A tool for evaluating strategies for grouping of biological data

KitEGA

Tool*Kit* for *E*valuation of *G*rouping *A*lgorithms

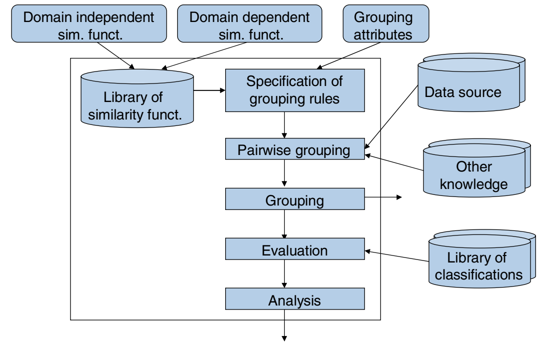
During the last decade an enormous amount of biological data has been generated and techniques and tools to analyze this data have been developed. Many of these tools use some form of grouping and are used in, for instance, data integration, data cleaning, prediction of protein functionality, and correlation of genes based on microarray data. Grouping of biological data is not a trivial task:

- a number of aspects influence the quality of the results: the data sources, the grouping attributes and the algorithms implementing the grouping procedure
- a variety of grouping algorithms is available, but it is often not clear which method performs best for which grouping tasks
- existing grouping algorithms may not be applied straightforward

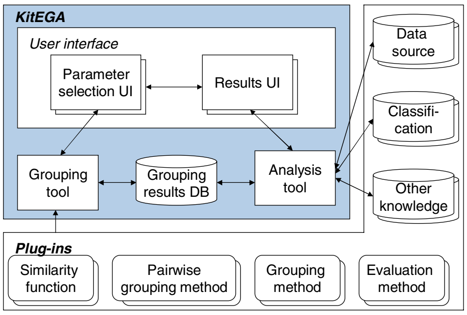
Environments that support comparison and evaluation of different grouping strategies for different grouping tasks on different data sets are needed to:

- support study of the properties of biological data sources
- select the suitable grouping procedures
- get an insight in how the grouping procedures could be used in the best way
- lead to recommendations on how to improve the current procedures and develop new procedures

Method for similarity-based grouping


Domain independent sim. funct.
Domain dependent sim. funct.
Grouping attributes
Library of similarity funct.
Specification of grouping rules
Data source
Pairwise grouping
Other knowledge
Grouping
Evaluation
Library of classifications
Analysis

The KitEGA framework



References

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', *Proceedings of the International Workshop on Data Integration in the Life Sciences - DILS06*, LNBI 4075, pp 136-151, 2006.

Contact: P. Lambrix
Department of Computer and Information Science
Linköping universitet, Sweden
kitega@ida.liu.se
http://www.ida.liu.se/~iislab/projects/KitEGA/

Example use

Grouping task. Grouping of proteins with respect to

1. biological function
2. class of isozymes they belong to

Formation of data sources:

- analyzed human proteins involved in glycolysis
- retrieved 190 data entries via Entrez
- used GO Consortium mappings ec2go and spkw2go to extend the set of available GO terms

The first prototype includes:

Library of similarity functions

- EditDist(v1,v2)
- SeqSim(v1,v2)
- SemSim(v1,v2)

Other knowledge

- GO ontology

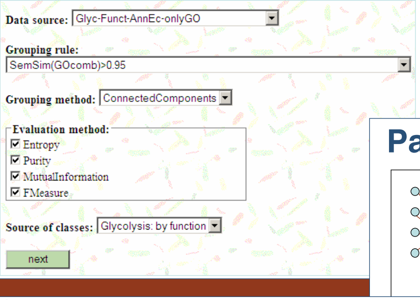
Classifications. Manual classification according to

- biological function
- classes of isozymes

Grouping metohds

- connected components (transitive similarity)
- cliques (overlapping groups)

Specification of grouping rules



Data source: Glyc-Funct-AnnEc-onlyGO

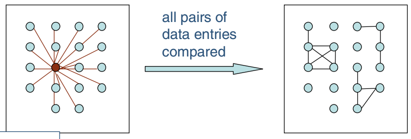
Grouping rule: SemSim(GOcomb)>0.95

Grouping method: ConnectedComponents

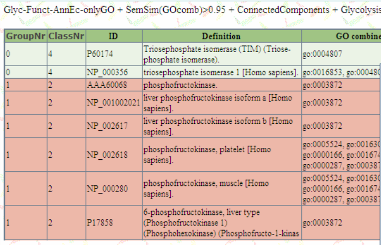
Evaluation method:
☑ Entropy
☑ Purity
☑ MutualInformation
☑ FMeasure

Source of classes: Glycolysis: by function

Pairwise grouping



Grouping

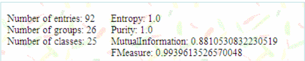


Glyc-Funct-AnnEc-onlyGO + SemSim(GOcomb)>0.95 + ConnectedComponents + Glycolysis: by function

GroupNr	ClassNr	ID	Definition	GO combined
0	4	P60174	Triosephosphate isomerase (TIM) (Triose-phosphate isomerase).	go:0004807
0	4	NP_000356	triosephosphate isomerase 1 [Homo sapiens].	go:0016853, go:000480
1	2	AAA60068	phosphofructokinase.	go:0003872
1	2	NP_001002021	liver phosphofructokinase isoform a [Homo sapiens].	go:0003872
1	2	NP_002617	liver phosphofructokinase isoform b [Homo sapiens].	go:0003872
1	2	NP_002618	phosphofructokinase, platelet [Homo sapiens].	go:0005524, go:001630 go:0000166, go:001674 go:0000287, go:000387
1	2	NP_000280	phosphofructokinase, muscle [Homo sapiens].	go:0005524, go:001630 go:0000166, go:001674 go:0000287, go:000387
1	2	P17858	6-phosphofructokinase, liver type (Phosphofructokinase 1) (Phosphohexokinase) (Phosphofructo-1-kinas	go:0003872

Evaluation

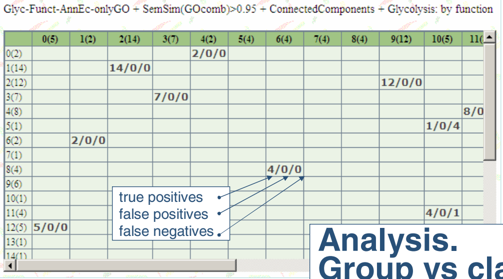
External quality measures, i.e. with respect to known classes of the grouped data



Number of entries: 92
Number of groups: 26
Number of classes: 25

Entropy: 1.0
Purity: 1.0
MutualInformation: 0.8810530832230519
FMeasure: 0.9939613526570048

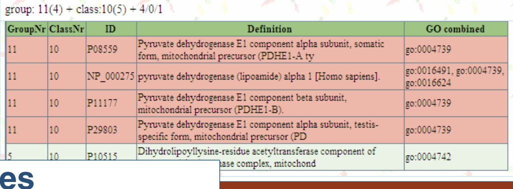
Analysis. Distribution of data entries in a test case



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

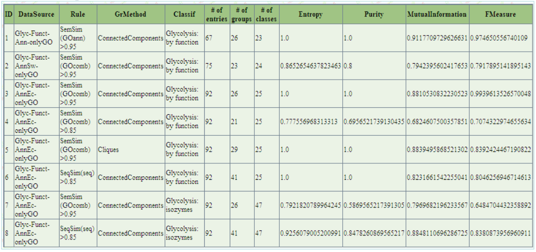
Analysis. Group vs class data entries



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

GroupNr	ClassNr	ID	Definition	GO combined
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	Dihydrolipoyllysine-residue acetyltransferase component of ase complex, mitochond	go:0004742

Analysis. All test cases



ID	DataSource	Rule	GrMethod	Classif	# of entries	# of groups	# of classes	Entropy	Purity	MutualInformation	FMeasure
1	Glyc-Funct-Ann-onlyGO	SemSim (GOann) >0.95	ConnectedComponents	Glycolysis: by function	67	26	23	1.0	1.0	0.911770972962663l	0.974650556740109
2	Glyc-Funct-AnnSw-onlyGO	SemSim (GOcomb) >0.95	ConnectedComponents	Glycolysis: by function	75	23	24	0.8652654637823463	0.8	0.7942395602417653	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) >0.85	ConnectedComponents	Glycolysis: by function	92	21	25	0.77556968313313	0.6956521739130435	0.6824607500357851	0.7074322974655634
5	Glyc-Funct-AnnEc-onlyGO	SemSim (GOcomb) >0.95	Cliques	Glycolysis: by function	92	29	25	1.0	1.0	0.8839495868521302	0.8392424467190822
6	Glyc-Funct-AnnEc-onlyGO	SeqSim(seq) >0.85	ConnectedComponents	Glycolysis: by function	92	41	25	1.0	1.0	0.8231661542255041	0.8046256946714613
7	Glyc-Funct-AnnEc-onlyGO	SemSim (GOcomb) >0.95	ConnectedComponents	Glycolysis: isozymes	92	26	47	0.7921820789964245	0.5869565217391305	0.7969682196233567	0.6484704432358892
8	Glyc-Funct-AnnEc-onlyGO	SeqSim(seq) >0.85	ConnectedComponents	Glycolysis: isozymes	92	41	47	0.9256079005200991	0.8478260869565217	0.8848110696286725	0.8380873956960911

Studied aspects, e.g. use of different data sources, grouping algorithms, and classifications, grouping on different attributes, impact of threshold.

Conclusions for test cases obtained by using KitEGA:

- best suited grouping approaches for data source Glyc-Funct-AnnEc-onlyGO
  SemSim(GOcomb) for grouping on biological function
  SeqSim(Sequence) for grouping on classes of isozymes
- for the used grouping tasks spkw2go – too general, ec2go – specific enough