Evolution-based Analysis of functional Protein Annotation

Anika Groß[1], Michael Hartung[1], Toralf Kirsten[1,3], Erhard Rahm[1,2]

[1] Interdisciplinary Centre for Bioinformatics, University of Leipzig http://www.izbi.de
[2] Department of Computer Science, University of Leipzig http://dbs.uni-leipzig.de
[3] Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig http://imise.uni-leipzig.de

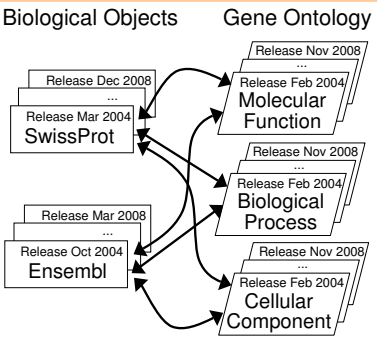
Motivation

Current State

- **Various and increasing number** of data sources storing molecular biological data, such as Ensembl and SwissProt for protein data
- **Annotations: ontology-based and semantic description** of biological objects, e.g., proteins are annotated with molecular functions or biological processes
- **Frequent changes** of both, data sources about biological objects and ontologies resulting in **different versions**
  - Addition of new experimental findings
  - Revision of existing knowledge

Problems

⇒ Evolution-based influences on dependent software systems and data, e.g., **outdated annotations**



Open Questions

- How different is the evolution in ontologies, protein data and annotations?
- **How stable** are annotations in different sources?
- **Which changes** exhibits a single annotation during its evolution process?
- How can **quality of annotations** be assessed to ensure enhanced quality in further analysis results?

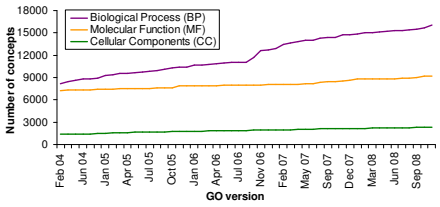
Goals

- **Quality-based ranking** of data sources
- **Filtering** of source-specific annotation data

⇒ **Evolution-based quantitative analysis** of biological data in the Gene Ontology (GO), Ensembl, SwissProt

Analysis Results

Evolution of Concepts in GO *



- High increase on the concept level in MF, BP, CC
- 97% / 26% / 69% more ontology concepts in MF / BP / CC
- 3.9 / 5.7 / 1.6 concept deletions and fusions on average per month in MF / BP / CC

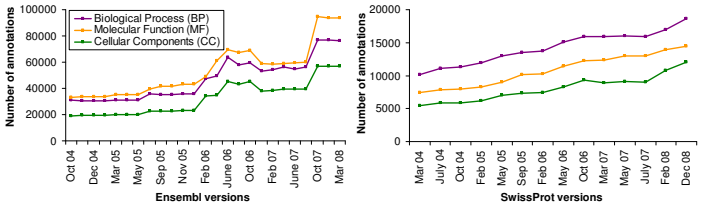
* Hartung, M.; Kirsten, T.; Rahm, E: Analyzing the Evolution of Life Science Ontologies and Mappings. Proc. 5th International Workshop on Data Integration in the Life Sciences (DILS), Paris, 2008

Evolution of Protein Data (Human Proteins)

	Ensembl	SwissProt
Number of proteins 2004	34111	10404
Number of proteins 2008	46742	20069
Growth rate	1.37 (37%)	1.93 (93%)
Percentage of annotated proteins 2004	52%	68%
Percentage of annotated proteins 2008	79%	59%

- **Significant increase of ontology concepts and proteins from 2004 to 2008**
- ***Add* is the dominant operation but there is also a significant number of *deletes***

Quantity Structure Evolution of GO Annotations in Ensembl & SwissProt



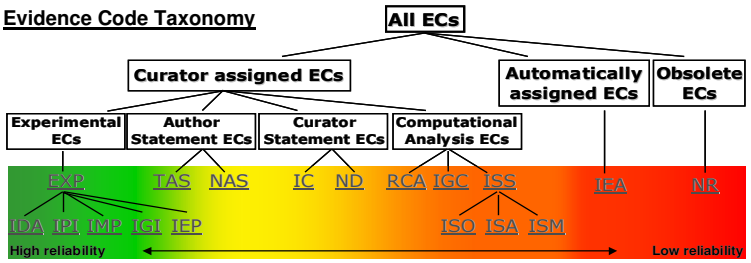
- Different evolution (2004 - 2008) in Ensembl and SwissProt
- Ensembl (SwissProt) has a growth rate of 2.73 (1.96)
- The most frequently used sub-ontology is MF (BP) in Ensembl (SwissProt)

Annotation Evaluation by Evidence Codes (EC)**

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

- Specifies the type of experiment or analysis that resulted in a GO annotation
- ECs are arranged in a taxonomy describing the reliability of an annotation

Evidence Code Taxonomy



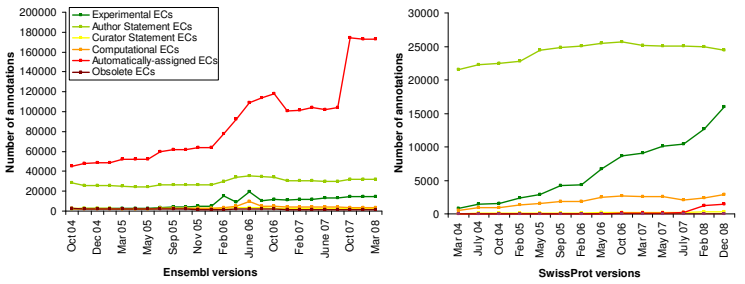
Example: Cytochrome b5 reductase 4 (SwissProt)

Concept id	Concept name	GO sub-ontology	v_{54}	v_{55}	v_{56}
GO:0004128	cytochrome-b5 reductase activity	MF	IEA	TAS	IDA
GO:0006091	generation of precursor metabolites and energy	BP		TAS	IDA
GO:0005737	cytoplasm	CC		TAS	
GO:0003032	detection of oxygen	BP			NAS
GO:0016174	NAD(P)H oxidase activity	MF			IDA

GO:0004128	GO:0005737
1 Add (v_{54}) 2 EC-Changes (v_{54}-v_{55}, v_{55}-v_{56})	1 Add (v_{55}) 1 Delete (v_{56})
"positive" annotation evolution (from automatically assigned to curator assigned)	"negative" annotation evolution (existed only in one version)

IDA	Inferred from Direct Assay
TAS	Traceable Author Statement
NAS	Non-traceable Author Statement
IEA	Inferred from Electronic Annotation

Annotation Evolution in different EC Groups



	Ensembl	SwissProt
Highest number of annotations	Automatically assigned (172648)	Author statement (24394)
Highest growth rate	Automatically assigned (3.85)	Experimental (18.85)
Degree of automation	High part of automatic assignent	Mostly manually curated

Aggregated EC-Changes in SwissProt (2004-2008)

	From	To	Diff
EXP	0	19	19
IDA	54	1052	998
IEP	97	17	-80
IGI	4	6	2
IMP	14	150	136
IPI	26	137	111
TAS	1240	187	**-1053**
NAS	238	273	35
IC	17	36	19
ISS	341	174	-167
IEA	21	1	-20

What is an EC-Change ($v_i \rightarrow v_{i+1}$)?

→ Persisting annotation, but its EC is revised from v_i to v_{i+1}

- Most EC-Changes occur towards IDA (Experimental)
- Most EC-Changes "leave" TAS (Author Statement)
- In SwissProt EC-Changes predominantly occur in order to annotate with experimental ECs

Conclusion & Future Work

Conclusions

- Annotations in **Ensembl** are highly **volatile**
- Ensembl covers significantly more annotations than SwissProt due to a high amount of additional **automatically assigned** annotations
- SwissProt is a **manually curated** data source and especially annotates with "high quality ECs" (author statement and experimental)
- Usage of annotations depends on the purpose of an application
  - **High quality but lower number of annotations** (e.g. Automatic annotation of new biological objects, Computation of ontology mappings)
  - **Low quality and very high number of annotations** (e.g. Annotation of protein networks with the objective of high coverage)

Future Challenges

- Computation of **stability values** by means of evolutionary information
  - to quantify the degree of evolution in ontologies and protein data
  - to determine the reliability of annotations (additional use of Evidence Code information)

Printed by Universitätsrechenzentrum Leipzig