A tool for evaluating strategies for grouping of biological data

Vaida Jakonienė, Patrick Lambrix

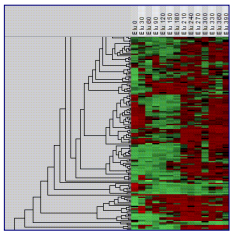
Linköpings universitet

Outline

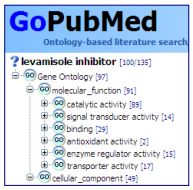
- Motivation
- Method for similarity based grouping
- KitEGA – illustration
- Summary and future work

Tools for biological data analysis

Hierarchical microarray clustering (J-Express Pro)

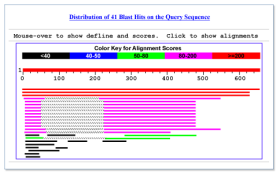


Classification of abstracts



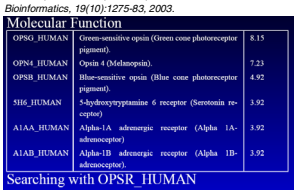
Similarity of biological data

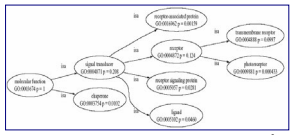
Sequence alignment (BLAST)



Similarity between data entries

Lord PW, Stevens RD, Brass A, Goble CA. Bioinformatics, 19(10):1275-83, 2003.





- Basic task – computation of a similarity value between objects

Similarity-based grouping

- Not a trivial task
  - data is complex
  - many grouping algorithms available: which algorithm performs best for which grouping task?
  - grouping on which attributes?
  - existing grouping algorithms may not be applied straightforward to new data sets

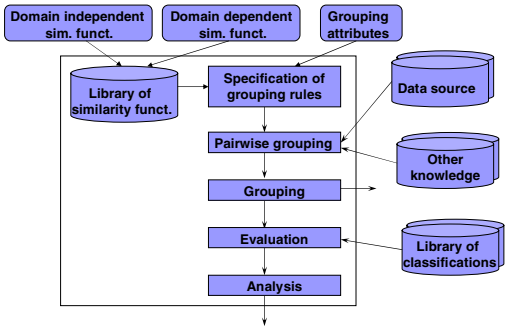
Similarity-based grouping

- Environments that support study, comparison and evaluation of different grouping strategies are needed

Outline

- Motivation
- Method for similarity based grouping
- KitEGA – illustration
- Summary and future work

Method for similarity-based grouping



Outline

- Motivation
- Method for similarity based grouping
- KitEGA – illustration
- Summary and future work

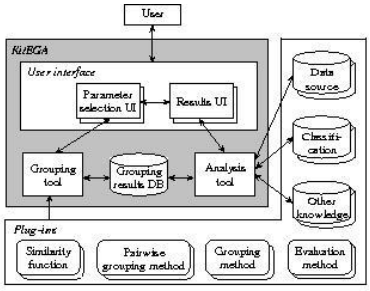
KitEGA

- A toolKit for Evaluating Grouping Algorithms

Idea

- Input components (plug-ins)
  - grouping procedures to be evaluated
  - data sources
  - evaluation methods
  - classifications
  - other knowledge
- Tool executes algorithms and stores results
- User analyzes results using different views on the result data

KitEGA framework



Illustration

- Grouping task. Grouping of proteins with respect to
  - biological function
  - class of isozymes they belong to
- Data source(s)
  - human proteins involved in glycolysis
  - via Entrez retrieved 190 data entries

Data entry

```
LOCUS       NP_000275               390 aa            linear   PRI 16-APR-2006
DEFINITION  pyruvate dehydrogenase (lipoamide) alpha 1 [Homo sapiens].
ACCESSION   NP_000275
VERSION     NP_000275.1  GI:4505685
DBSOURCE    REFSEQ: accession NM_000284.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 390)
  AUTHORS   Hiromasa,Y., Fujisawa,T., Aso,Y. and Roche,T.E.
  TITLE     Organization of the cores of the mammalian pyruvate dehydrogenase
            complex formed by E2 and E2 plus the E3-binding protein and their
            capacities to bind the E1 and E3 components
  JOURNAL   J. Biol. Chem. 279 (8), 6921-6933 (2004)
   PUBMED   14638692
  REMARK    GeneRIF: model of the pyruvate dehydrogenase complex formed by E2
            and E2 plus the E3-binding protein and binding of the E1 and E3
            components
```

Entrez. Protein database

Data entry

```
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from L13318.1.

            Summary: The pyruvate dehydrogenase complex is a nuclear-encoded
            mitochondrial matrix multienzyme complex that provides the primary
            link between glycolysis and the tricarboxylic acid (TCA) cycle by
            catalyzing the irreversible conversion of pyruvate into acetyl-CoA.
            The PDH complex is composed of multiple copies of 3 enzymes: E1
            (PDHA1); dihydrolipoyl transacetylase (DLAT; MIM 608770) (E2; EC
            2.3.1.12); and dihydrolipoyl dehydrogenase (DLD; MIM 238331) (E3;
            EC 1.8.1.4). The E1 enzyme is a heterotetramer of 2 alpha and 2
            beta subunits. The E1-alpha subunit contains the E1 active site and
            plays a key role in the function of the PDH complex (Brown et al.,
            1994).[supplied by OMIM].
FEATURES             Location/Qualifiers
     source          1..390
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="X"
                     /map="Xp22.2-p22.1"
     Protein         1..390
                     /product="pyruvate dehydrogenase (lipoamide) alpha 1"
                     /EC_number="1.2.4.1"
                     /note="pyruvate dehydrogenase alpha subunit; pyruvate
                     dehydrogenase E1 alpha subunit"
```


Data entry

```
     CDS             1..390
                     /gene="PDHA1"
                     /coded_by="NM_000284.1:106..1278"
GO_ann               /go_component="mitochondrion [pmid 3034892]"
                     /go_function="oxidoreductase activity; oxidoreductase
                     activity, acting on the aldehyde or oxo group of donors,
                     disulfide as acceptor; pyruvate dehydrogenase
                     (acetyl-transferring) activity"
                     /go_process="acetyl-CoA metabolism; glycolysis;
                     metabolism"
                     /db_xref="GeneID:5160"
                     /db_xref="HGNC:8806"
                     /db_xref="HPRD:02420"
                     /db_xref="MIM:300502"
Sequence
ORIGIN
        1 mrkmlaavsr vlagasqkpa srvlvasrnf andatfeikk cdlhrleegp pvttvltred
       61 glkyyrmmqt vrrmelkadq lykqkiirgf chlcdqqeac cvgleaginp tdhlitayra
      121 hgftftrgls vreilaeltg rkggcakgkg gsmhmyaknf yggngivgaq vplgagiala
      181 ckyngkdevc ltlygdgaan qgqifeaynm aalwklpcif icennrygmg tsveraaast
      241 dyykrgdfip glrvdgmdil cvreatrfaa aycrsgkgpi lmelqtyryh ghsmsdpgvs
      301 yrtreeiqev rsksdpimll kdrmvnsnla sveelkeidv evrkeiedaa qfatadpepp
      361 leelgyhiys sdppfevrga nqwikfksvs
//
```


Data sources

DS1: GO_{ann}, 67 data entries

- only terms of GO function ontology analyzed
- only data entries having GO terms

GO Consortium. Mappings between data values and ontological terms:
ec2go – ec_numbers translated into GO terms
spkw2go – swissprot keywords translated into GO terms

DS2:
Keywords —*spkw2go*→
Ec_number —*ec2go*→ GO_{comb}, 93 data entries
GO_{ann} →

DS3:
Ec_number —*ec2go*→ GO_{comb}, 92 data entries
GO_{ann} →

Grouping components

- Library of similarity functions
  - EditDist(v_1,v_2)
  - SeqSim(v_1,v_2)
  - SemSim(v_1,v_2) (→ GO ontology)
- Grouping rules
  - GO ontology
- Grouping methods
  - Connected components
  - Cliques

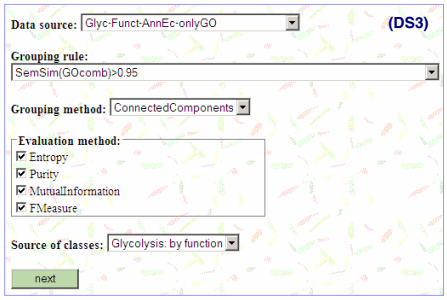
Evaluation methods

- Types of quality measures
  - internal – based on information obtained during the grouping
  - external – with respect to known classes of the grouped data
- In this illustration: external
  - Purity
  - F-measure
  - Entropy
  - Mutual information

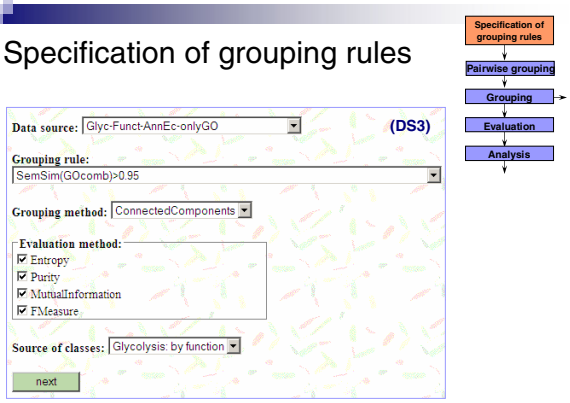
Classifications

- Manual classification according to
  - biological function
  - classes of isozymes

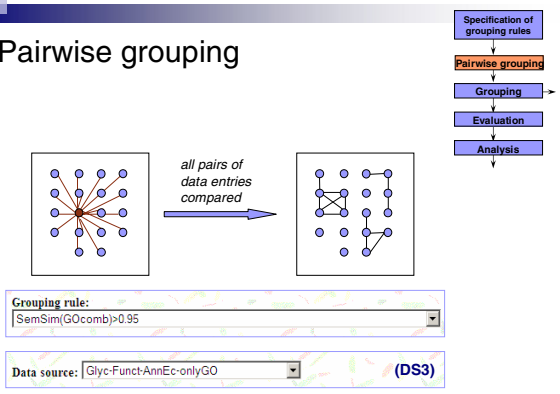
Selection of test case



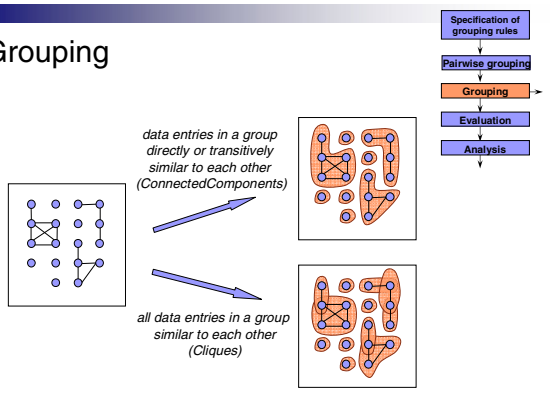
Specification of grouping rules



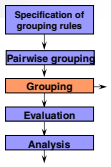
Pairwise grouping



Grouping



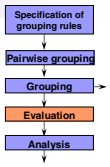
Grouping



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:0004807
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:0016301, go:0000166, go:0016740, go:0000287, go:0003872
1	2	NP_000280	phosphofructokinase, muscle [Homo sapiens].	go:0005524, go:0016301, go:0000166, go:0016740, go:0000287, go:0003872
1	2	P17858	6-phosphofructokinase, liver type (Phosphofructokinase 1) (Phosphohexokinase) (Phosphofructo-1-kinas	go:0003872

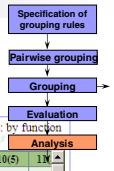
Evaluation



Number of entries: 92
Number of groups: 26
Number of classes: 25
Entropy: 1.0
Purity: 1.0
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

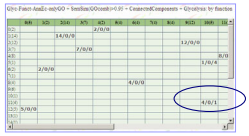


Glyc-Funct-AnnEc-onlyGO + SemSim(GOcomb)>0.95 + ConnectedComponents + Glycolysis: by function

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

true positives
false positives
false negatives

Analysis



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 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-Ann-onlyGO	SemSim (OAnn) >0.95	ConnectedComponents	Glycolysis: by function	67	26	23	1.0	1.0	0.9117709729626631	0.9746055674010
2	Glyc-Funct-AnnSw-onlyGO	SemSim (GOcomb) >0.95	ConnectedComponents	Glycolysis: by function	75	23	24	0.8652654637823463	0.8	0.7942956024176 53	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	0.7775569683133 13	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	0.8380739596090911

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

Test cases. Observations

- Best suited grouping approaches. For data source Glyc-Funct-AnnEc-onlyGO (DS3)
  - SemSim(GOcomb) for grouping on biological function
  - SeqSim(Sequence) for grouping on classes of isozymes
- Suitability of mappings for the used grouping approaches
  - spkw2go – too general, e.g. 'Glycolysis'
  - ec2go – specific enough, e.g. '6-phosphofructokinase activity'

Summary and future work

- Motivated need for environments that support the evaluation and comparison of similarity-based grouping procedures
- Implemented the KitEGA tool based on a method for evaluating similarity-based grouping algorithms
- Illustrated KitEGA using test cases based on different strategies and classifications

- Extend the Kitega implementation