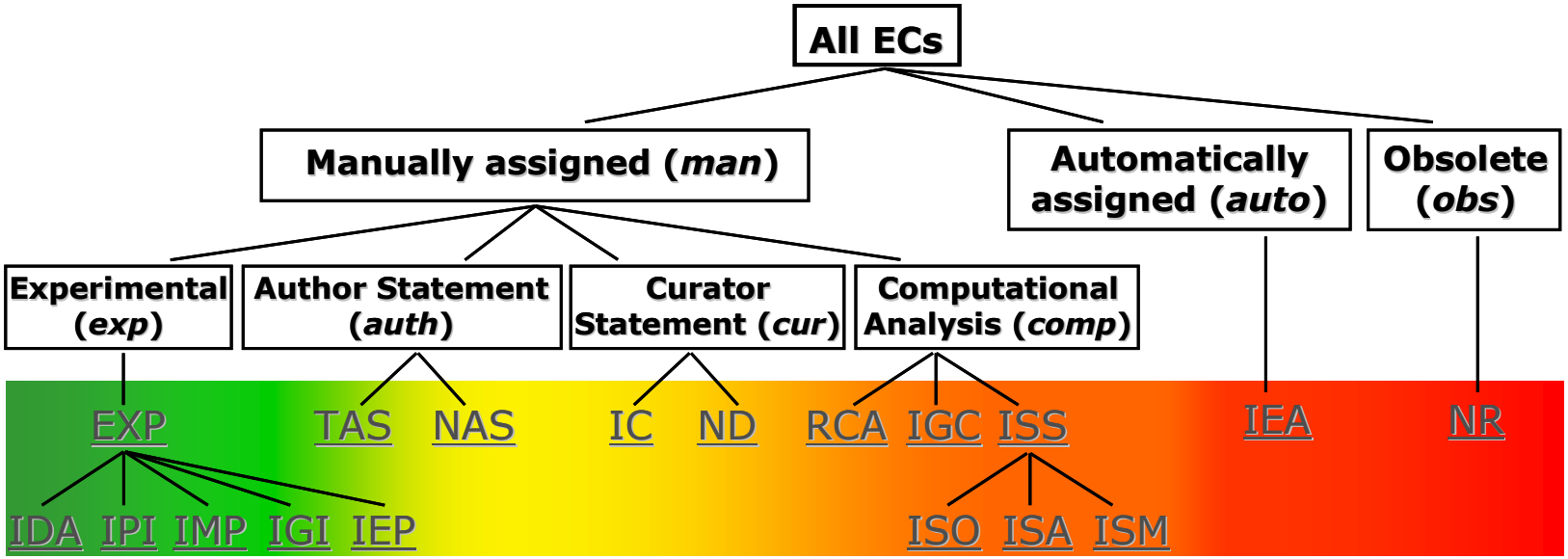


t	timestamp
I	set of instances
C	set of concepts
R	set of relationships
A	set of correspondences
Q	quality taxonomies

- Annotation $a \in A$, $a=(i,c,\{q\})$
- Linear versioning scheme
- Refers to quality taxonomies (e.g., provenance)

Provenance Taxonomy - Evidence Codes

- Evidence Code (EC) * = indicates how the annotation to a particular term has been derived, e.g., by which type of experiment or analysis



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

- Other taxonomies are possible
  - age
  - stability
  - ...

Evolution Operations

- Evolution operations for proteins, ontologies, annotations
  - Add
  - Delete
  - Change
- Propagation of evolutionary changes when ontologies and proteins evolve!

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

Change

Delete

Add

Dec 2007 – Dec 2008

Quantitative Evolution Analysis

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

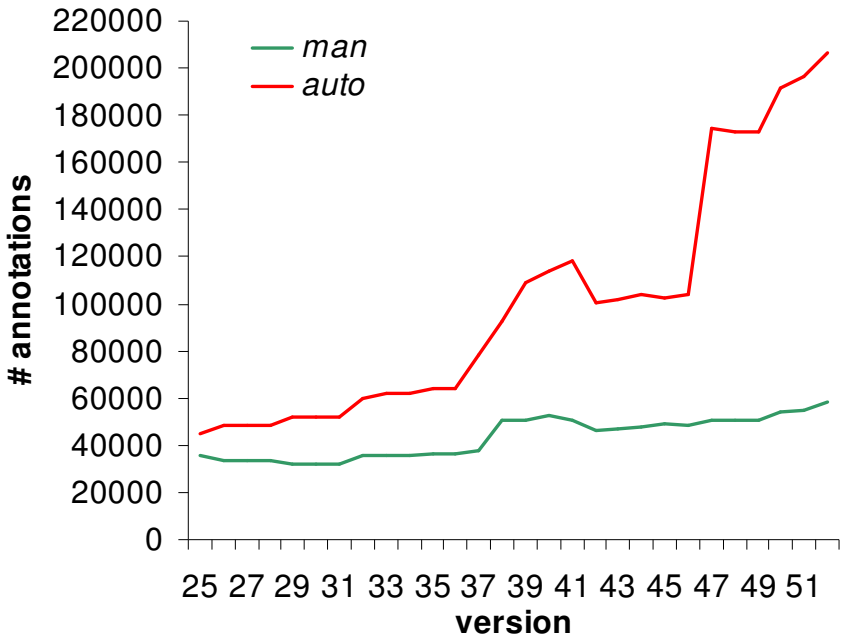
Ensembl v_{31}–v_{52}

Swiss-Prot v_{47}–v_{56}

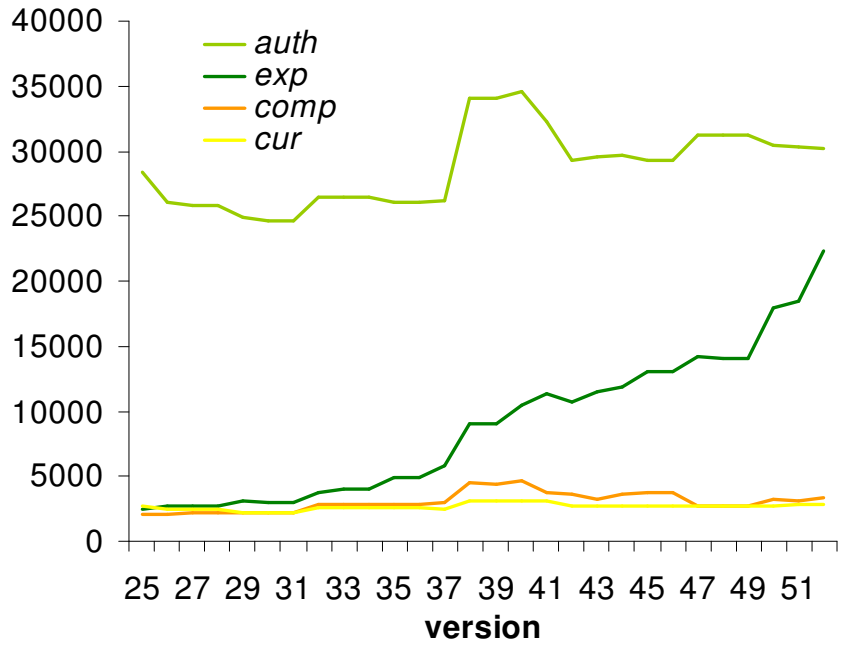
- Analysis of annotation evolution w.r.t.
  - Different provenance
  - Change operations

Evolution of Annotations

Manually assigned
vs. automatically assigned



"Subclasses"
of manually assigned



- 78% of 265,000 automatically assigned
- $growth_{auto}$ 4.6
- $v_{40} - v_{42}$ considerable number of deletions

- 22% manually assigned
- $growth_{man}$ 1.7
  - $growth_{exp}$ 8.9
  - $growth_{auth}$ 1.1

Evolution Operations

	Add		Change		Delete	
exp	25,979	6.6%	5,826	12.2%	7,575	3.6%
auth	34,046	8.7%	**16,381**	**34.3%**	29,148	14.0%
cur	6,362	1.6%	300	0.6%	6,318	3.0%
comp	6,734	1.7%	5,720	12.0%	4,362	2.1%
auto	**316,979**	**80.9%**	**18,344**	**38.4%**	**157,632**	**75.6%**
obs	1,826	0.5%	1,234	2.6%	3,550	1.7%
sum	**391,926 (60%)**		**47,805 (8%)**		**208,585 (32%)**	

>80% of all additions are *auto* annotations

Changes (8%) mainly *auth* and *auto*

Deletions (32%) mainly *auto* annotations

- **Instabilities for *auth* and *auto***

Provenance Changes

- How many annotations 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	13,004	347	4,654	18,875	142	44,201	
	16%	29%	1%	11%	43%	0%		

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

Assessing Annotation Quality

Seen so far

- Most annotation changes are additions of new annotations
- Also many deletions and changes
- Instabilities for *auto* and *auth* annotations

Idea

- Assessing the quality of annotations based on their <u>history</u> and occurred <u>changes</u> (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

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

- Combined stability

$$stab_{comb}(a) = min\ (\ stab_{qual}(a),\ stab_{exis}(a)\ )$$

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

Evaluation Scenario

- All currently available annotations for human proteins within the last three years

Ensembl v_{31}–v_{52}

Swiss-Prot v_{47}–v_{56}

- Quality / classification criteria:
  - Provenance *exp, auth, cur, comp, auto*
  - Age *old* (> 1.5 years)
    middle (0.5 to 1.5 years)
    novel (< 0.5 years)
  - Stability *stable* (stab ≥ 0.9)
    unstable (stab < 0.9)

Age Analysis

	old	*middle*	*novel*
exp	10,114	3,387	8,808
auth	23,445	4,253	2,492
cur	253	67	157
comp	1,913	477	942
auto	92,474	63,954	49,909
sum	128,199	72,138	62,308

49% old annotations

Novel, manual annotations are predominantly *exp*

- Novel and middle aged annotations rarely classified as unstable (results see paper)

Stability Analysis

	$\lvert stab_{exis} \rvert$	$\lvert stab_{qual} \rvert$	$\lvert stab_{comb} \rvert$
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
unstable

High share of temporal absence

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

Use – Quality Analysis

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, …
- Filter less/more reliable annotations (e.g. stable, old, manually assigned)

How was the annotation created?

How reliable is the annotation?

Use – Quality Analysis

- 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 biological findings

Which annotations fit best for my application?

Conclusion and Future Work

- Generic approach to estimate the quality of ontology-based annotations by taking their evolution history into account
- Evaluation in two large life sciences sources

  → Instabilities for *auth* or *auto* annotations
- Different quality criteria: provenance, age, and stability to classify annotations

- Investigate other quality aspects
- Explore the impact of unstable annotations on dependent applications