BiGGEsTS: integrated environment for biclustering analysis of time series gene expression data

Joana P. Gonçalves [1,2,3], Sara C. Madeira[1,2,3] and Arlindo L. Oliveira[1,2]

[1] KDBIO group, INESC-ID, [2] IST-UTL, [3] UBI

PORTUGAL

1 ABSTRACT

BiGGEsTS is a software tool for analyzing time series gene expression data and revealing local co-regulations of genes in specific intervals of time using biclustering algorithms. GO annotations can be integrated and used to assess which of the identified patterns are biologically relevant. The analysis is additionally supported by a number of preprocessing and post-processing methods, together with a visualization module capable of displaying informative representations of the data.

2 ANALYSIS OF TIME SERIES GENE EXPRESSION DATA

INPUT

- Universal format: delimited text file
- Log transformation
- Support for a wide range of organisms

	C1	C2	C3	C4	C5
G1	0.07	0.73	-0.54	0.45	0.25
G2	-0.34	0.46	-0.38	0.76	-0.44
G3	0.22	0.17	-0.11	0.44	-0.11
G4	0.70	0.71	-0.41	0.33	0.35

PREPROCESS

- **Filter genes**
  eliminate missing values
- **Treat missing values**
  fill missings with a given value
- **Normalize**
  compensate experimental differences between samples and different expression rates bettwen genes
- **Smooth**
  eliminate outliers
- **Discretize**
  reduce range of values

	C1	C2	C3	C4	C5
G1	N	U	D	U	N
G2	D	U	D	U	D
G3	N	N	N	U	N
G4	U	U	D	U	U

IDENTIFY BICLUSTERS

CCC-Biclustering & e-CCC-Biclustering

- combinatorial approaches based on suffix trees
- identify all occurrences of potential patterns (maximal contiguous column biclusters)
- scoring schema based on statistical significance of patterns and similarity with overlapping biclusters

CC-TSB

- greedy approach: iteratively add/remove genes/conditions
- goal: minimize mean squared residue score
- predefined # of biclusters

	C1	C2	C3	C4	C5
G1	N	U	D	U	N
G2	D	U	D	U	D
G3	N	N	N	U	N
G4	U	U	D	U	U

	CCC	e-CCC	CC-TSB
Type of data			
Real-valued data	-	-	●
Discrete data	●	●	-
Biclusters identified			
Exact patterns	●	-	-
Approximate patterns	-	●	●
Parameters			
Requires user-defined	-	●	●
Variations			
Anti-correlated patterns	●	●	-
Shifted patterns	●	●	-
Time-lagged patterns	●	●	-
Behavior with missings			
Jump	●	●	-
Consider as errors	-	●	-

POST-PROCESS

- **Filter biclusters**
- **Sort biclusters**

Bicluster1
Bicluster 2
Bicluster 3
Bicluster 4

Bicluster 3
Bicluster 1
Bicluster 4

- **Criteria:**
- triviality
- size, # genes, #conditions
- mean squared residue
- avg row/column variance
- similarity
- significance of GO terms
- significance of bicluster pattern

VALIDATE

- Statistical significance of GO terms
- Statistical significance of bicluster patterns

ANALYZE & VISUALIZE

- Heatmaps
- Dendrograms
- GO terms
- Expression profile/pattern charts
- Term-for-term analysis (p-values + correction for multiple tests)
- Graphs of significant terms

http://kdbio.inesc-id.pt/software/biggests

3 References

[1] Madeira SC and Oliveira AL, IEEE/ACM TCBB 2004, 1:24-25
[2] Madeira SC, Teixeira MC, Sá-Correia I, Oliveira AL, IEEE/ACM TCBB 2008, http://doi.ieeecomputersociety.org/10.1109/TCBB.2008.34
[3] Madeira SC and Oliveira AL, Proc of the 5th APBC, Imperial College Press 2007, 67-80
[4] Zhang Y, Zha H, and Chu CH, Proc of the 5th IEEE International Conference on Information Technology: Coding and Computing 2005, 32-37

Acknowledgements