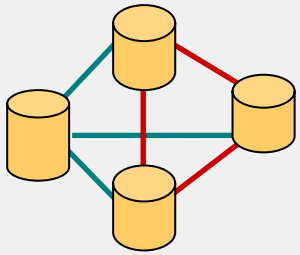
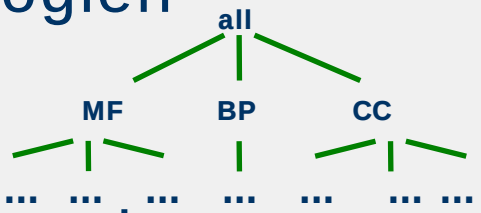
Untersuchung von Annotations- und Ontologie-Mappings

Motivation

- Sehr große, zunehmende Menge an Daten und Datenquellen in den Biowissenschaften



→Wissensstrukturierung mit Hilfe von Ontologien z.B. Gene Ontology (GO)



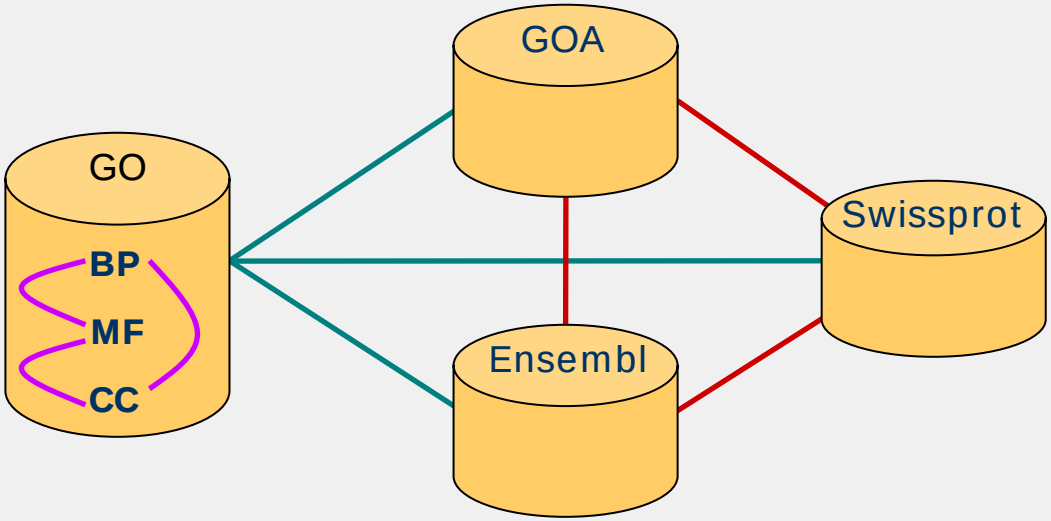
- Untersuchung von Instanz-, Annotations- und Ontologie-Mappings

- Probleme:
  - Unvollständige Annotationen
  - Fehlerhafte Annotationen
  - Versionierung von Datenquellen → inkonsistente Daten

Übersicht

- Einführung - Annotationen und Evidence Codes
- Untersuchung existierender Annotations-Mappings zur Berechnung neuer und verbesserter Ontologie-Mappings
  1) Vergleich unterschiedlicher Annotations-Quellen (Ensembl, Swissprot)
  2) Untersuchung der Evolution von Annotations-Mappings
- Erstellung neuer Annotations-Mappings für bisher nicht annotierte biologische Objekte (miRNAs)
- Zusammenfassung
- Ausblick

Annotationsmappings und Datenquellen



__ Instanzmapping
__ Annotationsmapping
__ Ontologiemapping

Beispiel: Annotation aus GOA

TAP1_HUMAN

TAP1, ABCB2, PSF1, RING4, Y3: Antigen peptide transporter 1

Accession, Term	Ontology	Evidence	Reference	Assigned by
GO:0016021 : integral to membrane	cellular component	NAS	PMID:1946428	UniProtKB
GO:0005515 : protein binding	molecular function	IPI With UniProtKB:Q03519	PMID:17055437	IntAct (via UniProtKB)
GO:0046982 : protein heterodimerization activity	molecular function	IPI With UniProtKB:Q03519	PMID:11133832	UniProtKB

Gene Ontology - Evidence Codes (ECs)

Evidence Code* = gibt an, auf welchem Experiment- bzw. welcher Analyseart eine Annotation beruht

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

Curator-assigned ECs

Experimental ECs	Computational Analysis ECs	Author Statement ECs	Curator Statement ECs
EXP	ISS	TAS	IC
IDA	ISO	NAS	ND
IPI	ISA		
IMP	ISM		
IGI	IGC		
IEP	RCA		

Automatically-assigned ECs

IEA

Obsolete ECs

NR

CURATED*

NOT CURATED

Engl.: *evidence: Beleg, Beweis; **curated ≈ geprüft, kontrolliert, abgesegnet

Evidence Codes – Beispiel IEA

- **IEA (Inferred from Electronic Annotation)**
  - Automatische Annotation über Algorithmus, keine Kontrolle durch einen „Curator“

    z.B. Sequenz-Ähnlichkeit
    Keyword-Mapping
  - Abgrenzung von ISS (Inferred from Sequence Similarity)

    Annotationen, welche auf Sequenz-ähnlichkeit beruhen und von einem Curator geprüft wurden, sollten auf ISS geändert werden

Evidence Codes

WICHTIG

- ! Kein Maß für die Qualität der Annotation
- ! Keine zwingende Klassifizierung von Experimenten
- ! Annotationen mit gleichen GO-IDs und unterschiedlicher Evidence möglich, falls die gleiche Information durch multiple Methoden erhalten wurde

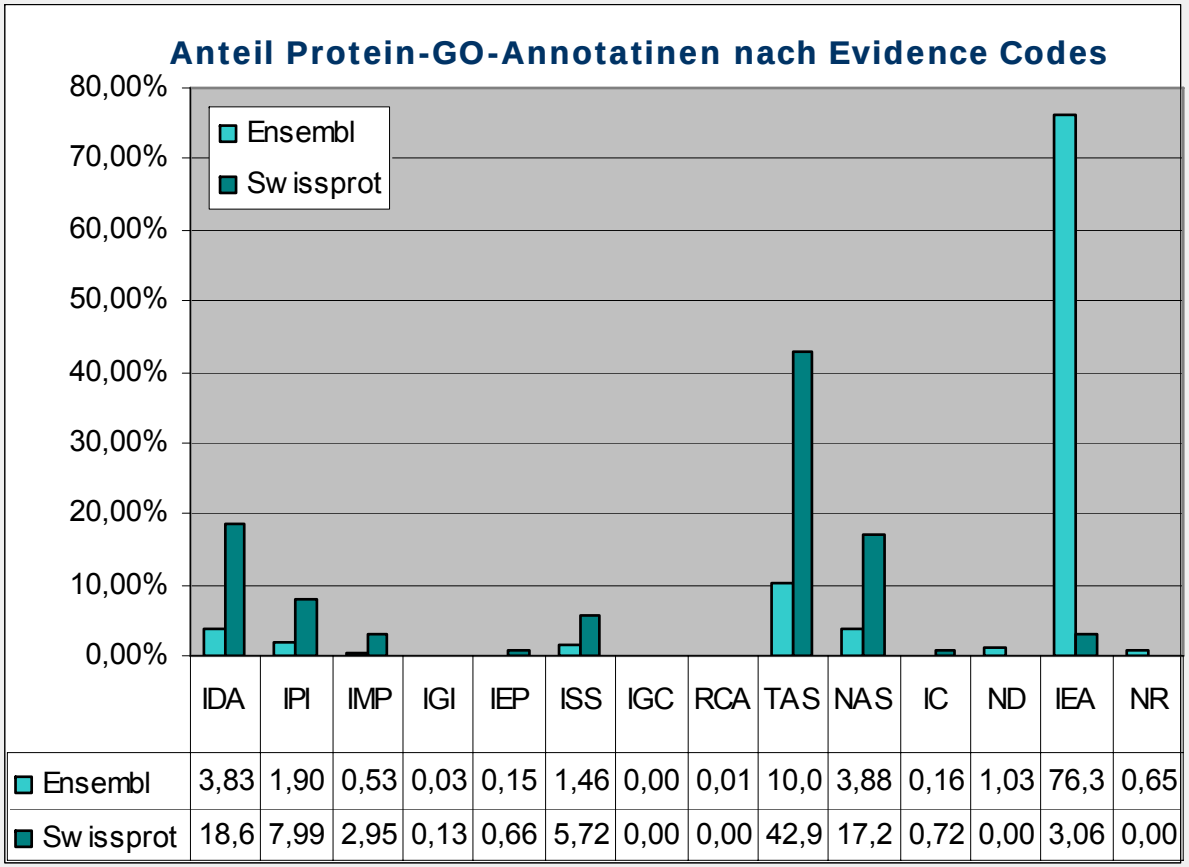
Übersicht

- Einführung - Annotationen und Evidence Codes
- Untersuchung existierender Annotations-Mappings zur Berechnung neuer und verbesserter Ontologie-Mappings
  1) Vergleich unterschiedlicher Annotations-Quellen (Ensembl, Swissprot)
  2) Untersuchung der Evolution von Annotations-Mappings
- Erstellung neuer Annotations-Mappings für bisher nicht annotierte biologische Objekte (z.B. miRNAs)
- Zusammenfassung
- Ausblick

Vergleich von Swissprot und Ensembl

- Fragestellungen
  - Daten höherer Qualität aus Swissprot gewinnen? (Mehr „curated“ statt „not curated“?)
  - Gewinnen zusätzlicher Informationen aus Swissprot im Vergleich zu Ensembl?
  - Ähnliche Ergebnisse für Ontologiemappings mit einer anderen Annotationsquelle als bisher (Swissprot statt Ensembl)?

Vergleich von Swissprot und Ensembl



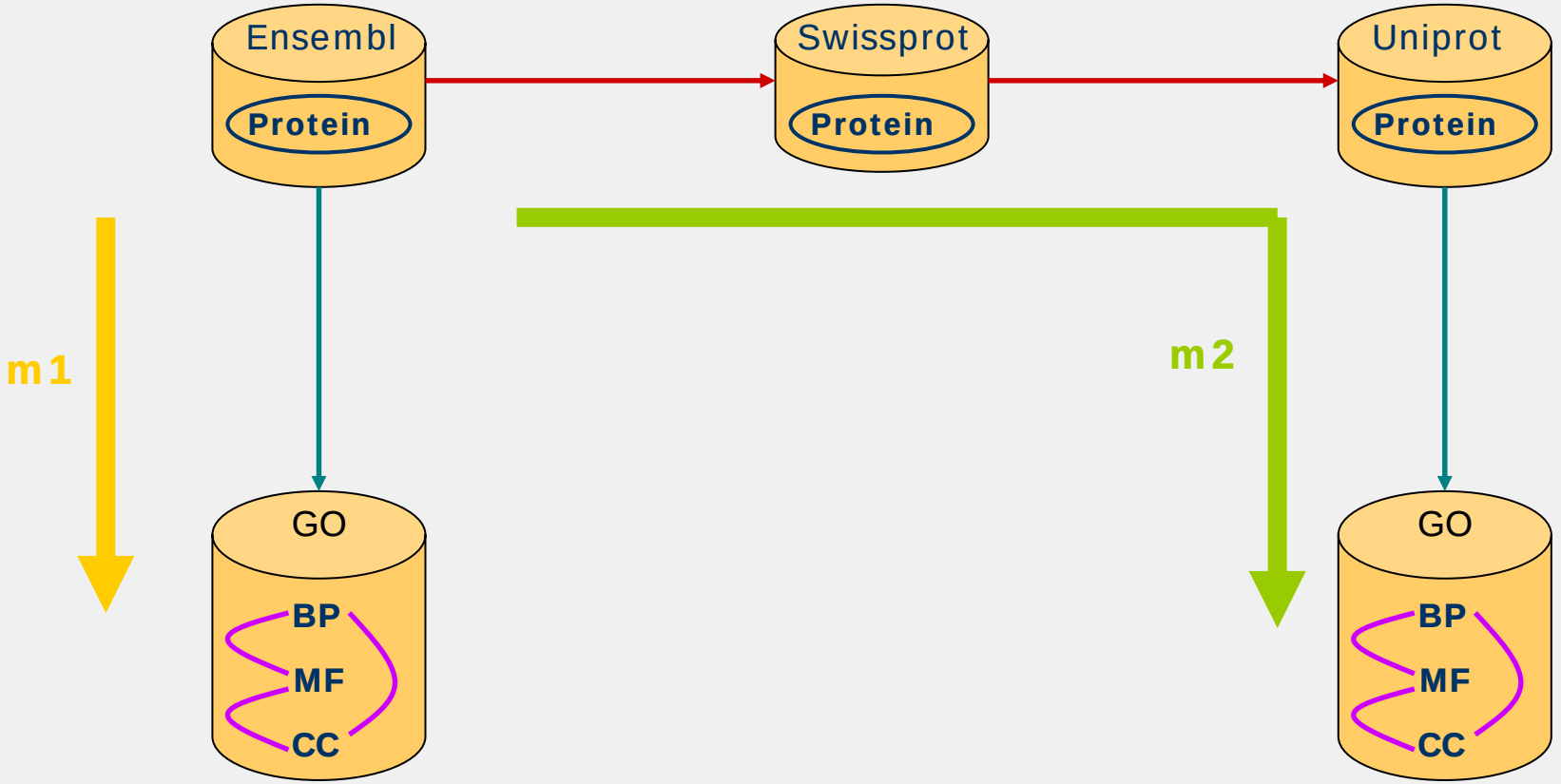
Absolute Anzahl Annotationen

Evidence Code	Ensembl	Uniprot
IDA	8746	7717
IPI	4340	3309
IMP	1222	1220
IGI	75	54
IEP	346	275
ISS	3332	2370
IGC	0	0
RCA	17	0
TAS	23020	17771
NAS	8857	7120
IC	357	298
ND	2359	0
IEA	174359	1265
NR	1488	0
Summe	228518	41399
Summe ohne IEA,NR	52671	40134

- Swissprot: deutlich höherer Anteil „curated“ (insbes. IDA, TAS, NAS)
- Ensembl: größter Anteil automatische Annotationen (IEA)
- Absolute Anzahl ist ähnlich (außer IEA) → unterschiedliche Informationen?

Mapping-Szenario

- Informationsgewinn bezüglich der Annotations-Mappings in Swissprot im Vergleich zu Ensembl?



	m1	m2	$Info_{same}$	$Info_{new}$
Protein-BP	76297	20972	19294	1678
Domain	28021	9597	9564	33
Range	3585	2532	2427	105
Protein-MF	93837	17097	15728	1369
Domain	31905	10660	10607	53
Range	2641	1962	1917	45
Protein-CC	56554	12762	11238	1524
Domain	26744	9360	9322	38
Range	681	568	547	21

$Info_{same}$ = intersect(m1, m2)

$Info_{new}$ = m2 - $Info_{same}$

Informationsgewinn?

- Kann auch durch Versionsunterschiede entstanden sein (Ensembl 47 Oktober 2007, Swissprot 55 März 2008)
- Ensembl integriert u.a. Informationen aus Swissprot

Ontologiemappings

- größter Teil Korrespondenzen aus Swissprot auch in Ensembl enthalten
- Ensembl 2-3fache Anzahl der Korrespondenzen im Vgl. zu Swissprot (nur „curated“-Annotationen, base5, min, dice)

Übersicht

- Einführung - Annotationen und Evidence Codes
- Untersuchung existierender Annotations-Mappings zur Berechnung neuer und verbesserter Ontologie-Mappings
  1) Vergleich unterschiedlicher Annotations-Quellen (Ensembl, Swissprot)
  2) Untersuchung der Evolution von Annotations-Mappings
- Erstellung neuer Annotations-Mappings für bisher nicht annotierte biologische Objekte (z.B. miRNAs)
- Zusammenfassung
- Ausblick

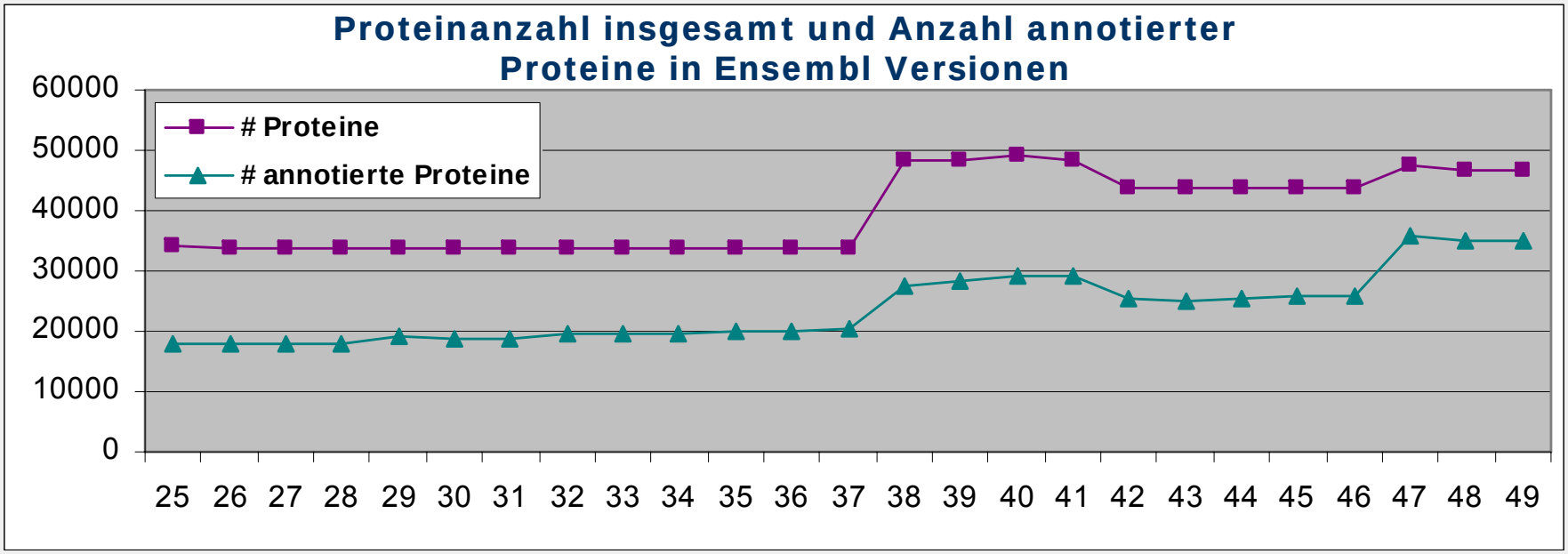
Untersuchung existierender Annotations-Mappings

- Untersuchung der Annotations-Mappings in Ensembl Versionen 25-49
  (Mapping-Evolution Oktober 2004 - März 2008)
- Berechnung von Hinzufügungen, Löschungen, Migrationen
- Einbeziehen der Informationen aus Evidence Codes

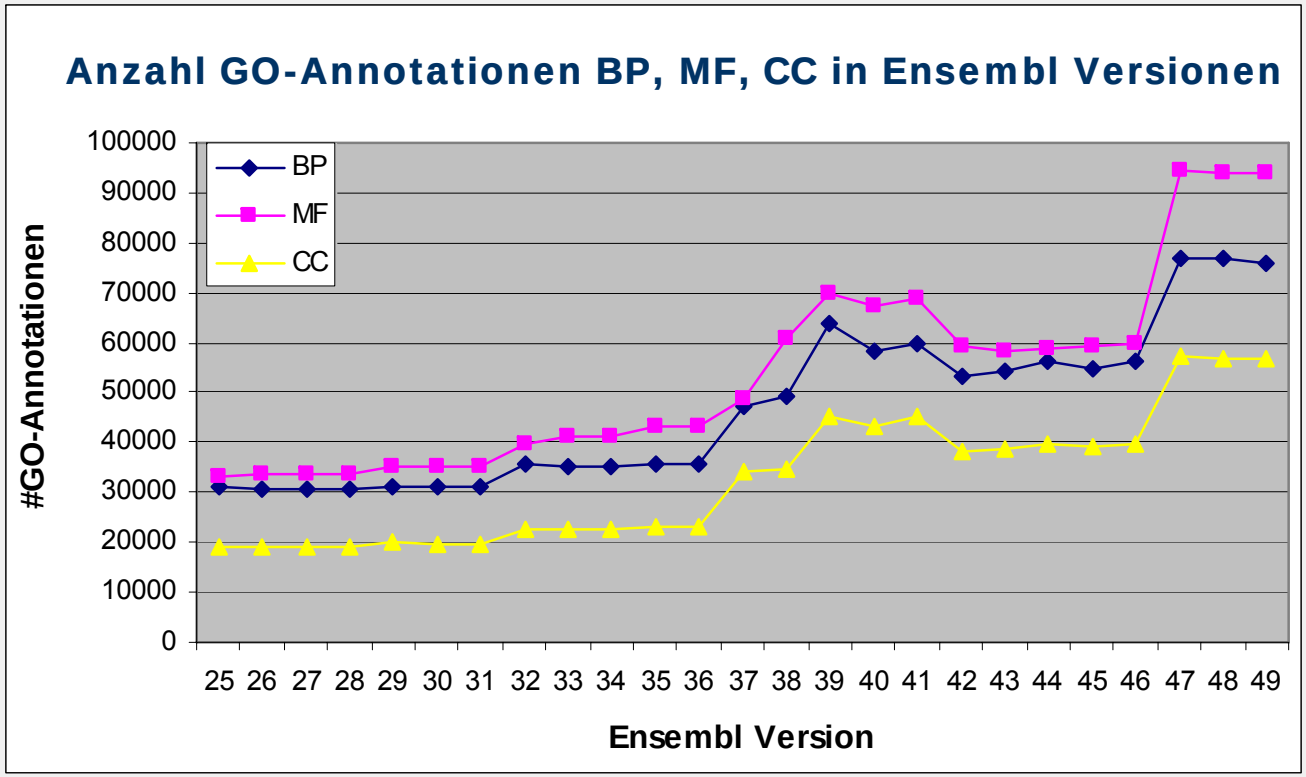
Ziele

- Erreichen stabilerer Mappings
- Quantitatives und qualitatives Aufbessern der Annotations-Mappings

  → Bessere Ontologie-Mappings



- Version 25 (Okt 2004) 52,01% der Proteine annotiert
- Version 49 (März 2008) 75,48% der Proteine annotiert



- Höchste Anzahl Annotationen: Molekulare Funktionen
- Zuwachs insgesamt
  CC 2,99 MF 2,82 BP 2,46

Gene Ontology - Evidence Codes (ECs)

- Zur Erinnerung – Evidence Code Gruppen

Curator-assigned ECs

Experimental ECs	Computational Analysis ECs	Author Statement ECs	Curator Statement ECs
EXP	ISS	TAS	IC
IDA	ISO	NAS	ND
IPI	ISA		
IMP	ISM		
IGI	IGC		
IEP	RCA		

Automatically-assigned ECs

IEA

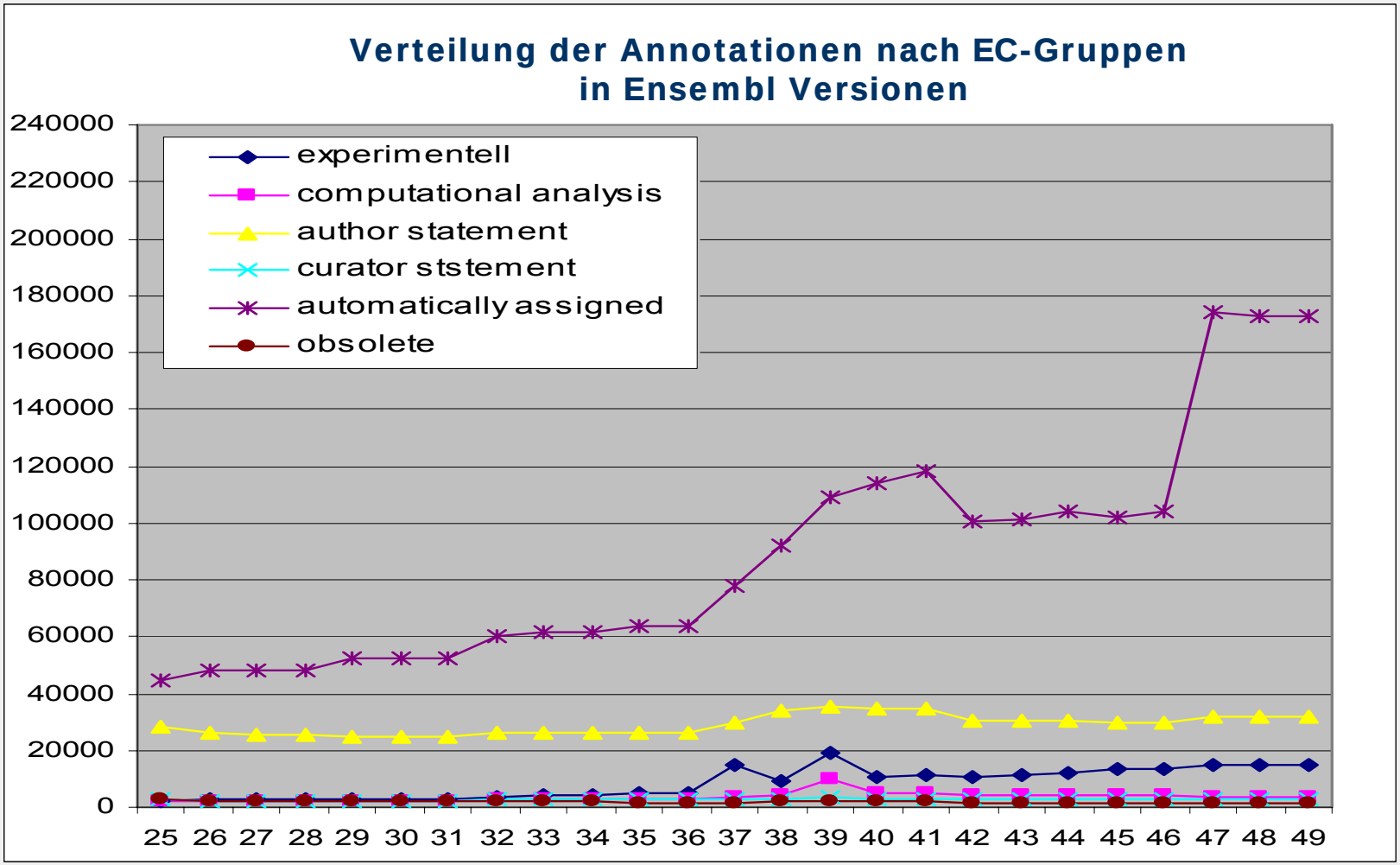
Obsolete ECs

NR

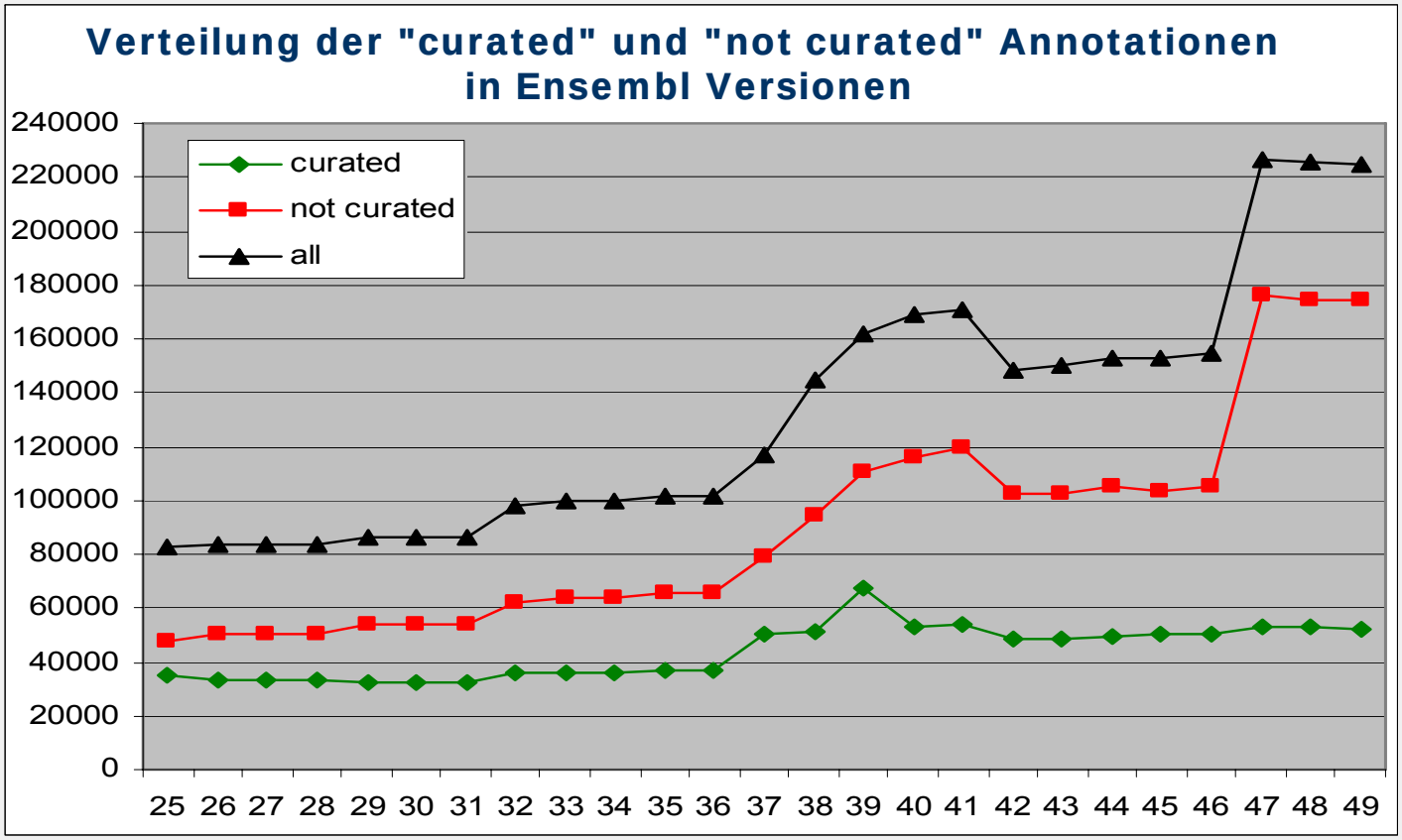
CURATED*

NOT CURATED

Engl.: *evidence: Beleg, Beweis; **curated ≈ geprüft, kontrolliert, abgesegnet



- Häufigste Annotationen und Größter Zuwachs: “automatically assigned” (IEA)



- Version 25-49 Zuwachs um Faktor

“not curated”	3,7
“curated”	1,5
“all”	2,7

- Starker Zuwachs bei “not curated” beeinflusst deutlich das Wachstum der Gesamt-Annotationsanzahl

- Bisher noch keine vollständige Auswertung und Darstellung der Add-, Delete- und Migrationsraten über alle Versionen
- Interessant:
  - Gibt es Migrationen von “not curated” nach “curated”?
  - Gibt es Migrationen innerhalb der “curated”-Annotationen?

Veränderungen von Version 46 zu Version 47

Beispiel: Biologische Prozesse, „curated“ Annotationen

	46	47	Intersect (46,47)	Deletes	Adds	Delete fraction	Add fraction
experimental	3742	4254	3227	515	1027	13,76 %	24,14 %
computational	2020	1517	1116	904	401	44,75 %	26,43 %
author statement	14126	15062	12084	2042	2978	14,46 %	19,77 %
curator statement	847	877	716	131	161	15,47 %	18,36 %
curated	20735	21710	17344	3391	4366	16,35 %	20,11 %
				3592			

Wieso entspricht die Summe der Deletes nicht der Gesamt-Delete-Anzahl in der „curated“-Gruppe?

46 \ 47	exp	comp	authst	curast
exp		29	10	0
comp	27		102	1
authst	25	6		1
curast	0	0	0	

exp = experimentell
comp = computational
authst = author statement
curast = curator statement

29 „experimentell“-BP-Annotationen in Version 46 haben in Version 47 einen „computational“ Evidence Code
→ECs werden bei gleich bleibender Annotation von Version zu Version geändert = Migration

Unstimmigkeiten von Version 37 bis Version 49

Addition der EC-Gruppen ergeben anderes Ergebnis als Union

→ Existieren gleiche Annotationen mit verschiedenen ECs ("Duplikat"-Problem)

	48	49
EXP	0	0
IDA	8749	8743
IPI	4339	4339
IMP	1224	1221
IGI	78	78
IEP	346	343
ISS	3337	3333
ISO	0	0
ISA	0	0
ISM	0	0
IGC	0	0
RCA	17	17
TAS	23007	22855
NAS	8849	8789
IC	357	357
ND	2345	2345
IEA	173021	172648
NR	1487	1477
Summe	**227156**	**226545**
Union all	225240	224633
Differenz	**1916**	**1912**

Protein	GO-Term	EC
ENSP00000375699	GO:0005515	IPI
ENSP00000375699	GO:0005515	IEA
ENSP00000375699	GO:0005634	IEA
ENSP00000375699	GO:0005635	IDA
ENSP00000375699	GO:0005764	IEA
ENSP00000375699	GO:0006355	IEA
ENSP00000375699	GO:0007166	ISS
ENSP00000375699	GO:0008285	ISS
ENSP00000375699	GO:0008285	IEA
ENSP00000375699	GO:0009072	IEA
ENSP00000375699	GO:0009755	NAS

"curated" und "not curated" Evidence für die gleiche Annotation sollten nicht existieren

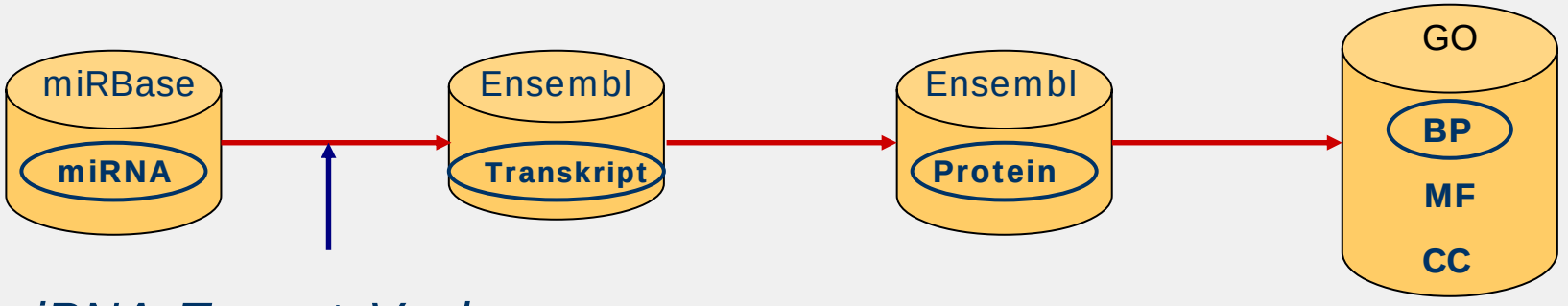
- Erkenntnisse
  - Neben Adds und Deletes existieren Migrationen
    - Innerhalb “curated”
    - Zwischen “not curated” und “curated”
  - Duplikatproblem: gleiche Annotation unterschiedliche Evidence Codes

- Konsequenzen für Ontologie-Mappings
  - “Not curated”-Annotationen nicht für Mappings verwenden
  - Migrationen beachten
  - Duplikate weglassen oder für höher qualitativen EC entscheiden?

Übersicht

- Einführung - Annotationen und Evidence Codes
- Untersuchung existierender Annotations-Mappings zur Berechnung neuer und verbesserter Ontologie-Mappings
  1) Vergleich unterschiedlicher Annotations-Quellen (Ensembl, Swissprot)
  2) Untersuchung der Evolution von Annotations-Mappings
- Erstellung neuer Annotations-Mappings für bisher nicht annotierte biologische Objekte (z.B. miRNAs)
- Zusammenfassung
- Ausblick

Erstellung neuer Annotationsmappings



miRNA-Target-Vorhersage

Algorithmen z.B.
PicTar 4,5
miRanda(miRBase)
TargetScanS
DIANA-microT
miRanda (microRNA.org)

Hohe Falsch-Positiv-Rate

miRNA-GO-Term-Mappings

Filter

→ Mehrheitsentscheid (mind.2-3 Programme sagen Target vorher)

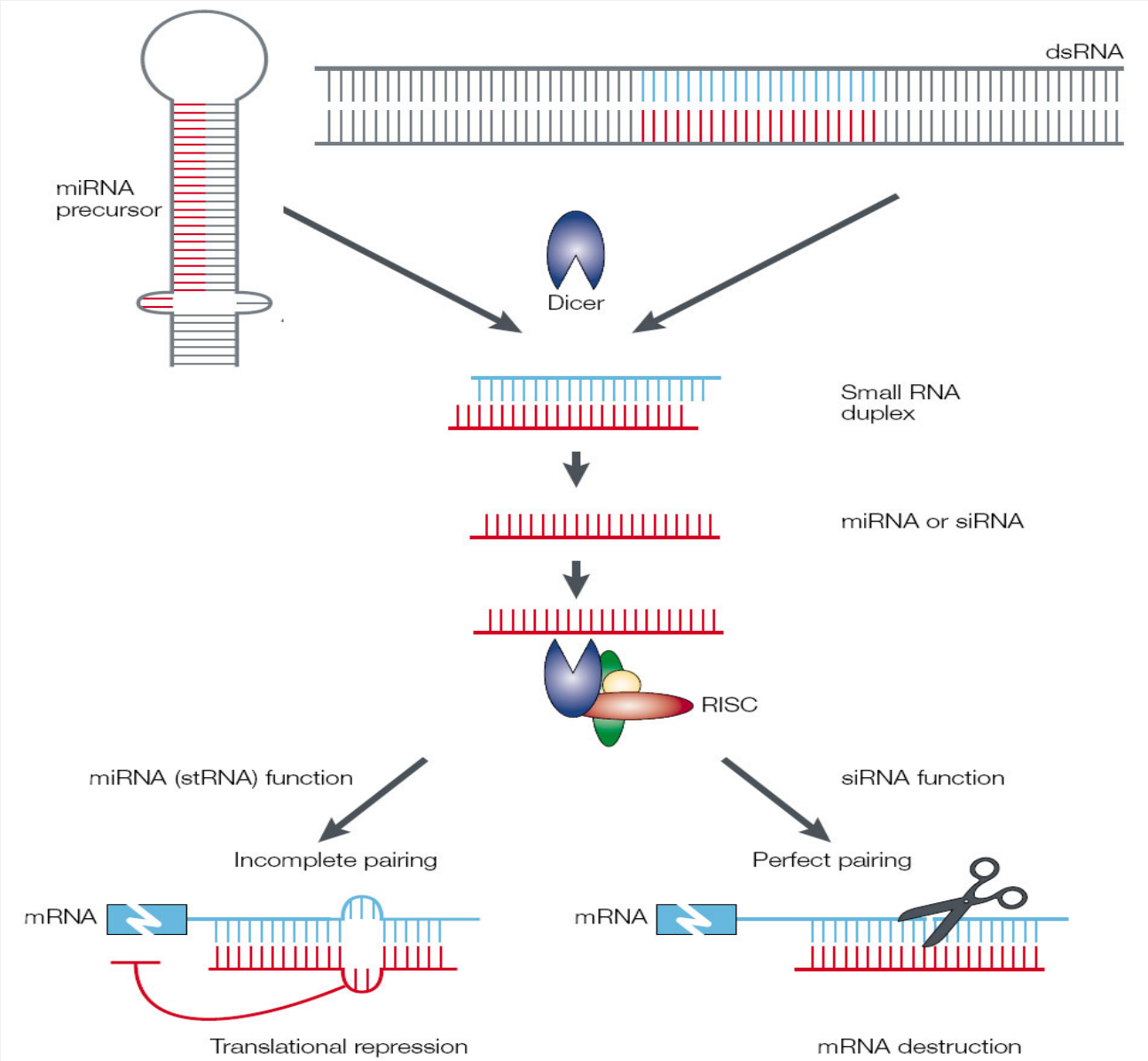
→ Nur Targets mit hohem Evolutions-Konservierungs-score (Orientierung an TarBase)

Filter

→ Occurrence-Filter

→ „Anzahl miRNA“ pro Konzept (GO-Konzepte, die bei allen miRNAs vorkommen sind nicht interessant)

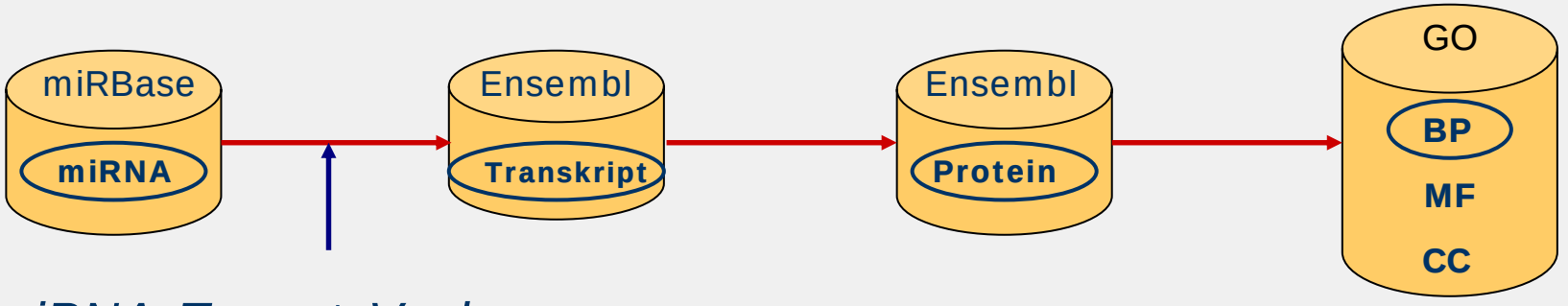
Was ist eine miRNA?



- Kleine biologische Moleküle
- Einfluss auf die Regulation der Translation

→Kontrolle der Genexpression

Erstellung neuer Annotationsmappings



miRNA-Target-Vorhersage

Algorithmen z.B.
PicTar 4,5
miRanda(miRBase)
TargetScanS
DIANA-microT
miRanda (microRNA.org)

Hohe Falsch-Positiv-Rate

miRNA-GO-Term-Mappings

Filter

→ Mehrheitsentscheid (mind.2-3 Programme sagen Target vorher)

→ Nur Targets mit hohem Evolutions-Konservierungs-score (Orientierung an TarBase)

Filter

→ Occurrence-Filter

→ „Anzahl miRNA“ pro Konzept (GO-Konzepte, die bei allen miRNAs vorkommen sind nicht interessant)

miRNA-GO-Mapping – erste Ergebnisse

- Bisher konzeptionell, einige erste Berechnungen
  - Automatische Annotation von 711 humanen miRNAs mit Biologischen Prozessen
  - Bisher ein Algorithmus: miRanda(miRBase)
  - Dabei nur Verwendung von „curated“-Annotationen
  - Einsatz der Filter (Konservierungs-Score, Occurrence, „Anzahl miRNAs“)

	curated BP-miRNA Annotationen	occurrence >5	occurence > 5 cnt_miRNA < 200
ohne Konservierungs-Filter	207844	14570	4528
Konservierungs-Filter 0.588 (Durchschnitt TarBase)	65007	1169	755
Konservierungs-Filter 0.9	44014	453	243

Übersicht

- Einführung
  - Annotationen und Evidence Codes
- Untersuchung existierender Annotations-Mappings zur Berechnung neuer und verbesserter Ontologie-Mappings
  - Vergleich unterschiedlicher Annotations-Quellen (Ensembl, Swissprot)
  - Untersuchung der Evolution von Annotations-Mappings
- Erstellung neuer Annotations-Mappings für bisher nicht annotierte biologische Objekte (z.B. miRNAs)
- Zusammenfassung
- Ausblick

Zusammenfassung / Schlussfolgerungen

- Vergleich von Swissprot und Ensembl
  - kein deutlicher Informationsgewinn aus Swissprot
  - wichtig: mit „passenden“ Versionen arbeiten
- Evolutionäre Analyse der Annotations-Mappings der Datenquelle Ensembl
  - Bedeutung und Einfluss der Evidence Codes auf Annotations- und Ontologie-Mappings
- Annotation biologischer Objekte (miRNAs)
- Generell
  Für automatische Berechnungen sollten automatische(„IEA“) oder veraltete(„NR“) Annotationen nicht verwendet werden

Übersicht

- Einführung
  - Annotationen und Evidence Codes
- Untersuchung existierender Annotations-Mappings zur Berechnung neuer und verbesserter Ontologie-Mappings
  - Vergleich unterschiedlicher Annotations-Quellen (Ensembl, Swissprot)
  - Untersuchung der Evolution von Annotations-Mappings
- Erstellung neuer Annotations-Mappings für bisher nicht annotierte biologische Objekte (z.B. miRNAs)
- Zusammenfassung
- Ausblick

Ausblick

- Vervollständigen der Evolutionsanalyse
  - Ontologie-Mappings unter Beachtung der verbesserten Annotations-Mappings (Evidence Codes)
- Annotation der miRNAs
  - Weitere Algorithmen (z.B. PicTar4,5) verwenden
  - Verschiedene Einstellungen der Filter prüfen

<u>Weiter blickend</u>

- Andere Quellen z.B. GOA einbeziehen
- Bei den Mappings bisher Instanzdaten „allein“ betrachtet

  Aber: In Wirklichkeit existieren Homologie- und Interaktionsbeziehungen zwischen biologischen Objekten
- Untersuchung Struktur von Mappings (Hub-Konzepte und ihr Einfluss)

VIELEN DANK FÜR EURE/IHRE AUFMERKSAMKEIT