

REVERSE A2

Linköpings universitet

28.11.2007

1. Ontology alignment

Patrick Lambrich
Linköpings universitet

GENE ONTOLOGY (GO)	SIGNAL-ONTOLOGY (SigO)
immune response	Immune Response
- acute phase response	- Allergic Response
- anaphylaxis	- Antigen Processing and Presentation
- antigen presentation	- B Cell Activation
- antigen processing	- B Cell Development
- cellular defense response	- Complement Signaling
- cytokine metabolism	synonym complement activation
- cytokine biosynthesis	- Cytokine Response
synonym cytokine production	- Immune Suppression
...	- Inflammation
p- regulation of cytokine biosynthesis	- Intestinal Immunity
...	- Leukotriene Response
- B-cell activation	- Leukotriene Metabolism
- B-cell differentiation	- Natural Killer Cell Response
- B-cell proliferation	- T Cell Activation
- cellular defense response	- T Cell Development
...	- T Cell Selection in Thymus
- T-cell activation	
- activation of natural killer cell activity	
...	

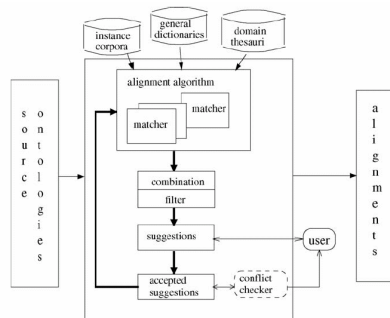
Ontology Alignment

GENE ONTOLOGY (GO)	SIGNAL-ONTOLOGY (SigO)
immune response	Immune Response
- acute phase response	- Allergic Response
- anaphylaxis	- Antigen Processing and Presentation
- antigen presentation	- B Cell Activation
- antigen processing	- B Cell Development
- cellular defense response	- Complement Signaling
- cytokine metabolism	synonym complement activation
- cytokine biosynthesis	- Cytokine Response
synonym cytokine production	- Immune Suppression
...	- Inflammation
p- regulation of cytokine biosynthesis	- Intestinal Immunity
...	- Leukotriene Response
- B-cell activation	- Leukotriene Metabolism
- B-cell differentiation	- Natural Killer Cell Response
- B-cell proliferation	- T Cell Activation
- cellular defense response	- T Cell Development
...	- T Cell Selection in Thymus
- T-cell activation	
- activation of natural killer cell activity	
...	

Legend:
- - - - - equivalent concepts
- equivalent relations
- is-a relation

- To define the relations between the terms in different ontologies

An Alignment Framework



Ontology alignment system - SAMBO

Example alignment system SAMBO – matchers, combination, filter

Align Concept in *Ontology-1* and *Ontology-2*

Terminology
 Term. + WordNet
 Domain (UMLS)
 Learning
 Structure

matchers: Terminology Term. + WordNet Domain (UMLS) Learning Structure

threshold:

Example alignment system SAMBO – suggestion mode

nose_MA	nose_MeSH
nasal_cavity_epithelium	nasal_mucosa
definition: MA:0001324	definition: MESH:A.04.531.520
synonym: nasal mucosa	synonym: nasal epithelium
part-of: nasal_cavity	part-of:

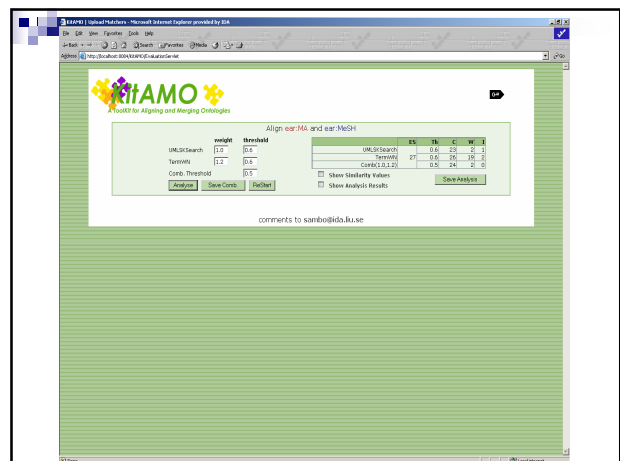
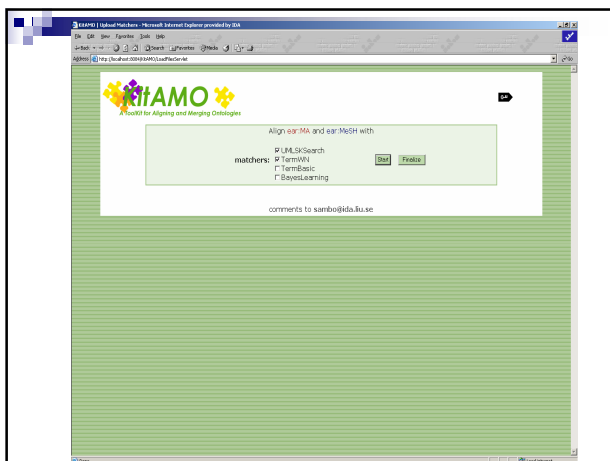
nasal_cavity_epithelium
nasal_mucosa

new name for the equivalent concepts:

Example alignment system SAMBO – manual mode

nose_MA	nose_MeSH
<ul style="list-style-type: none"> O nasalis O external_naris O internal_naris O nasal_capsule M nasal_cavity (nasal_cavity) O nasal_cavity_epithelium O nasal_septum O nasal_turbinate O olfactory_gland O olfactory_nerves O vomeronasal_organ 	<ul style="list-style-type: none"> O nasal_bone M nasal_cavity (nasal_cavity) O nasal_mucosa O olfactory_mucosa O olfact_loo1 O olfactory_receptor_neuron O nasal_septum O paranasal_sinus O turbinate

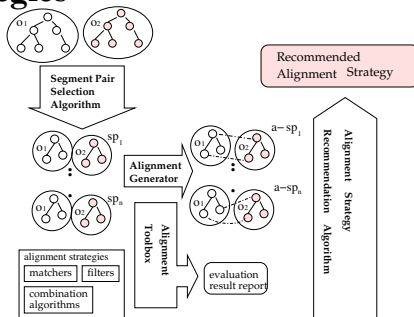
System for evaluation
of ontology alignment
strategies - KitAMO



NIA	MESH	UMLS	Search	Term	WN	(L,DTW)	Sim
basilar membrane	basilar membrane	1.0000	1.0000	1.0000			
tectorial membrane	tectorial membrane	1.0000	1.0000	1.0000			
stapes	stapes	1.0000	1.0000	1.0000			
scala tympani	scala tympani	1.0000	1.0000	1.0000			
vestibular aqueduct	vestibular aqueduct	1.0000	1.0000	1.0000			
utricle	sacule and utricle	1.0000	1.0000	1.0000	W		
tensor tympani	tensor tympani	1.0000	1.0000	1.0000			
middle ear	middle ear	1.0000	1.0000	1.0000			
ear	ear	1.0000	1.0000	1.0000			
spiral organ	organ of corti	1.0000	1.0000	1.0000			
tympanic membrane	tympanic membrane	1.0000	1.0000	1.0000			
auditory bone	ear ossicle	1.0000	1.0000	1.0000			
cochlea	cochlea	1.0000	1.0000	1.0000			
sacule	sacule and utricle	1.0000	1.0000	1.0000	W		
incus	incus	1.0000	1.0000	1.0000	C		
auditory tube	eustachian tube	1.0000	1.0000	1.0000			
vestibular apparatus	vestibule	1.0000	1.0000	1.0000			
endolymphatic duct	endolymphatic duct	1.0000	1.0000	1.0000			
stria vascularis	stria vascularis	1.0000	1.0000	1.0000			
cochlear duct	cochlear duct	1.0000	1.0000	1.0000			
outer ear	external ear	1.0000	1.0000	1.0000			
inner ear	labyrinth	1.0000	1.0000	1.0000			
semicircular canal	semicircular canal	1.0000	1.0000	1.0000			
endolymphatic sac	endolymphatic sac	1.0000	1.0000	1.0000			
external acoustic meatus	ear canal	0.6000	0.7360	0.6742			
malleus	malleus	0.0000	0.9235	0.5937			
stapes	stapes	0.0000	0.9095	0.4961			
perilymphatic channel	cochlear aqueduct	0.0000	0.8554	0.4666			

	Th	C	W	I
(L,DTW)	0.50	23	2	0
(L,DTW)	0.50	24	2	0
TermWN	0.40	26	110	19
TermWN	0.50	26	65	8
TermWN	0.60	26	19	2
UMLSSearch	0.40	23	2	1
UMLSSearch	0.50	23	2	1
UMLSSearch	0.60	23	2	1

Recommendation of alignment strategies



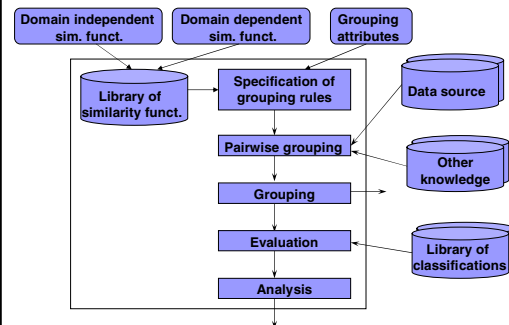
Further reading

- Lambrix, Tan, SAMBO – a system for aligning and merging biomedical ontologies, *Journal of Web Semantics*, 4(3):196-206, 2006.
- Lambrix, Tan, A tool for evaluating ontology alignment strategies, *Journal on Data Semantics*, VIII:182-202, 2007.
- Lambrix P, Tan H, Ontology alignment and merging, chapter 6 in Burger, Davidson, Baldock, (eds), *Anatomy Ontologies for Bioinformatics: Principles and Practice*, pp 133-150, Springer, 2007.
- Tan H, Lambrix P, 'A method for recommending ontology alignment strategies', *International Semantic Web Conference*, 494-507, 2007.

2. Evaluation of grouping strategies for biological data

Patrick Lambrix
Linköpings universitet

Method for similarity-based grouping





■ A toolKit for Evaluating Grouping Algorithms

Selection of test case

Data source: Glyc-Funct-AnnEc-onlyGO (DS3)

Grouping rule: SemSim(GOcomb)>0.95

Grouping method: ConnectedComponents

Evaluation method:

- Entropy
- Purity
- MutualInformation
- FMeasure

Source of classes: Glycolysis: by function

next

Specification of grouping rules

Data source: Glyc-Funct-AnnEc-onlyGO (DS3)

Grouping rule: SemSim(GOcomb)>0.95

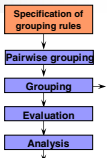
Grouping method: ConnectedComponents

Evaluation method:

- Entropy
- Purity
- MutualInformation
- FMeasure

Source of classes: Glycolysis: by function

next

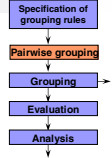


Pairwise grouping

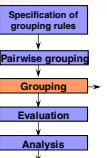
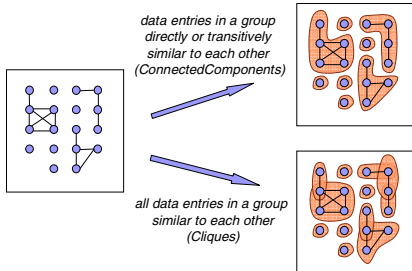


Grouping rule: SemSim(GOcomb)>0.95

Data source: Glyc-Funct-AnnEc-onlyGO (DS3)



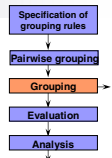
Grouping



Grouping

Glyc-Funct-AnnEc-onlyGO - SemSim(GOcomb)>0.95 - ConnectedComponents - Glycolysis: by function

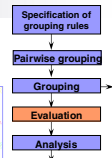
GroupNr	ClassNr	ID	Definition	GO combined
0	4	P60174	Trisophosphate isomerase (TIM) (Trisophosphate isomerase)	go:0004807
0	4	NP_000356	trisophosphate isomerase 1 [Homo sapiens]	go:0016853, go:0004807
1	2	AAA46068	phosphofruktokinase	go:0003872
1	2	NP_001002002	liver phosphofruktokinase isoform a [Homo sapiens]	go:0003872
1	2	NP_002617	liver phosphofruktokinase isoform b [Homo sapiens]	go:0003872
1	2	NP_002618	phosphofruktokinase, platelet [Homo sapiens]	go:0005724, go:0016301, go:0000166, go:0016740, go:0000287, go:0003872
1	2	NP_000280	phosphofruktokinase, muscle [Homo sapiens]	go:0005724, go:0016301, go:0000166, go:0016740, go:0000287, go:0003872
1	2	P17838	6-phosphofruktokinase, liver type (Phosphofruktokinase 1) (Phosphofruktokinase)	go:0003872



Evaluation

Number of entries: 92 Entropy: 1.0
 Number of groups: 26 Purity: 1.0
 Number of classes: 25 MutualInformation: 0.8810530832230519
 FMeasure: 0.9939613526570048

- Entropy: average distribution of the data entries in each group among the classes
- Purity: average precision of the groups with respect to their best matching classes
- Mutual information: correspondence on average between each group and class
- F-measure: precision and recall of the classes with respect to their best matching groups on average



Analysis

Glyc-Funct-AnnEc-onlyGO = SemSim(GOcomb)>0.95 - ConnectedComponents + Glycolysis: by Function

	0(5)	1(2)	2(14)	3(7)	4(2)	5(4)	6(4)	7(4)	8(4)	9(12)	10(5)	11(4)
0(2)					2/0/0							
1(14)			14/0/0									
2(12)										12/0/0		
3(7)					7/0/0							
4(8)												8/0
5(1)											1/0/4	
6(2)			2/0/0									
7(4)							4/0/0					
8(6)												
9(12)												
10(1)												
11(4)												4/0/1
12(5)		5/0/0										
13(1)												
14(1)												

Annotations: true positives (arrow to 11(4)/4(8)), false positives (arrow to 11(4)/1(14)), false negatives (arrow to 11(4)/11(4))

Analysis

group: 11(4) - class:10(5) + 4/0/1

GroupNr	ClassNr	ID	Definition	GO term(s)
11	10	P08559	Pyruvate dehydrogenase E1 component alpha subunit, somatic form, mitochondrial precursor (PDHE1-A ty)	go:0004739
11	10	NP_000275	pyruvate dehydrogenase (lipoamide) alpha 1 [Homo sapiens]	go:0016491, go:0004739, go:0016624
11	10	P11177	Pyruvate dehydrogenase E1 component beta subunit, mitochondrial precursor (PDHE1-B)	go:0004739
11	10	P29803	Pyruvate dehydrogenase E1 component alpha subunit, testis-specific form, mitochondrial precursor (PD)	go:0004739
5	10	P10515	Dihydropyridine-residue acetyltransferase component of pyruvate dehydrogenase complex, mitochond	go:0004742

Analysis - comparison

ID	DataSource	Rule	GrMethod	Classif	# of entries	# of groups	# of classes	Entropy	Purity	MutualInformation	FMeasure
1	Glyc-Funct-AnnEc-onlyGO	SemSim (GOcomb) > 0.95	ConnectedComponents	Glycolysis: by function	67	26	23	1.0	1.0	0.9117709736626631	0.974650558740109
2	Glyc-Funct-AnnEc-onlyGO	SemSim (GOcomb) > 0.95	ConnectedComponents	Glycolysis: by function	75	23	24	0.8652654637823463	0.8	0.7942395602417655	0.7917895141895143
3	Glyc-Funct-AnnEc-onlyGO	SemSim (GOcomb) > 0.95	ConnectedComponents	Glycolysis: by function	92	26	25	1.0	1.0	0.8810530832230523	0.9939613526570048
4	Glyc-Funct-AnnEc-onlyGO	SemSim (GOcomb) > 0.95	ConnectedComponents	Glycolysis: by function	92	21	25	0.777556968313313	0.6956521739130435	0.6824607300337851	0.7074322974655634
5	Glyc-Funct-AnnEc-onlyGO	SemSim (GOcomb) > 0.95	Cliques	Glycolysis: by function	92	29	25	1.0	1.0	0.8839495868213102	0.8392424467190822
6	Glyc-Funct-AnnEc-onlyGO	SeqSim (seq) > 0.85	ConnectedComponents	Glycolysis: by function	92	41	25	1.0	1.0	0.823166142253041	0.8046259464714613
7	Glyc-Funct-AnnEc-onlyGO	SemSim (GOcomb) > 0.95	ConnectedComponents	isozymes	92	26	47	0.7921820789964246	0.5869561217391305	0.7969682196235567	0.6484704432558992
8	Glyc-Funct-AnnEc-onlyGO	SeqSim (seq) > 0.85	ConnectedComponents	isozymes	92	41	47	0.9256079052200991	0.847826166956217	0.8848110696286732	0.8380839566609111

- Comparisons: use of different data sources, grouping algorithms, and classifications, grouping on different attributes, impact of threshold

Further reading

- Jakoniene V, Lambrix P. A tool for evaluating strategies for grouping of biological data, *Journal of Integrative Bioinformatics*, 4(3):83, 2007.
- Jakoniene V, Rundqvist D, Lambrix P. A method for similarity-based grouping of biological data. *DILS06*, LNBI 4075, 136-151, 2006.