Bio Data Management

Kapitel 7

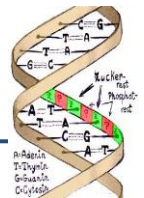
Annotationen

Wintersemester 2014/15
Dr. Anika Groß
Universität Leipzig, Institut für Informatik, Abteilung Datenbanken
http://dbs.uni-leipzig.de

UNIVERSITÄT LEIPZIG

Vorläufiges Inhaltsverzeichnis

1. Motivation und Grundlagen
2. Bio-Datenbanken
3. Datenmodelle und Anfragesprachen
4. Modellierung von Bio-Datenbanken
5. Sequenzierung und Alignments
6. Genexpressionsanalyse
7. Annotationen
8. Datenintegration: Ansätze und Systeme
9. Schema- und Ontologiematching
10. Versionierung von Datenbeständen



Informationsflut in den Lebenswissenschaften

- Zehntausende Gene und gen-regulatorische Elemente (in versch. Spezies)
- Zahlreiche Interaktionen in komplexen molekularen Netzwerken
- Millionen von Publikationen
- …

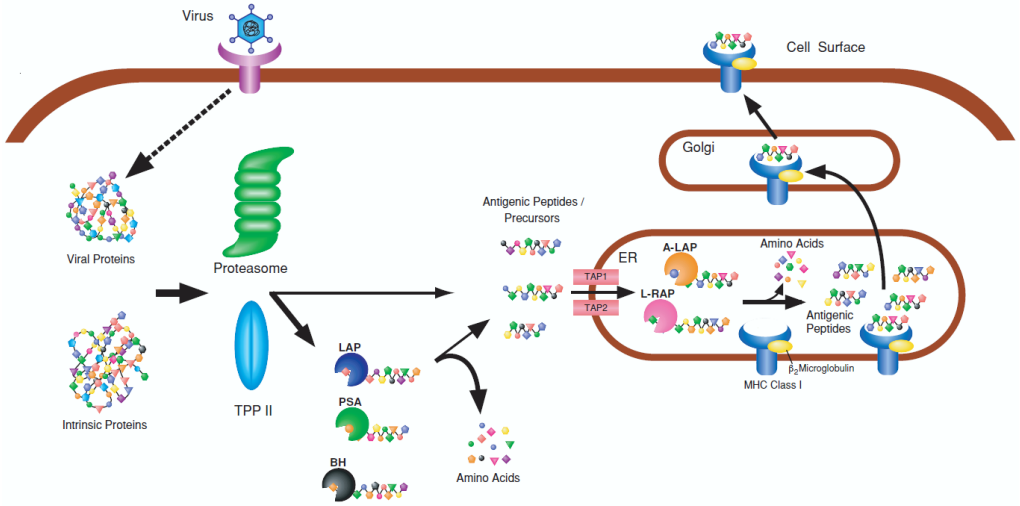
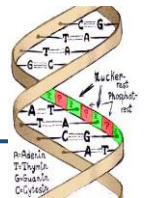


Fig. 1. Processing Pathway of Antigenic Peptides Presented to MHC Class I Molecules

Hattori A, Tsujimoto M.: Processing of antigenic peptides by aminopeptidases. Biol Pharm Bull., 2004

- Zuordnung
  - der Funktionen, Prozesse, Interaktionen zu biologischen Objekten
  - der enthaltenen Themen zu Publikationen
  - …

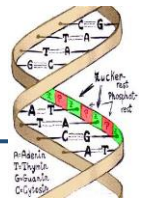
→ Nutzen maschinenverstehbarer Annotationen



Annotationen in den Lebenswissenschaften

- **Annotationen** dienen der semantischen Beschreibung der Eigenschaften von (biologischen) Objekten
- Zur Erfassung von Metadaten
- Möglichst einheitlicher Gebrauch der domänenspezifischen Begriffe
- Assoziationen eines zu beschreibenden Objekts zu den Begriffen eines Vokabulars bzw. den Konzepten einer Ontologie

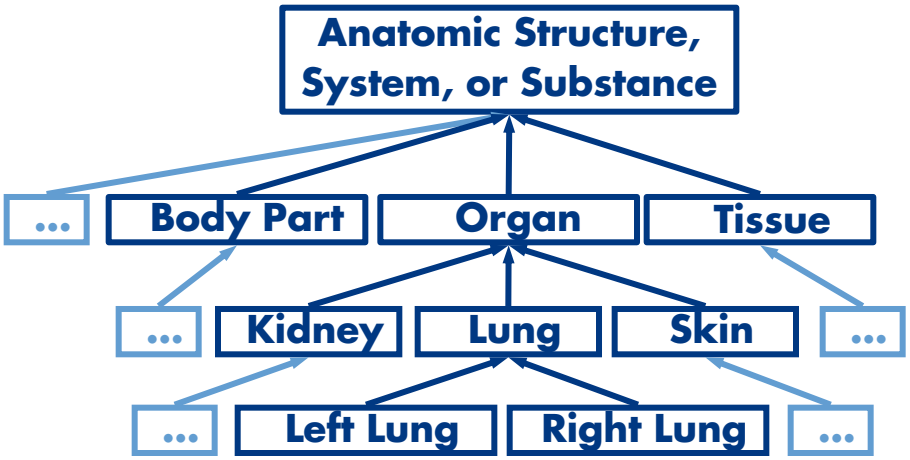
- Zunächst: Was ist eine Ontologie???



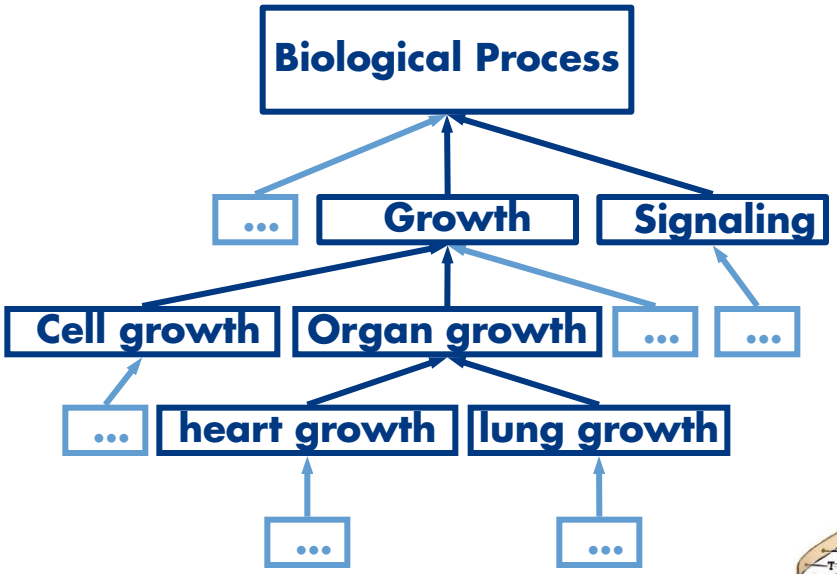
Ontologien

- Begriff in Philosophie ↔ **Informatik**
- "An ontology is an explicit specification of a conceptualization." (T. R. Gruber, 1993)
- Erweiterung: "An ontology is an explicit, formal specification of a shared conceptualization in a domain of interest."

Ausschnitt NCI Thesaurus



Ausschnitt Gene Ontology (GO)



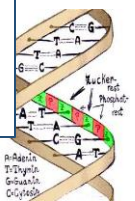
Repräsentation von Bio-Ontologien

```
[Term]
id: MA:0000072
name: heart
relationship: part_of MA:0000010 ! cardiovascular system
relationship: part_of MA:0002449 ! heart/pericardium
```

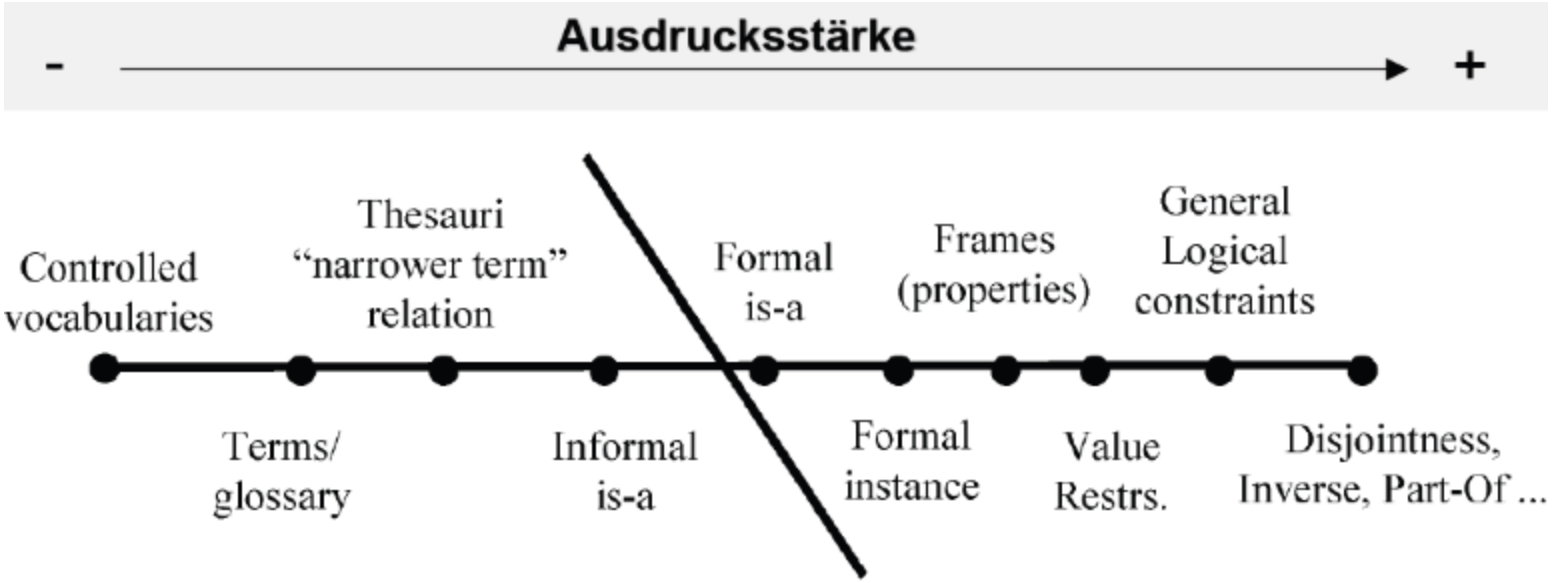
OBO = Open Biomedical Ontologies

```
<owl:Class rdf:about="http://purl.org/obo/owl/MA#MA_0000072">
  <rdfs:label xml:lang="en">heart</rdfs:label>
    <oboInOwl:hasOBONamespace>adult_mouse_anatomy.gxd</oboInOwl:hasOBONamespace>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty>
        <owl:ObjectProperty rdf:about="http://purl.org/obo/owl/OBO_REL#part_of"/>
      </owl:onProperty>
      <owl:someValuesFrom rdf:resource="http://purl.org/obo/owl/MA#MA_0000010"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty>
        <owl:ObjectProperty rdf:about="http://purl.org/obo/owl/OBO_REL#part_of"/>
      </owl:onProperty>
      <owl:someValuesFrom rdf:resource="http://purl.org/obo/owl/MA#MA_0002449"/>
    </owl:Restriction>
  </rdfs:subClassOf>
 </owl:Class><
```

OWL = Web Ontology Language

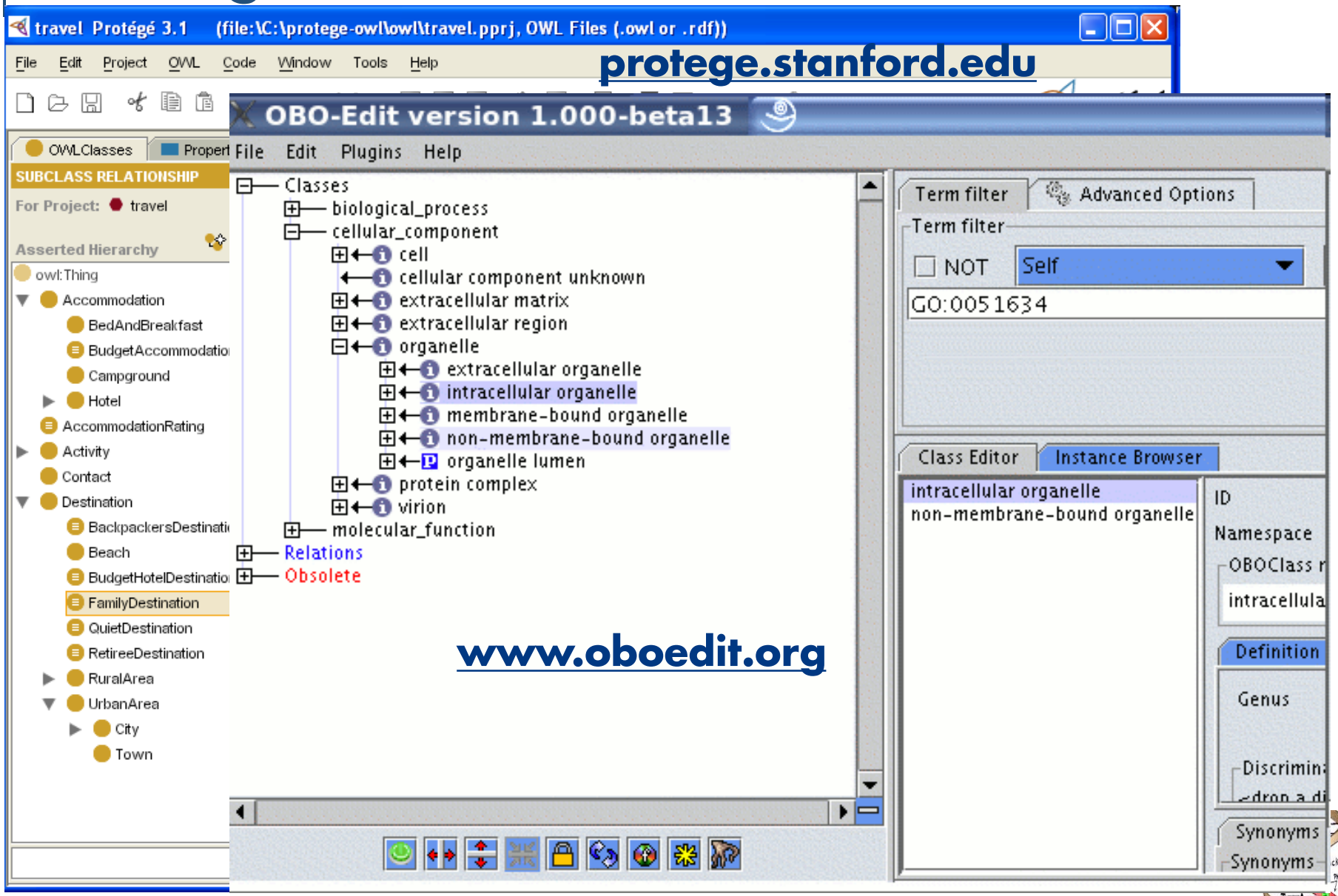


Ontologietypen



O.Lassila, D. McGuinness: *The Role of Frame-Based Representation on the Semantic Web*. Stanford Knowledge Systems Laboratory, 2001.

Ontologie-Editoren



Ontologie-Browser

the Gene Ontology http://amigo.geneontology.org

BioPortal | Browse | Search | Projects | Annotate | All Mappings | All Resources Alpha

bioportal.bioontology.org

OnEX Ontology Evolution Explorer

www.izbi.de/onex

Quantitative Analysis | Concept-based Analysis | Annotation Migration | Help | Contact

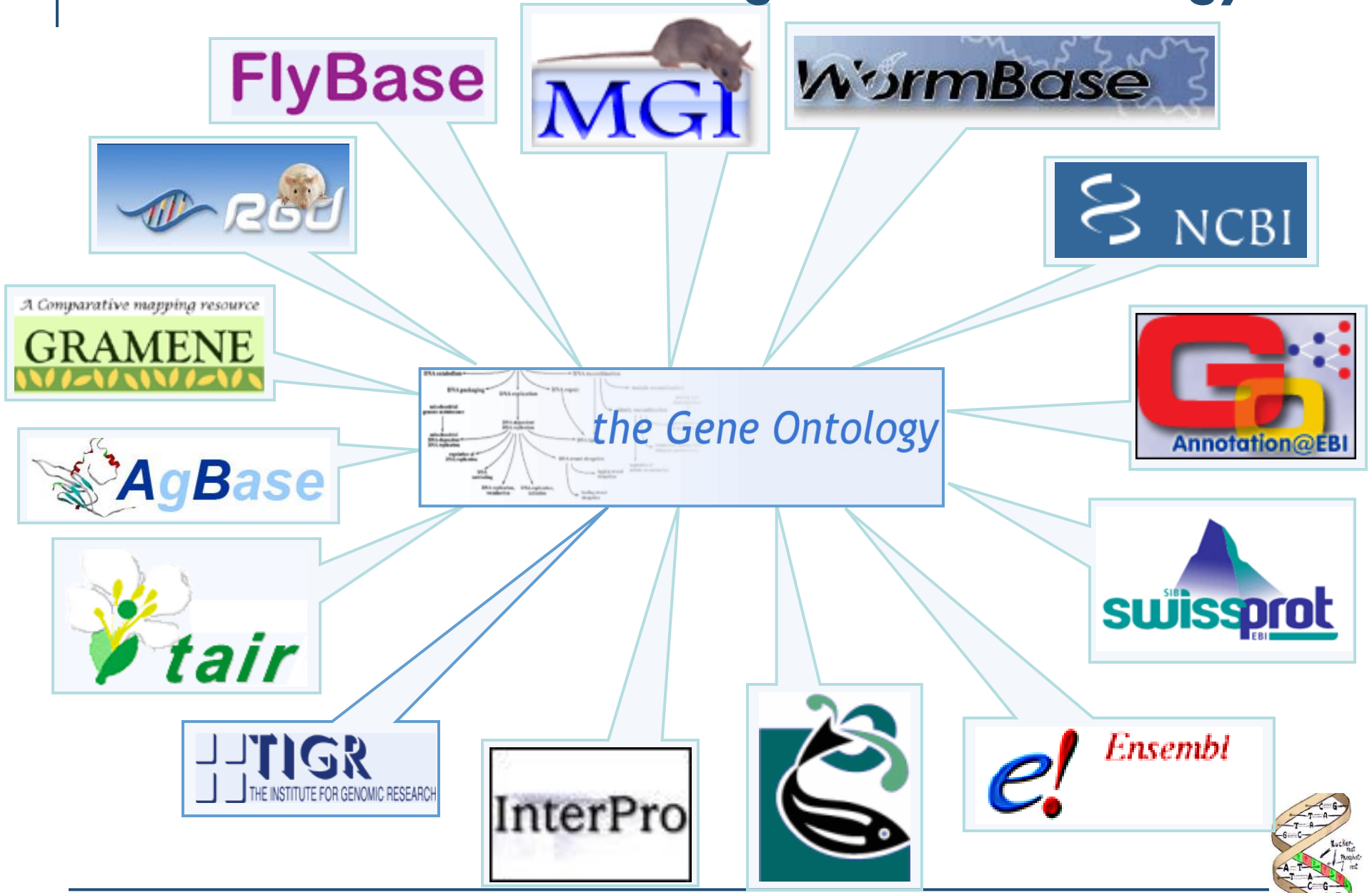
Ontology Evolution Overview

Ontology Evolution Chart - Process@GeneOntology (time, number of concepts and relationships)

concepts relationships

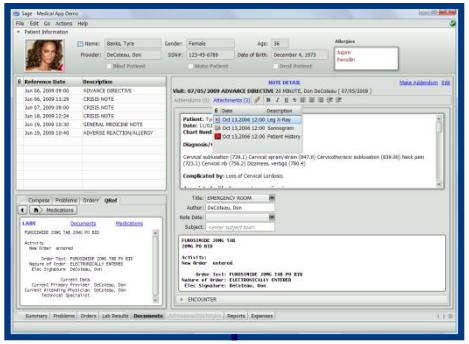
Ontology							
BioChemistryEntity@ChemicalEntitiesOfBiomedic							
AnatomicalEntity@CellTypeOntology							
AnatomicalEntity@FlyAnatomyOntology							
Concept@FlyBaseControlledVocabulary							
Process@GeneOntology							
Component@GeneOntology							
Function@GeneOntology							
AnatomicalEntity@AdultMouseAnatomyOntology							
Experiment@ProteinProteinInteractionOntology							
PhenotypeEntity@MammalianPhenotypeOntology							
AnatomicalEntity@PlantStructureOntology			2005-07	2009-08	681	868	980
Process@PathwayOntology			2005-11	2010-03	427	751	478
Concept@SequenceOntology			2005-08	2010-03	981	1764	1181
AnatomicalEntity@ZebraFishAnatomyOntology			2005-11	2009-12	1389	2431	4272
HealthEntity@NCIThesaurus			2003-10	2009-12	28740	77448	33847
Protein@ProteinModificationOntology			2006-06	2009-02	1074	1338	1568

Weit verbreitete Anwendung der Gene Ontology



Arten von Annotationen

- Biologische Objekte (Gene, Proteine, …)
- Pathways, Netzwerke
- …



Experiment-Annotation
MIAME, MGED…

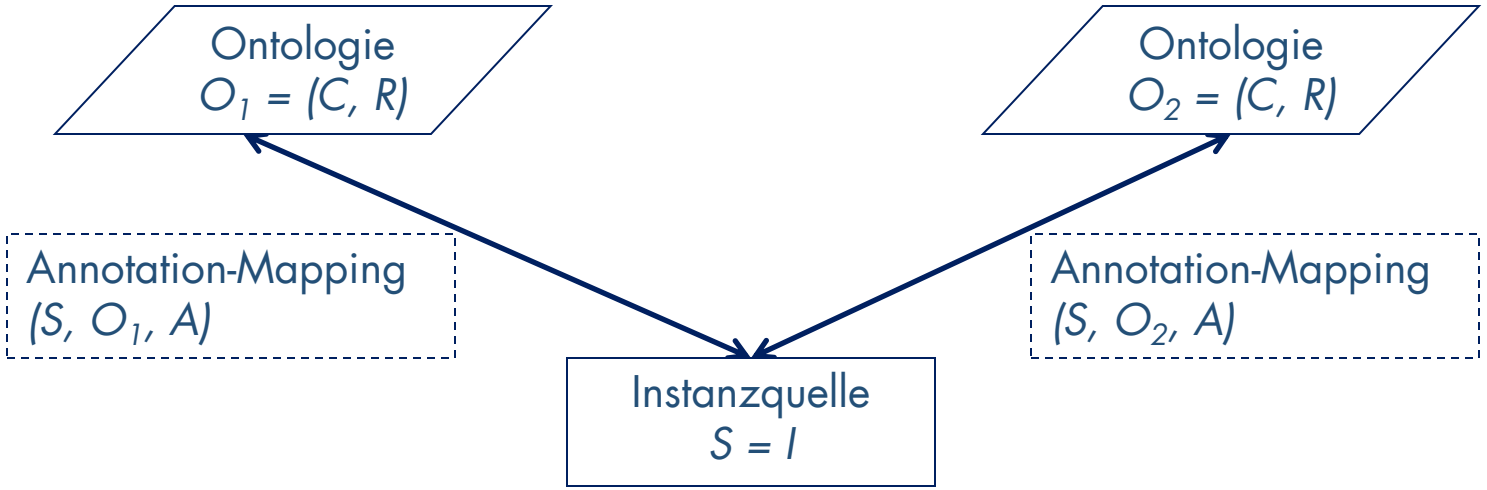
Annotation biologischer Objekte / Pathways
GO, KEGG, …

Annotation von Publikationen (indexing)
MeSH, GO,…

Annotation von klinischen Dokumenten
ICD, SNOMED, LOINC,…

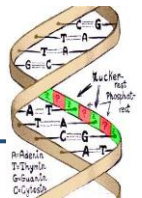
MIAME	*Minimum Information About a Microarray Experiment*
MGED	*Microarray Gene Expression Data*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
MeSH	*Medical Subject Headings*
ICD	*International Classification of Diseases*
SNOMED	*Systematized Nomenclature of Medicine Clinical Terms*
LOINC	*Logical Observation Identifiers Names and Codes*
NCIT	*NCI (National Cancer Institute) Thesaurus*

Annotationsmodell



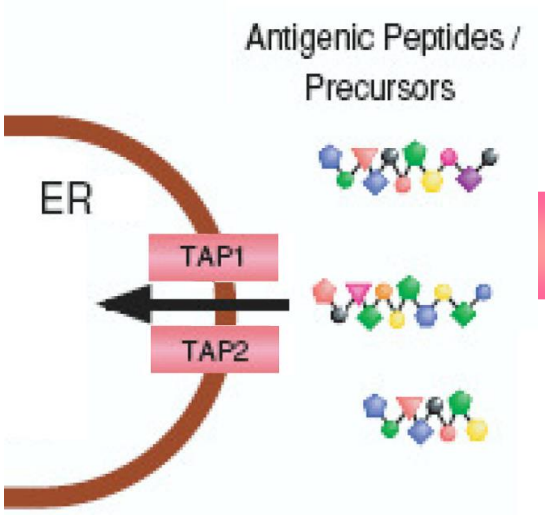
C Menge von Konzepten
R Menge von Relationen
I Menge von Instanzen
A Menge von Annotationen / Korrepondenzen

- Annotation $a=(i,c)$, $a \in A$, $i \in I$, $c \in C$,
- *object_id - concept_id*



Beispiel: GO-Annotationen

- Assoziation eines Proteins zu einem/r bestimmten biologischen Prozess / zellulären Komponente / molekularen Funktion



TAP1

Component: is integral to membrane

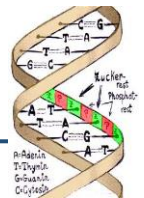
Process: peptide transport

Function: protein heterodimerization activity

GO-Konzept "is integral to membrane" - Details

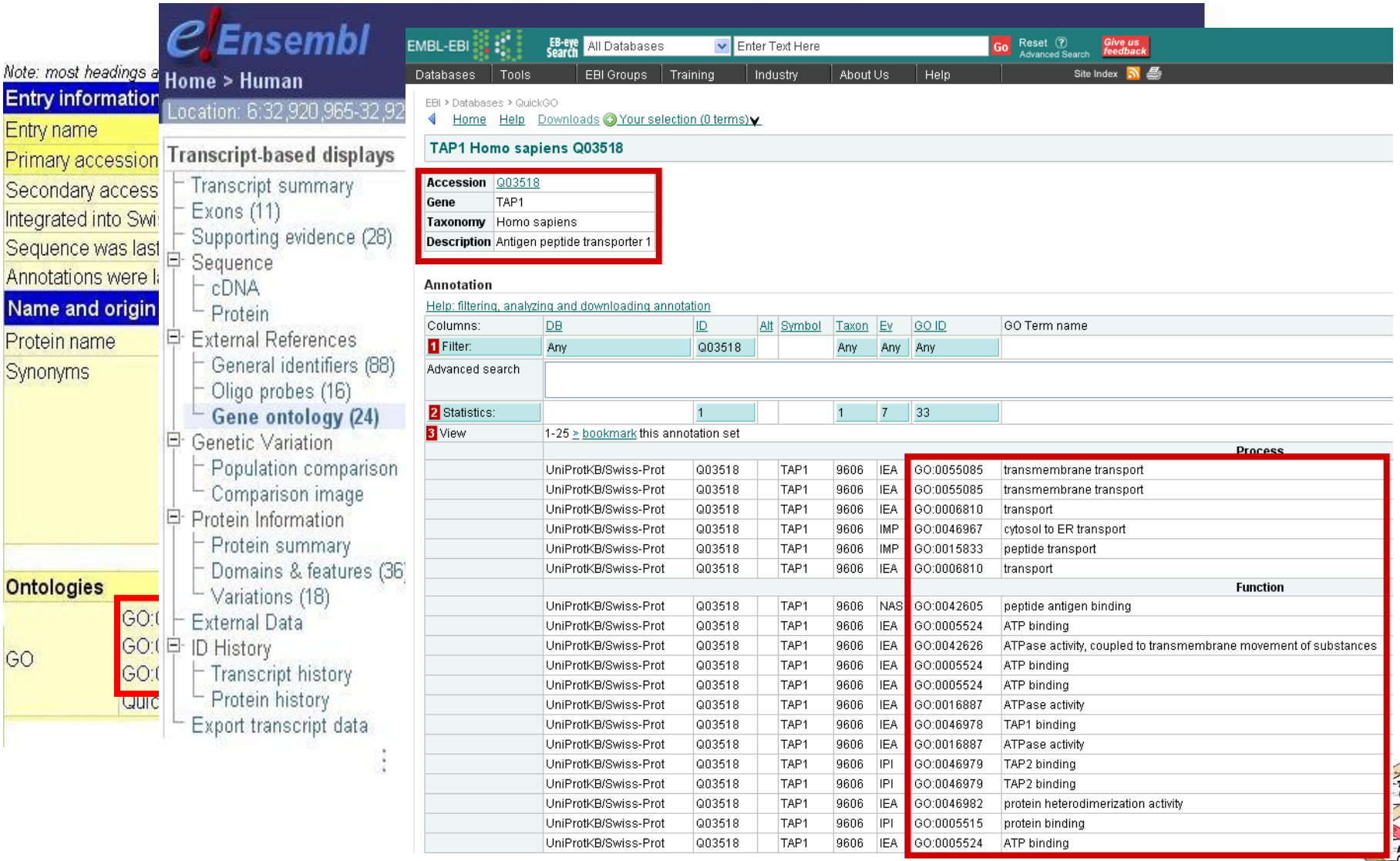
Accession	GO:0016021
Name	is integral to membrane
Definition	Penetrating at least one phospholipid bilayer of a membrane. [...]
Synonym	transmembrane
is_obsolete	false

…



Beispiel: GO-Annotationen

UniProtKB/Swiss-Prot entry Q03518



Note: most headings a

Entry information
Entry name
Primary accession
Secondary access
Integrated into Swi
Sequence was last
Annotations were l
Name and origin
Protein name
Synonyms

Ontologies
GO
GO:
GO:
GO:
Quic

e!Ensembl
Home > Human
Location: 6:32,920,965-32,92

Transcript-based displays
- Transcript summary
- Exons (11)
- Supporting evidence (28)
- Sequence
  - cDNA
  - Protein
- External References
  - General identifiers (88)
  - Oligo probes (16)
  - Gene ontology (24)
- Genetic Variation
  - Population comparison
  - Comparison image
- Protein Information
  - Protein summary
  - Domains & features (36
  - Variations (18)
- External Data
- ID History
  - Transcript history
  - Protein history
- Export transcript data

EMBL-EBI | EB-eye Search | All Databases | Enter Text Here | Go | Reset | Advanced Search | Give us feedback
Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

EBI > Databases > QuickGO
Home | Help | Downloads | Your selection (0 terms)

TAP1 Homo sapiens Q03518

Accession	Q03518
Gene	TAP1
Taxonomy	Homo sapiens
Description	Antigen peptide transporter 1

Annotation

Help: filtering, analyzing and downloading annotation

Columns:	DB	ID	Alt	Symbol	Taxon	Ev	GO ID	GO Term name
1 Filter:	Any	Q03518			Any	Any	Any	
Advanced search								
2 Statistics:		1			1	7	33	
3 View	1-25 > bookmark this annotation set							
								Process
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IEA	GO:0055085	transmembrane transport
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IEA	GO:0055085	transmembrane transport
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IEA	GO:0006810	transport
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IMP	GO:0046967	cytosol to ER transport
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IMP	GO:0015833	peptide transport
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IEA	GO:0006810	transport
								Function
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	NAS	GO:0042605	peptide antigen binding
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IEA	GO:0005524	ATP binding
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IEA	GO:0042626	ATPase activity, coupled to transmembrane movement of substances
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IEA	GO:0005524	ATP binding
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IEA	GO:0005524	ATP binding
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IEA	GO:0016887	ATPase activity
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IEA	GO:0046978	TAP1 binding
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IEA	GO:0016887	ATPase activity
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IPI	GO:0046979	TAP2 binding
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IPI	GO:0046979	TAP2 binding
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IEA	GO:0046982	protein heterodimerization activity
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IPI	GO:0005515	protein binding
	UniProtKB/Swiss-Prot	Q03518		TAP1	9606	IEA	GO:0005524	ATP binding

Beispiel: MeSH-Annotationen

NCBI Resources How To

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health

Search: PubMed Limits Advanced search Help

Search Clear

Display Settings: Abstract Send to:

BMC Bioinformatics. 2009 Aug 13;10:250.

OnEX: Exploring changes in life science ontologies.

Hartung M, Kirsten T, Gross A, Rahm E.

Interdisciplinary Centre for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, 04107 Leipzig, Germany. hartung@izbi.uni-leipzig.de

Abstract

BACKGROUND: Numerous ontologies have recently been developed in life sciences to support a consistent annotation of biological objects, such as genes or proteins. These ontologies underlie continuous changes which can impact existing annotations. Therefore, it is valuable for users of ontologies to study the stability of ontologies and to see how many and what kind of ontology changes occurred. RESULTS: We present OnEX (Ontology Evolution EXplorer) a system for exploring ontology changes. Currently, OnEX provides access to about 560 versions of 16 well-known life science ontologies. The system is based on a three-tier architecture including an ontology version repository, a middleware component and the OnEX web application. Interactive workflows allow a systematic and explorative change analysis of ontologies and their concepts as well as the semi-automatic migration of out-dated annotations to the current version of an ontology. CONCLUSION: OnEX provides a user-friendly web interface to explore information about changes in current life science ontologies. It is available at http://www.izbi.de/onex.

PMID: 19678926 [PubMed - indexed for MEDLINE] PMCID: PMC2746816 **Free PMC Article**

Publication Types, MeSH Terms

Publication Types:

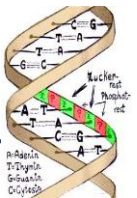
Research Support, Non-U.S. Gov't

MeSH Terms:

Computational Biology/methods*
Databases, Factual
Gene Expression Profiling
Internet
Software*
User-Computer Interface
Vocabulary, Controlled

LinkOut - more resources

MeSH = Medical Subject Headings



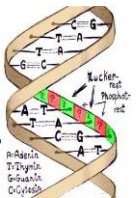
Repräsentation von Annotationen

- Nicht normalisiert, z.B. CSV-Dateien von GOA
- Entry-Modell, z.B. Entry-Modell von SwissProt
- Normalisiert, ER, z.B. MySQL-DB von Ensembl

GOA

```
!CVS Version: Revision: 1.84 $
!GOC Validation Date: 03/13/2008 $
!Submission Date: 3/13/2008
....
UniProtKB   Q03468  ERCC6_HUMAN     GO:0005654  PMID:16107709   IDA     C   ERCC6, CSB: DNA excision repair protein ERCC-6  IPI00414779 protein taxon:9606  20071002    UniProtKB
UniProtKB   Q03468  ERCC6_HUMAN     GO:0005730  PMID:16107709   IDA     C   ERCC6, CSB: DNA excision repair protein ERCC-6  IPI00414779 protein taxon:9606  20071002    UniProtKB
UniProtKB   Q03468  ERCC6_HUMAN     GO:0008023  PMID:9326587    IDA     C   ERCC6, CSB: DNA excision repair protein ERCC-6  IPI00414779 protein taxon:9606  20071003    UniProtKB
UniProtKB   Q03518  TAP1_HUMAN      GO:0000166  GOA:interpro|GO_REF:0000002 IEA InterPro:IPR003593  F   TAP1, ABCB2, PSF1, RING4, Y3: Antigen peptide transporter 1 IPI00646625 protein taxon:9606  20080310    UniProtKB
UniProtKB   Q03518  TAP1_HUMAN      GO:0005215  GOA:interpro|GO_REF:0000002 IEA InterPro:IPR005293  F   TAP1, ABCB2, PSF1, RING4, Y3: Antigen peptide transporter 1 IPI00646625 protein taxon:9606  20080310    UniProtKB
UniProtKB   Q03518  TAP1_HUMAN      GO:0005515  PMID:17055437   IPI UniProtKB:Q03519    F   TAP1, ABCB2, PSF1, RING4, Y3: Antigen peptide transporter 1 IPI00646625 protein taxon:9606  20080310    IntAct
UniProtKB   Q03518  TAP1_HUMAN      GO:0005515  PMID:17055437   IPI UniProtKB:P30101    F   TAP1, ABCB2, PSF1, RING4, Y3: Antigen peptide transporter 1 IPI00646625 protein taxon:9606  20080310    IntAct
UniProtKB   Q03518  TAP1_HUMAN      GO:0015198  GOA:spkw|GO_REF:0000004 IEA SP_KW:KW-0571   F   TAP1, ABCB2, PSF1, RING4, Y3: Antigen peptide transporter 1 IPI00646625 protein taxon:9606  20080310    UniProtKB
```

db	- UniProtKB
db_object_id	- Q03518
db_object_symbol	- TAP1_HUMAN
qualifier	-
go_id	- GO:0015198
db_reference	- GOA:spkw|GO_REF:0000004
Evidence	- IEA
EC_with	- SP_KW:KW-0571
aspect	- F
db_object_name	- Antigen peptide transporter 1
synonym	- TAP1, ABCB2, PSF1, RING4, Y3: Antigen peptide transporter 1, IPI00646625
db_object_type	- Protein
tax_id	- taxon:9606
ins_date	- 20080310
assigned_by	- UniProtKB

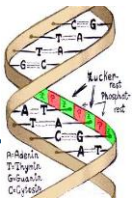


Repräsentation von Annotationen

```
ID   TAP1_HUMAN              Reviewed;         808 AA.
AC   Q03518; Q16149; Q96CP4;
DT   01-JUN-1994, integrated into UniProtKB/Swiss-Prot.
DT   25-NOV-2008, sequence version 2.
DT   18-MAY-2010, entry version 117.
DE   RecName: Full=Antigen peptide transporter 1;
DE            Short=APT1;
...          ...
```

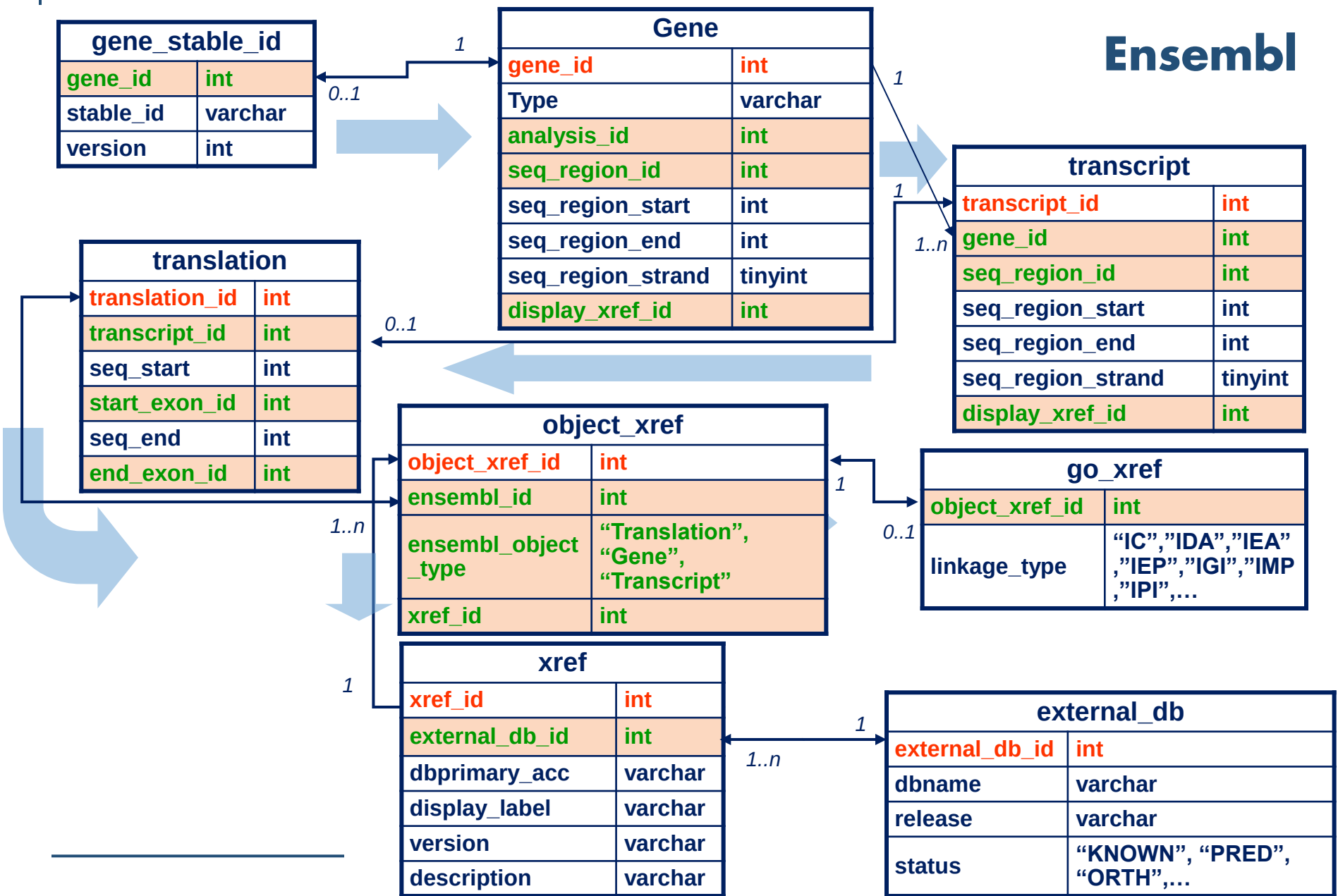
SwissProt

```
DR   GO; GO:0005829; C:cytosol; NAS:UniProtKB.
DR   GO; GO:0030176; C:integral to endoplasmic reticulum membrane; IDA:UniProtKB.
DR   GO; GO:0042825; C:TAP complex; IDA:UniProtKB.
DR   GO; GO:0043531; F:ADP binding; IDA:UniProtKB.
DR   GO; GO:0005524; F:ATP binding; IDA:UniProtKB.
DR   GO; GO:0042626; F:ATPase activity, coupled to transmembrane m...; IEA:InterPro.
DR   GO; GO:0042605; F:peptide antigen binding; NAS:UniProtKB.
DR   GO; GO:0015197; F:peptide transporter activity; IMP:UniProtKB.
DR   GO; GO:0042803; F:protein homodimerization activity; ISS:UniProtKB.
DR   GO; GO:0046978; F:TAP1 binding; ISS:UniProtKB.
DR   GO; GO:0046979; F:TAP2 binding; IPI:UniProtKB.
DR   GO; GO:0019885; P:antigen processing and presentation of endo...; IMP:UniProtKB.
DR   GO; GO:0046967; P:cytosol to ER transport; IMP:UniProtKB.
DR   GO; GO:0006955; P:immune response; IEA:UniProtKB-KW.
DR   GO; GO:0019060; P:intracellular transport of viral proteins i...; IMP:UniProtKB.
DR   GO; GO:0015833; P:peptide transport; IMP:UniProtKB.
DR   GO; GO:0055085; P:transmembrane transport; IEA:InterPro.
DR   InterPro; IPR013305; ABC_B2.
DR   InterPro; IPR003439; ABC_transporter-like.
DR   InterPro; IPR017871; ABC_transporter_CS.
DR   InterPro; IPR017940; ABC_transporter_type1.
DR   InterPro; IPR001140; ABC_transptr_TM_dom.
DR   InterPro; IPR011527; ABC_transptrTM_dom_typ1.
DR   InterPro; IPR005293; Ag_transporter2.
DR   InterPro; IPR003593; ATPase_AAA+_core.
```

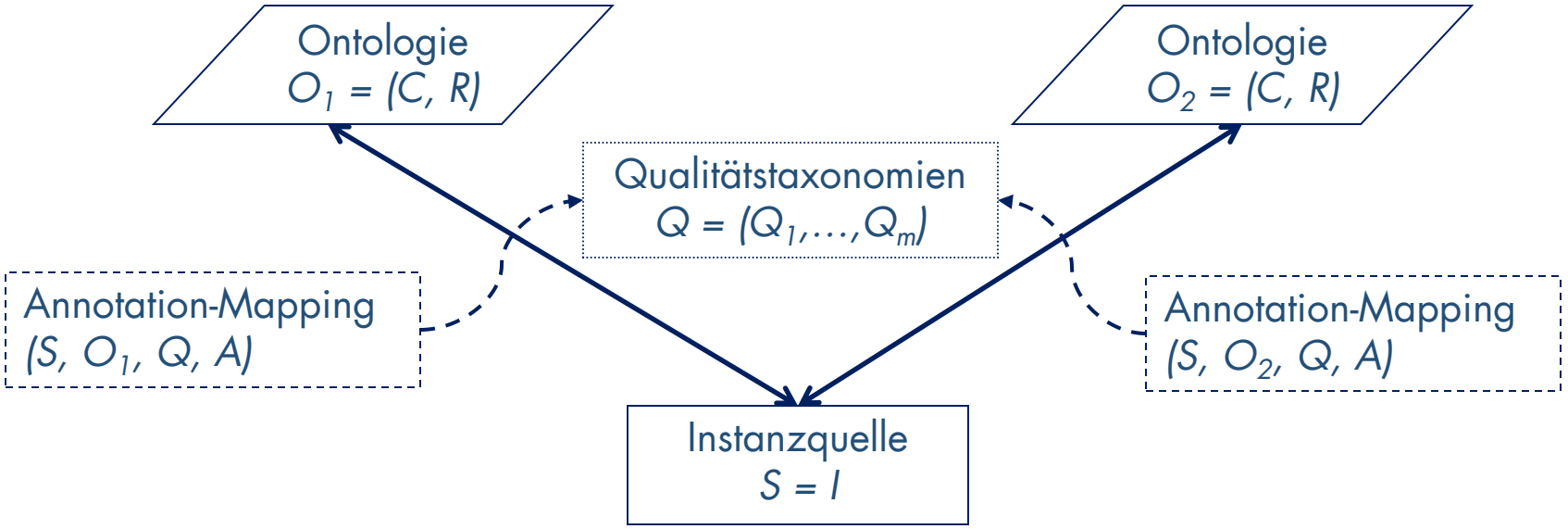


Repräsentation von Annotationen

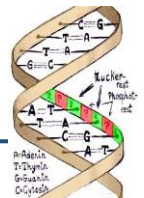
Ensembl



Erweitertes Annotationsmodell

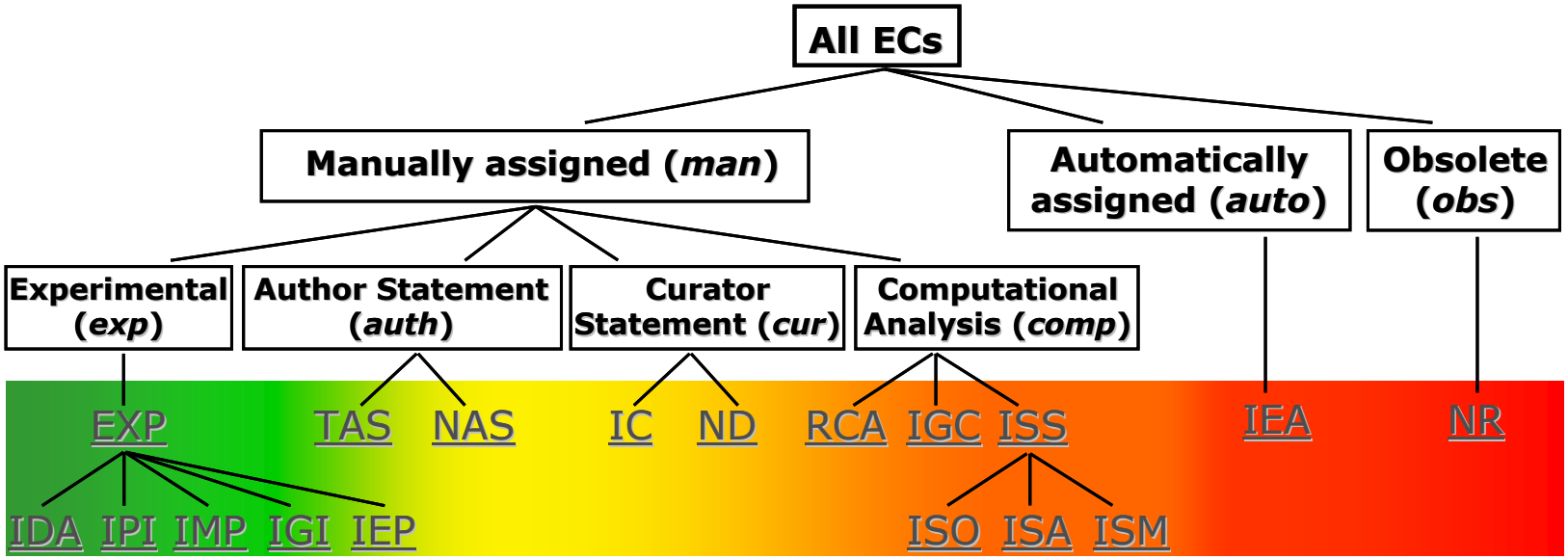


- Speichern zusätzlicher Korrespondenz-Informationen
- Verweis auf Qualitätstaxonomie(n) ⇨ Annotation $a \in A$, $a=(i,c,\{q\})$
- *concept_id – object_id – {qual_term}*
- z.B. Annotationsherkunft (*provenance*) ⇨ Evidence Codes
- Andere: Curator, Methode, Datum, Quelle …



Herkunft von GO-Annotationen - Evidence Codes

- Herkunft/Ursprung von Annotationen: engl. provenance
- Evidence Code (EC) * gibt an, woher die Annotation zu einem bestimmten Term stammt z.B. durch welche Art von Experiment oder Analyse wurde die Information nachgewiesen
- Repräsentation durch Taxonomie → kann als Qualitätsmerkmal genutzt werden, Zuverlässigkeit (engl. reliability)



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

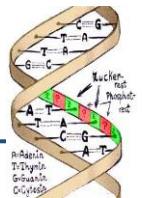
Methoden der automatischen Annotation

- Ziel: Vorschlagsgenerierung (Verifikation notwendig!)

- IEA (Inferred from Electronic Annotation): automatische Zuordnung, kein Curator; automatische Berechnung oder Übertragung von Annotationen
- ISS (Inferred from Sequence or Structural Similarity): sequenz-basierte Analyse, Zuordnung nach manueller Bewertung

Methoden

- Sequenzähnlichkeit, Homologie-Mappings
- Funktionale Gruppen (funktional ähnliche Proteine, InterPro)
- Keyword-Mappings (SwissProt, Enzyme Codes)
- Text-Mining: Automatische Extraktion von Annotationen aus Publikationen

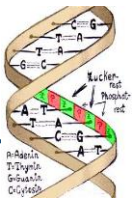


Text Mining zur Automatischen Annotation

- Text Mining: Automatisierter Prozess zur Analyse natürlicher Sprache in Texten

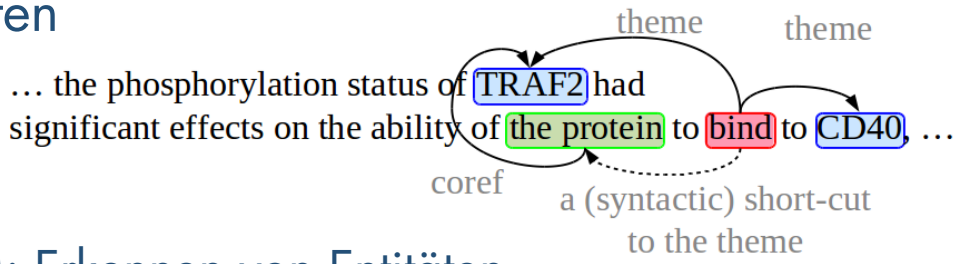
1) Auffinden relevanter Dokumente (Information Retrieval (IR))
   - z.B. alle Artikel zu einem Protein, alle neuen Artikel (letzter Monat/letztes Jahr/…), …
   - Suche in Literaturdatenbank (z.B. Pubmed)
     - Query auf indexierten Texten
     - Einbeziehen von Textstatistiken
       - z.B. inverse Dokumenthäufigkeit (Inverse document frequency)

$$IDF_t = \log\left(\frac{Gesamtanzahl\ an\ Dokumenten\ in\ DB}{Anzahl\ der\ Dokumente\ mit\ dem\ Term\ t}\right)$$



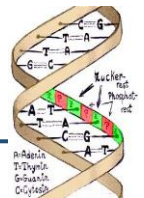
Text Mining zur Automatischen Annotation

2) Auffinden biologisch bedeutsamer semantischer Strukturen in Texten (Information Extraction (IE))
   a) Tagging biologischer Entitäten
   b) Erkennen von Beziehungen



- Named Entity Recognition (NER): Erkennen von Entitäten (Proteine, Gene, Krankheiten, Funktionen, Prozesse …) in Texten
  - z.B. erwähnt als Genesymbol/ID/Accession, Name, Synonym, …
- Stemming (Identifizieren des Wortstamms)
- POS (Part of speech): Identifizieren der Wortart (z.B. Substantiv, Verb, Adjektiv, …)
- Verwenden von domänenspezifischen Wörterbüchern, Ontologien, …
- Identifikation von Mustern, Nutzen regelbasierter Ansätze, maschinelles Lernen (SVM, HMM, Entscheidungsbaum, …)
- …

Abb: http://www.kdnuggets.com/2013/01/bionlp-shared-task-text-mining-for-biology-competition.html

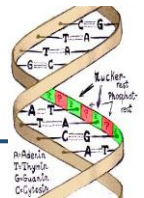


Nutzen / Anwendungen von Annotationen

- Strukturiertes Erfassen von Wissen statt natürlicher Sprache (einheitliche Begriffswelt, maschinenverstehbar)

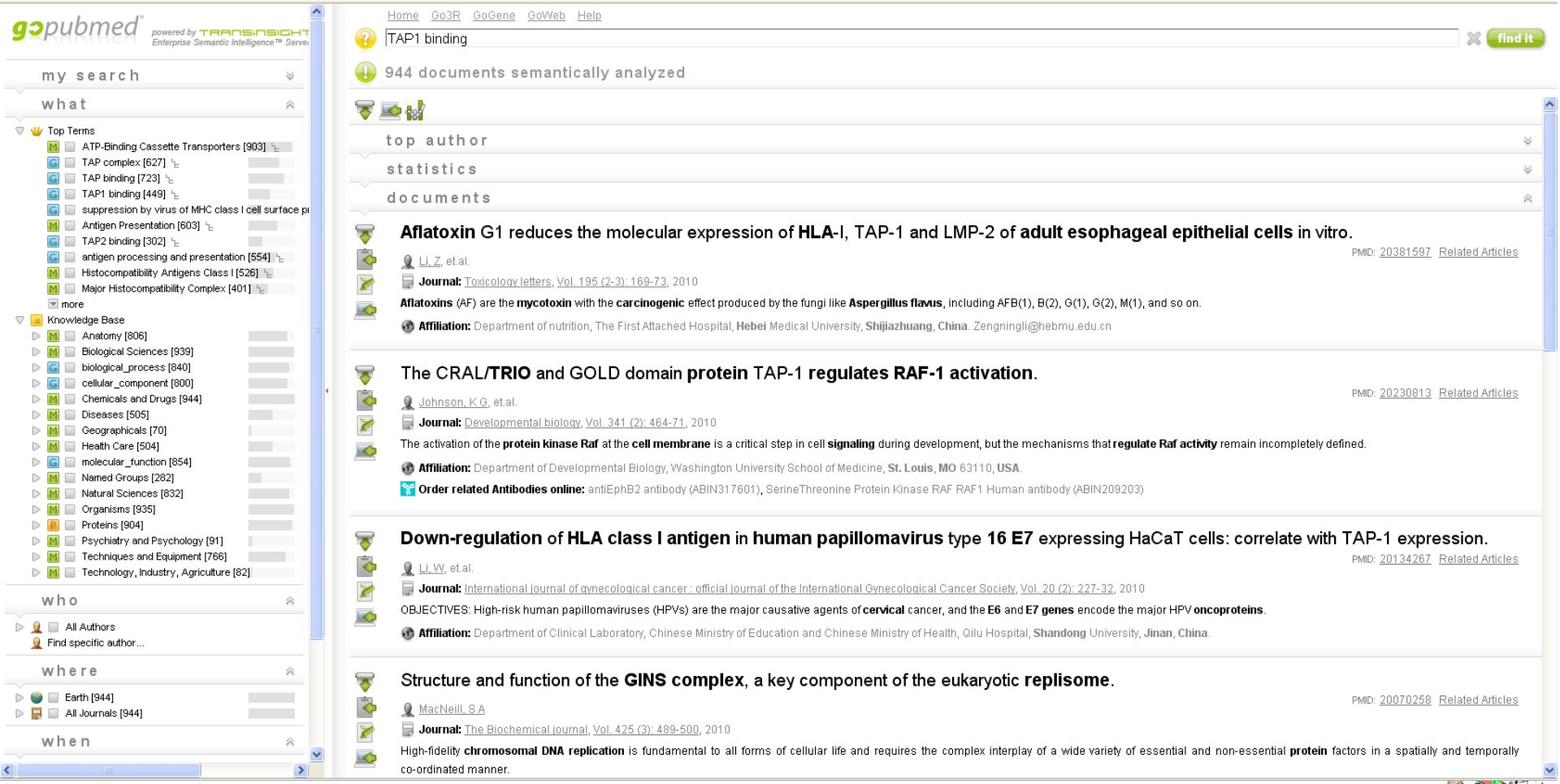
→ Erleichterte Suche & Navigation

- Ermittlung signifikanter über-/unterrepräsentierter GO-Terme in einer Menge von biologischen Objekten
- Unterstützung von Text-Mining / Natural Language Processing
- Ontology Matching (z.B. instanzbasiertes Matchen von GO Subontologien, GOPubmed…)
- Weitere Anwendung MeSH: 2 Mengen von Publikationen aus untersch. Jahren mit MeSH Annotationen → Ermittlung der „Hype"-Themen
- Nutzen in klinischen Studien
  - Beschreibung der Experimente – strukturierte Erfassung der Phänotypen von Teilnehmern (kurzfristig)
  - Therapie-Vorschlagsgenerierung (langfristig)



Applikation: GO-Pubmed

- Klassifikation von PubMed-Artikeln / Abstracts unter Verwendung von GO- und MeSH-Kategorien

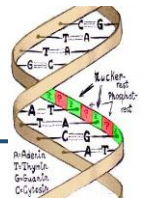


Applikation: Term Enrichment

- Auch: Funktionale Analysen, functional profiling
- Ermittlung signifikant über-/unterrepräsentierter GO-Terme in einer Menge von biologischen Objekten im Vergleich zu einer Hintergrundmenge (z.B. alle Gene einer Spezies)
- Automatische Zuordnung biologischer Funktionen in sehr großen Datensätzen
- Daten stammen z.B. aus Microarray Experimenten zur Genexpression, positiven Selektion von Arten (Evolutionsanalysen), …
- Hypergeometrische Verteilung, Wilcoxon Rank Test, Binomialverteilung…
- Tools: GO::Termfinder, FUNC, FatiGO …

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



Applikation: Term Enrichment

Eingabe

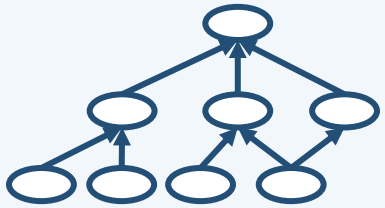
Liste von Genen/Proteinen mit GAV
(z.B. Expressionslevel)

Gene	GAV*
CTSR4	1
CA151	0
TNR1B	0
RPB4	1
MOGS	0

Menge von Annotationen

Gene	Concept
CTSR4	GO:0005216 (ion channel activity)
CTSR4	GO:0006811 (ion transport)
CTSR4	GO:0016020 (membrane)
CTSR4	GO:0016021 (integral to membrane)

Ontologie
(z.B. Gene Ontology)



Term Enrichment
(statistischer Signifikanztest)

Ausgabe

Liste von Ontologiekonzepten mit „Signifikanzwert“ bzgl. der untersuchten Gene/Proteine

(z.B. Liste von Funktionen und Prozessen aus *GO mit FWER)*

Concept		FWER**
GO:0004984	(olfactory receptor activity)	0
GO:0016021	(integral to membrane)	0
GO:0031224	(intrinsic to membrane)	0
GO:0044425	(membrane part)	0
GO:0016020	(membrane)	0.0001
GO:0007166	(cell surface receptor linked signal transduction)	0.0002
GO:0050877	(neurophysiological process)	0.0005
GO:0050874	(organismal physiological process)	0.0007
GO:0050896	(response to stimulus)	0.0033
GO:0004871	(signal transducer activity)	0.0046

** GAV = Gene-Associated Variable, ** FWER = Family Wise Error Rate*

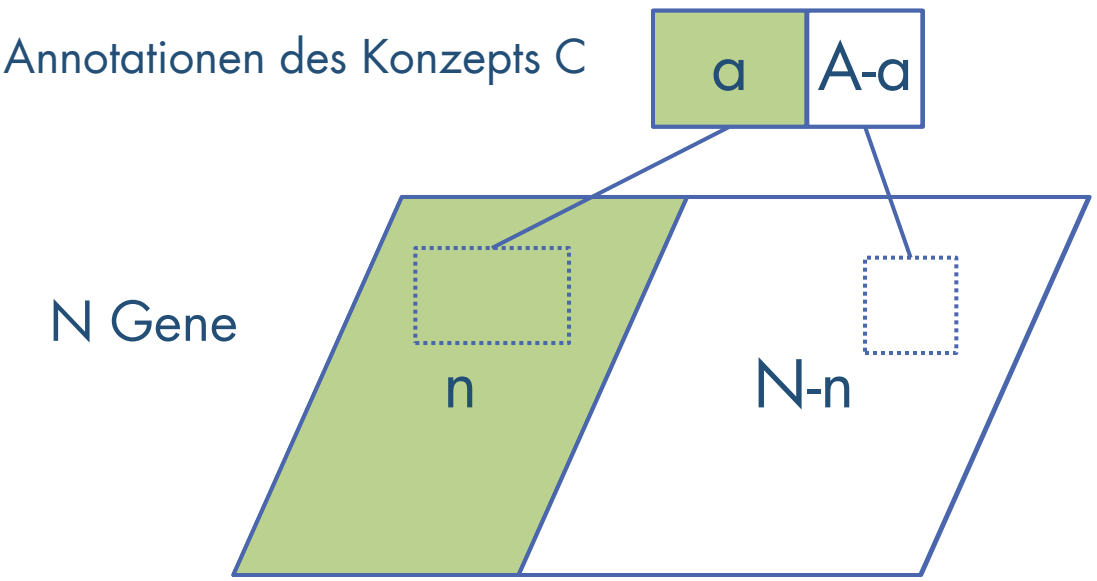
Hypergeometrischer Test

- Berechne den pValue für jedes Konzept C (signifikant: pValue≤0.05)

$$pValue(C) = 1 - \sum_{i=0}^{a} \frac{\binom{A}{i}\binom{N-A}{n-i}}{\binom{N}{n}}$$

A - |annotierte Gene zu Konzept C|
a - |annotierte "interessante" Gene * von C|
N - |alle Gene|
n - |alle "interessanten" Gene|

* Interessant = Gene mit einer bestimmten Eigenschaft,
z.B. "differentiell exprimiert", "positiv selektiert", …

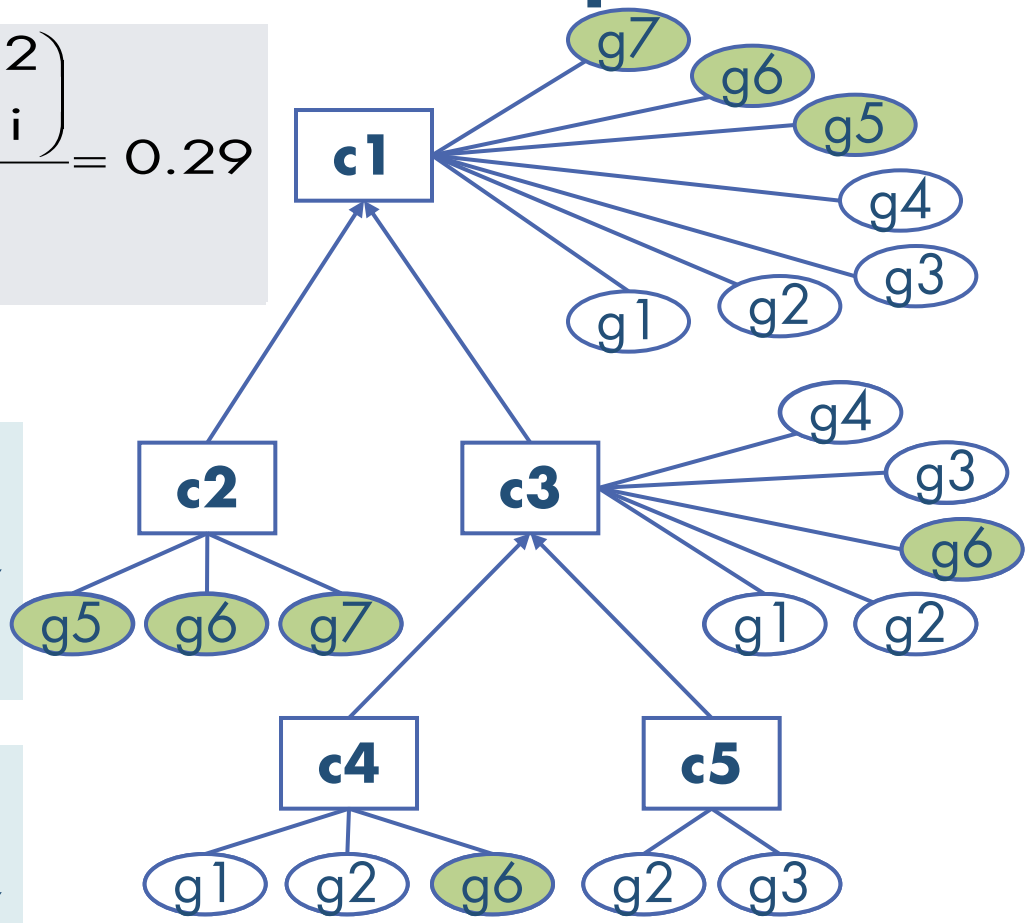


Binomialkoeffizient

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}.$$

Term Enrichment - Beispiel

$$1-\sum_{i=0}^{o}\frac{\binom{2}{i}\binom{7-2}{3-i}}{\binom{7}{3}}=0.29$$



A = 7
a = 3
N = 7
n = 3

p = 1

A = 3
a = 3
N = 7
n = 3

p = 0.03

A = 5
a = 1
N = 7
n = 3

p = 0.14

A = 3
a = 1
N = 7
n = 3

p = 0.51

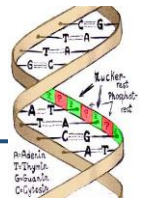
A = 2
a = 0
N = 7
n = 3

p = 0.29

g — Gene mit der „interessanten“ Eigenschaft

g — Gene ohne die „interessante“ Eigenschaft

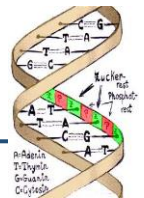
g1 g2 g3 g4 g5 g6 g7



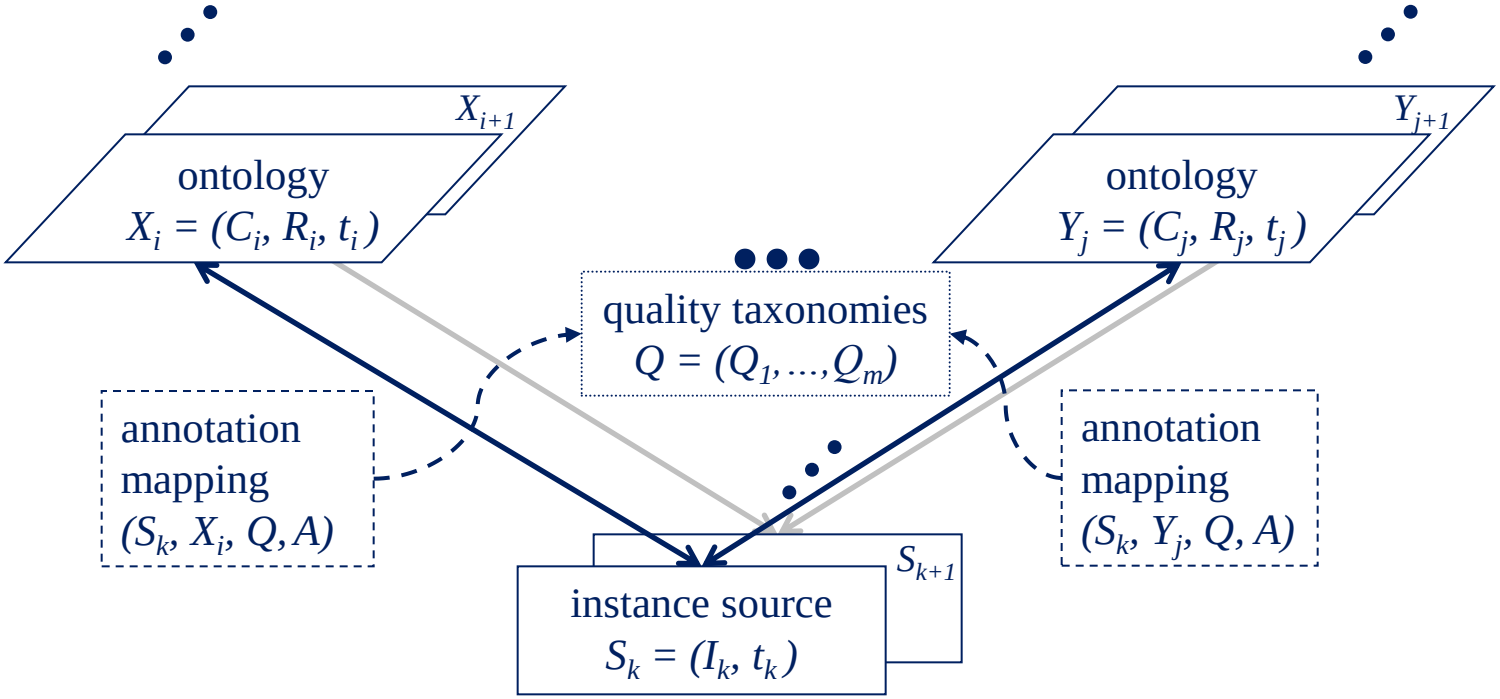
GO Slim-Terme

- "Abgespeckte" Versionen der GO Subontologien, enthalten eine Untermenge der GO Terme
- Geben einen Überblick zum Ontologieinhalt
- Keine Details aus feingranularen Termen
- Beispiel: GO:0002376 - immune system process
- Nutzen: Zusammenfassung der Resultate, wenn eine weitgefasste Klassifikation der Genproduktfunktion notwendig ist
    - GO-Annotation eines Genoms, Microarray, cDNA Sammlung
- Entsprechend der Bedürfnisse durch Nutzer selbst erstellt, z.B. spezifisch für bestimmte Spezies oder für bestimmte Bereiche der Ontologie
- GO stellt generische, nicht spezies-spezifische GO Slim-Terme zur Verfügung (~150)
- Außerdem stehen Modellorganismen-spezifische Slim-Terme zur Verfügung

- http://www.geneontology.org/GO.slims.shtml



Annotationsmodell mit Versionierung



$O_v = (C_v, R_v, t_v)$
v - Versionsnummer
t – release timestamp

Evolutionsanalyse von Annotationen*

- Vergleich von zwei großen Datenquellen im Zeitraum März 2004 – Dezember 2008
- GO Annotationen für humane Proteine
- Analyse der Annotationshäufigkeit in bestimmten Evidence-Gruppen über die Zeit / Versionshistorie

Ensembl v_{31}–v_{52}

Ensembl

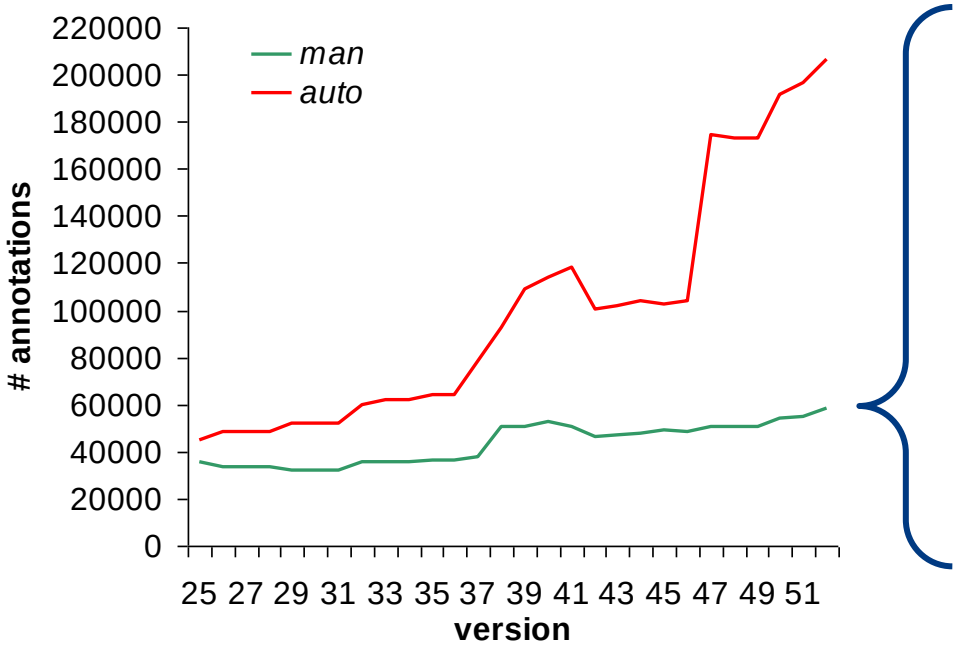
Swiss-Prot v_{47}–v_{56}

SIB swissprot EBI

* Groß, A.; Hartung, M.; Kirsten, T.; Rahm, E.: Estimating the Quality of Ontology-based Annotations by Considering Evolutionary Changes, Proc. 6th Data Integration in the Life Sciences (DILS) Conf., 2009

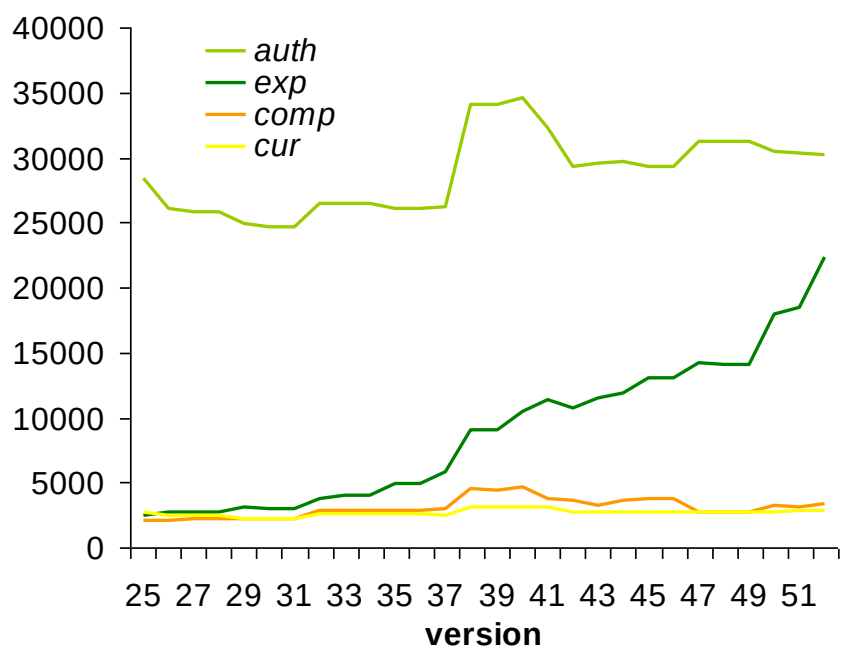
Quantitative Analyse über die Zeit (Ensembl)

Manuelle vs. automatische Annotationen



- 78% von 265,000 automatisch zugeordnet
- Auch Löschungen von Annotationen treten auf

Nur manuelle Annotationen



- 22% manuelle Annotationen

Fragen ?