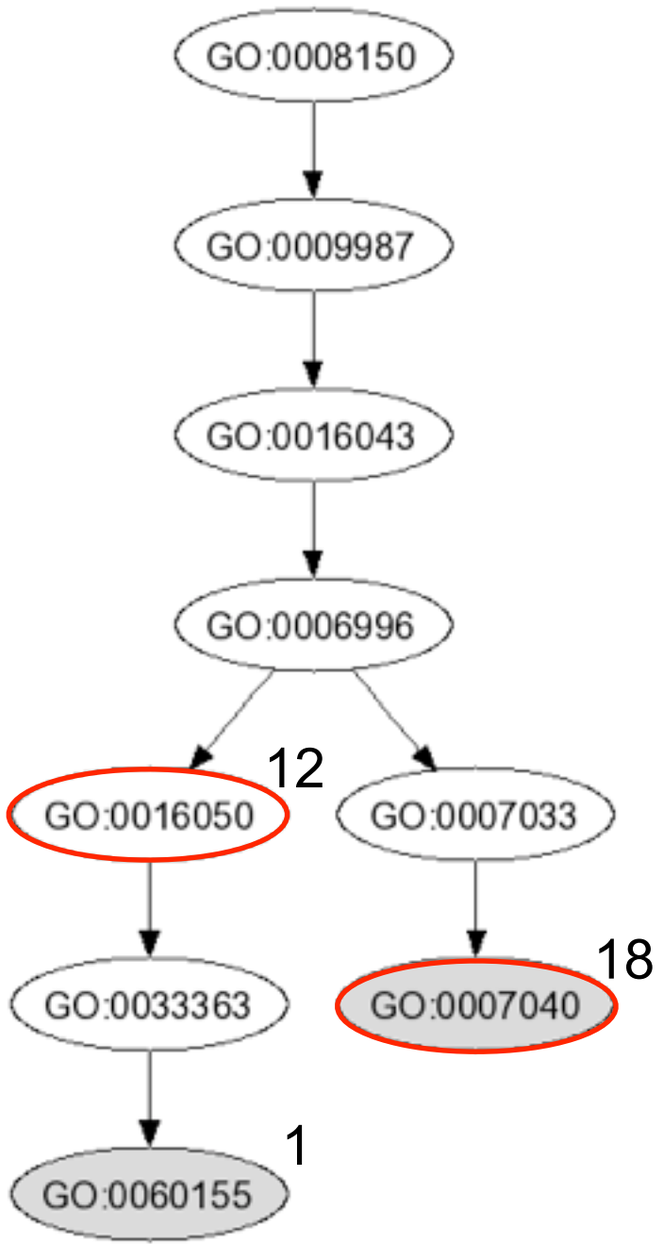
How to evaluate a quality of gene coexpression data in ATTED-II and COXPRESdb

Takeshi Obayashi

2010. 12. 23

GO term extraction for assessment

Extraction rule:
(1) All annotations are mapped to upper terms.
(2) Terms with low direct annotation are omitted (<10% of all mapped annotation).
(3) Terms with [5,20] genes are extracted.



Vesicle

Secretory granule

Platelet dense granule

Vacuole

Lysosome

Example:
A part of GO CC annotation
for Human ABCA1 Gene (EGI:19)

Typically, one or two GO terms are assigned
for each gene and for each of BP, CC, MF.

Definition of GOOD coexpression data

For 2 genes (X, Y) and 10 GO annotations (a .. j),
we can suppose 20 possible gene-GO associations.

Gene X - GO a real	Gene Y - GO a
Gene X - GO b real	Gene Y - GO b
Gene X - GO c	Gene Y - GO c real
Gene X - GO d	Gene Y - GO d real
Gene X - GO e	Gene Y - GO e
Gene X - GO f	Gene Y - GO f
Gene X - GO g	Gene Y - GO g
Gene X - GO h	Gene Y - GO h
Gene X - GO i	Gene Y - GO i
Gene X - GO j	Gene Y - GO j

GOOD coexpression data will correctly predict
real associations in the possible associations above.

Coexpressed genes for Gene X, Y

Prediction target (guide gene)

GeneX

Gene annotation

GO a GO b

GO c

GO a GO d

GO a GO b

GO e

Coex. gene list

Top1
Top2
Top3
Top4
Top5
Top6

Prediction target (guide gene)

GeneY

Gene annotation

GO a GO c

GO b

GO b

GO e

GO d GO f

GO e

Coex. gene list

Top1
Top2
Top3
Top4
Top5
Top6

Introduce a common threshold for valid coex.

Prediction target
(guide gene)

GeneX

Prediction target
(guide gene)

GeneY

Gene annotation

GO a

GO b

GO c

~~GO a~~

GO d

~~GO a~~

~~GO b~~

GO e

Coex. gene list

Top1

Top2

Top3

Top4

Top5

Top6

Gene annotation

GO a

GO c

GO b

~~GO b~~

GO e

GO d

GO f

~~GO e~~

Coex. gene list

Top1

Top2

Top3

Top4

Top5

Top6

Associated

Not Associated

Threshold of valid coexpression

Assign True and False to each predicted association

Prediction target (guide gene)

GO a *GO b*

GeneX

Gene annotation

True GO a

True GO b

False GO c

~~GO a~~

False GO d

~~GO a~~

~~GO b~~

True GO e

Coex. gene list

Top1

Top2

Top3

Top4

Top5

Top6

Threshold of valid coexpression

Prediction target (guide gene)

GO c *GO d*

GeneY

Gene annotation

False GO a

True GO c

False GO b

~~GO b~~

False GO e

False GO d

True GO f

~~GO e~~

Coex. gene list

Top1

Top2

Top3

Top4

Top5

Top6

Associated

Not Associated

Make contingency table for the prediction

	Associated (real annotation)	Not associated (real annotation)
Assocated (pred)	GeneX - GO a GeneX - GO b GeneY - GO d 3	GeneX - GO c GeneX - GO d GeneY - GO a GeneY - GO b GeneY - GO e 5
Not associated (pred)	GeneY - GO d 1	Gene X - GO e Gene X - GO f Gene X - GO g Gene X - GO h Gene X - GO i Gene X - GO j Gene Y - GO f Gene Y - GO g Gene Y - GO h Gene Y - GO i Gene Y - GO j 11

True positive rate = 0.75 (3 / 4)
False positive rate = 0.32 (5 / 16)

, under the given coex threshold (Top4 this time).

Move the threshold to draw ROC curve

Prediction target (guide gene)

GeneX

Prediction target (guide gene)

GeneY

Coex. gene list

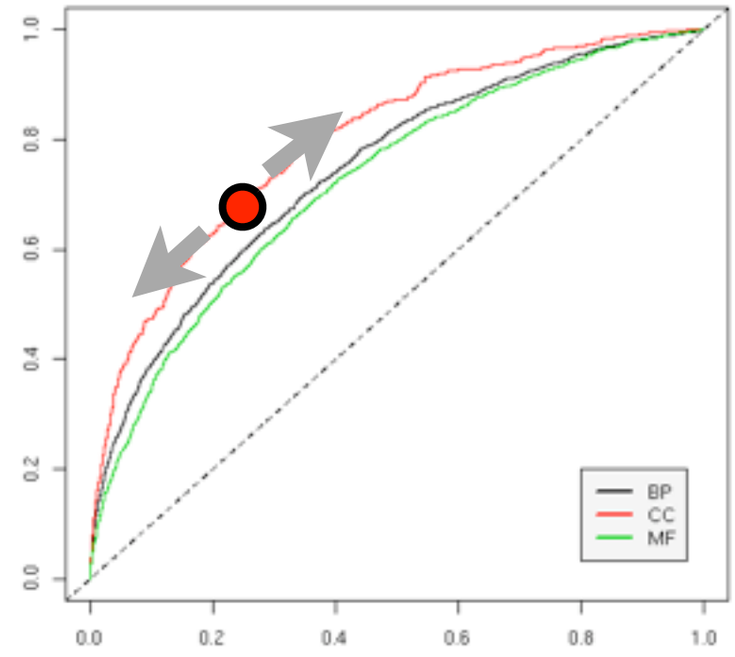
Top1
Top2
Top3
Top4
Top5
Top6

Coex. gene list

Top1
Top2
Top3
Top4
Top5
Top6

Threshold of valid coexpression

ROC curve



AUC