GENECODIS

A web-based tool for finding significant concurrent annotations in gene lists

Pedro Carmona-Saez et. al

GENECODIS: A web-based tool for finding significant concurrent annotations in gene lists

Genome Biology 2007, **8**:R3 doi:10.1186/gb-2007-8-1-r3

Pedro Carmona-Saez (pcarmona@cnb.uam.es)
Monica Chagoyen (monica@cnb.uam.es)
Francisco Tirado (ptirado@dacya.ucm.es)
Jose M Carazo (carazo@cnb.uam.es)
Alberto Pascual-Montano (pascual@fis.ucm.es)

Background

- Expression study or proteomics → list of potentially interesting genes or proteins.
  - e.g. genes transcribed only in pathological tissue.
- What is the molecular biology behind this?
- Next step: find out which functions are these genes associated with?
  - interpret and extract the knowledge from a large list of genes or proteins.
- Most applications find annotations that are significantly enriched in a list of genes compared to a reference set (genome, or genes used in microarray)
  - onto-Express was one of the first.
- GENECODIS takes it one step further.

GENECODIS

- A web based tool to find function of genes or proteins used in expression studies or proteomics.
- Find **combinations** annotations that are overrepresented in a list of genes compared to a reference set (genome, or genes used in microarray)
- Current tools
  - evaluate single annotations
  - don’t take into account their potential relationships
- Sources: KEGG, Swiss-Prot, GO, InterPro
- Result: rank scores for single annotations and their combinations.
- Potentially important extension to existing tools.

Example of advantage

- Co-occurence patterns add information.
- Single annotations have limitations
  - experiment result (other tool): “signal transduction”
    - concrete aspect of cell physiology
    - but used in many different biological processes
    - what is signalled?
    - co-occurs with “cell-proliferation”
    - this alone is insignificant – many such annotations in DB compared to our list.
    - GENECODIS result: genes related to signalling cell-proliferation
    - Relevant associations might be underestimated if single annotation are taken into account.

Algorithm I

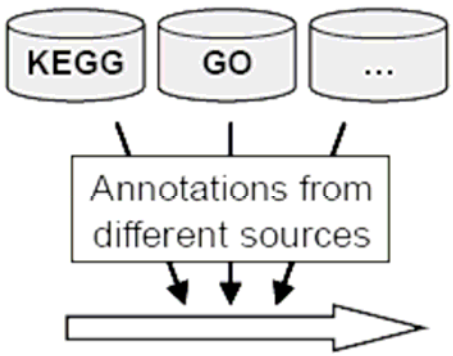
- User inputs a list of differentially expressed genes.
- Retrieve all DB (e.g. GO) annotations for every gene.
- Make a list of all frequently found annotations and their combinations (sets)
  - include only annotation combinations appearing in at least X genes
- Remove redundant sets (duplicate information)

Retrieve annotations for genes

- Retrieve annotations for each gene from the selected databases

List of genes

ACO1
CIT1
CIT2
CIT3
FUM1
IDH1
IDH2
KGD1
KGD2
LSC1
LSC2
YJL200C



genes	Annotations
ACO1	GO:0005759,GO:0005829,GO:0042645,sce00020,sce00630,sce00720
CIT1	GO:0005739,GO:0005759,sce00020,sce00630
CIT2	GO:0005739, sce00020,sce00630
CIT3	GO:0005759,sce00020,sce00630
FUM1	GO:0005759,GO:0005829,sce00020,sce00720
IDH1	GO:0005759,GO:0042645,sce00020
IDH2	GO:0005739,GO:0005759,sce00020
KGD1	GO:0005759,GO:0009353,GO:0042645,sce00020,sce00310,sce00380
KGD2	GO:0005759,GO:0009353,GO:0042645,sce00020,sce00310
LSC1	GO:0005739,GO:0042645,sce00020,sce00640
LSC2	GO:0005739,sce00020,sce00640
YJL200C	GO:0005739,sce00020,sce00630,sce00720

- Make a list of all frequently appearing annotations and their co-occurences containing at least **X** genes

Annotations	# genes	genes
GO:0005759	8	ACO1,CIT1,CIT3,FUM1,IDH1,IDH2,KGD1,KGD2
GO:0042645	5	ACO1,IDH1,KGD1,KGD2,LSC1
sce00020	12	ACO1,CIT1,CIT2,CIT3,FUM1,IDH1,IDH2,KGD1,KGD2,LSC1,LSC2,YJL200C
sce00630	5	ACO1,CIT1,CIT2,CIT3,YJL200C
sce00720	3	ACO1,FUM1,YJL200C
GO:0005739	6	CIT1,CIT2,IDH2,LSC1,LSC2,YJL200C

Finding frequently appearing sets

- To extract combinations of gene annotations GENECODIS uses a modification of the methodology reported in 2006 by Carmona-Saez et al., which implements **the *apriori* algorithm** to extract associations among gene annotations and expression patterns.

Carmona-Saez P, Chagoyen M, Rodriguez A, Trelles O, Carazo JM, Pascual-Montano A: **Integrated analysis of gene expression by Association Rules Discovery.** *BMC Bioinformatics* 2006, **7**:54.

- The *apriori* algorithm was originally introduced in 1993 by Agrawal *et al*.and has been extensively used to extract association rules from transaction databases.

Agrawal R, Imielinski T, Swami A: **Mining Association Rules between Sets of Items in Large Databases.** In *Proceedings of the ACM SIGMOD international conference on Management of data*; *Washington, D.C.*; 1993: 207-216.

1. Find frequent 1-item sets

FINDING FREQUENTLY APPEARING SETS

Annotations	# genes	genes
GO:0005759	8	ACO1,CIT1,CIT3,FUM1,IDH1,IDH2,KGD1,KGD2
GO:0042645	5	ACO1,IDH1,KGD1,KGD2,LSC1
sce00020	12	ACO1,CIT1,CIT2,CIT3,FUM1,IDH1,IDH2,KGD1,KGD2,LSC1,LSC2,YJL200C
sce00630	5	ACO1,CIT1,CIT2,CIT3,YJL200C
sce00720	3	ACO1,FUM1,YJL200C
GO:0005739	6	CIT1,CIT2,IDH2,LSC1,LSC2,YJL200C

2. Find frequent 2-item sets

Annotations	# genes	genes
GO:0005759,GO:0042645	4	ACO1,IDH1,KGD1,KGD2
GO:0005759,sce00020	8	ACO1,CIT1,CIT3,FUM1,IDH1,IDH2,KGD1,KGD2
GO:0005759,sce00630	3	ACO1,CIT1,CIT3
GO:0042645,sce00020	5	ACO1,IDH1,KGD1,KGD2,LSC1
sce00020,sce00630	5	ACO1,CIT1,CIT2,CIT3,YJL200C
sce00020,sce00720	3	ACO1,FUM1,YJL200C
sce00020,GO:0005739	6	CIT1,CIT2,IDH2,LSC1,LSC2,YJL200C
sce00630,GO:0005739	3	CIT1,CIT2,YJL200C

3. Find frequent 3-item sets

…repeat until no more itemsets with 3 genes

Annotations	# genes	genes
GO:0005759,GO:0042645,sce00020	4	ACO1,IDH1,KGD1,KGD2
GO:0005759,sce00020,sce00630	3	ACO1,CIT1,CIT3
sce00020,sce00630,GO:0005739	3	CIT1,CIT2,YJL200C

Remove redundant sets

Annotations	# genes	genes
GO:0005759,GO:0042645	4	ACO1,IDH1,KGD1,KGD2
GO:0005759,sce00020	8	ACO1,CIT1,CIT3,FUM1,IDH1,IDH2,KGD1,KGD2
GO:0005759,sce00630	3	ACO1,CIT1,CIT3
GO:0042645,sce00020	5	ACO1,IDH1,KGD1,KGD2,LSC1
sce00020,sce00630	5	ACO1,CIT1,CIT2,CIT3,YJL200C
sce00020,sce00720	3	ACO1,FUM1,YJL200C
sce00020,GO:0005739	6	CIT1,CIT2,IDH2,LSC1,LSC2,YJL200C
sce00630,GO:0005739	3	CIT1,CIT2,YJL200C

Annotations	# genes	genes
GO:0005759,GO:0042645,sce00020	4	ACO1,IDH1,KGD1,KGD2
GO:0005759,sce00020,sce00630	3	ACO1,CIT1,CIT3
sce00020,sce00630,GO:0005739	3	CIT1,CIT2,YJL200C

Redundant itemset - subset of a larger itemset that has >= support value (genes). No loss of information.

Algorithm II

- Test statistically which sets are overrepresented in genelist compared to the reference list
  - find frequency of occurences of each set in the genelist and reference list.
    - default reference list: NCBI Entrez Gene DB of corresponding genome.
  - statistical Tests in GENECODIS:
    - hypergeometric distribution
    - chi-square test of independence
  - get p-values
    - low p-value means an annotation shouldn’t appear in your genelist solely by chance.
    - many genelist genes have an annotation – low p-value
    - many reference list genes have an annotation – high p-value
  - correct p-values for multiple tests
    - simulation based method
    - false discovery method

Statistical tests

- Compute frequency of each set
- Calculate p-values
- Correct p-values
  - simulations based method
  - false discovery method

Annotation/s	# List	# Reference	*p*-value	Corrected *p*-value	Genes	Description/s
00020	12(12)	30(6194)	1.90e-28	1.90e-27	CIT2, ACO1, KGD2, LSC2, YJL200C, IDH2, LSC1, KGD1, IDH1, CIT1, FUM1, CIT3	(KEGG)Citrate cycle (TCA cycle)
00020, GO:0005759	8(12)	9(6194)	1.52e-21	1.52e-20	ACO1, KGD2, IDH2, KGD1, IDH1, CIT1, FUM1, CIT3	(KEGG)Citrate cycle (TCA cycle) | (CC)mitochondrial matrix
00020, GO:0005739	6(12)	9(6194)	5.43e-15	5.43e-14	CIT2, LSC2, YJL200C, IDH2, LSC1, CIT1	(KEGG)Citrate cycle (TCA cycle) | (CC)mitochondrion
00020, GO:0042645	5(12)	7(6194)	5.83e-13	5.83e-12	ACO1, KGD2, LSC1, KGD1, IDH1	(KEGG)Citrate cycle (TCA cycle) | (CC)mitochondrial nucleoid
00020, 00630	5(12)	8(6194)	2.62e-12	2.62e-11	CIT2, ACO1, YJL200C, CIT1, CIT3	(KEGG)Citrate cycle (TCA cycle) | (KEGG)Glyoxylate and dicarboxylate metabolism

Algorithm III

- User chooses a p-value threshold
- Consider annotation sets below threshold as biologically significant to your experiment.

Annotation/s	# List	# Reference	*p*-value	Corrected *p*-value	Genes	Description/s
00020	12(12)	30(6194)	1.90e-28	1.90e-27	CIT2, ACO1, KGD2, LSC2, YJL200C, IDH2, LSC1, KGD1, IDH1, CIT1, FUM1, CIT3	(KEGG)Citrate cycle (TCA cycle)
00020, GO:0005759	8(12)	9(6194)	1.52e-21	1.52e-20	ACO1, KGD2, IDH2, KGD1, IDH1, CIT1, FUM1, CIT3	(KEGG)Citrate cycle (TCA cycle) | (CC)mitochondrial matrix
00020, GO:0005739	6(12)	9(6194)	5.43e-15	5.43e-14	CIT2, LSC2, YJL200C, IDH2, LSC1, CIT1	(KEGG)Citrate cycle (TCA cycle) | (CC)mitochondrion
00020, GO:0042645	5(12)	7(6194)	5.83e-13	5.83e-12	ACO1, KGD2, LSC1, KGD1, IDH1	(KEGG)Citrate cycle (TCA cycle) | (CC)mitochondrial nucleoid
00020, 00630	5(12)	8(6194)	2.62e-12	2.62e-11	CIT2, ACO1, YJL200C, CIT1, CIT3	(KEGG)Citrate cycle (TCA cycle) | (KEGG)Glyoxylate and dicarboxylate metabolism

Algorithm characteristics

- Computation time is increased by:
  - searching larger / more databases
  - decreasing minimum support value → more algorithm cycles.
  - e.g. Searching for all possible combinations that appear in at least 1 gene is often computationally unfeasible.
- Suggested minimum support value: 3

GENECODIS at work

- Human data – 85 expressed genes
  - GTOM program by Zhang
    - data from GO Biological Processses
      - resulting annotations: 1) cell proliferation; 2) testis-specific developement; 3) protein phosphorylation; 4) glycerolipid metabolism.
  - GENECODIS
    - data: GO Biological Processes, InterPro motifs
    - X = 3
    - result differences:
      - no glycerolipid metabolism (appeared in 2 genes only)
      - extra info: co-annotation with “protein phosphorylation” + “cell cycle” + “protein kinase motifs”
      - Zhang related “phosphorylation” to sperm proteins in general
      - GENECODIS finding was confirmed by other studies in literature.

Implementation

- Free web-based tool
- Users can upload gene lists
  - gene Symbols, Entrez ID, Unigene ID etc
  - duplicated ID-s considered unique
- Sources of annotation
  - NCBI Entrez Gene database → GO annotations
    - Biological Process
    - Cellular Component
    - Molecular Function
  - KEGG database
    - metabolic pathways
  - Swiss-Prot database
    - Swiss-Prot keywords
    - InterPro motifs
- Supported organisms
  - Total 11; including Arabidopsis thaliana, Danio rerio, Homo sapiens
- Computing power: 16 processors

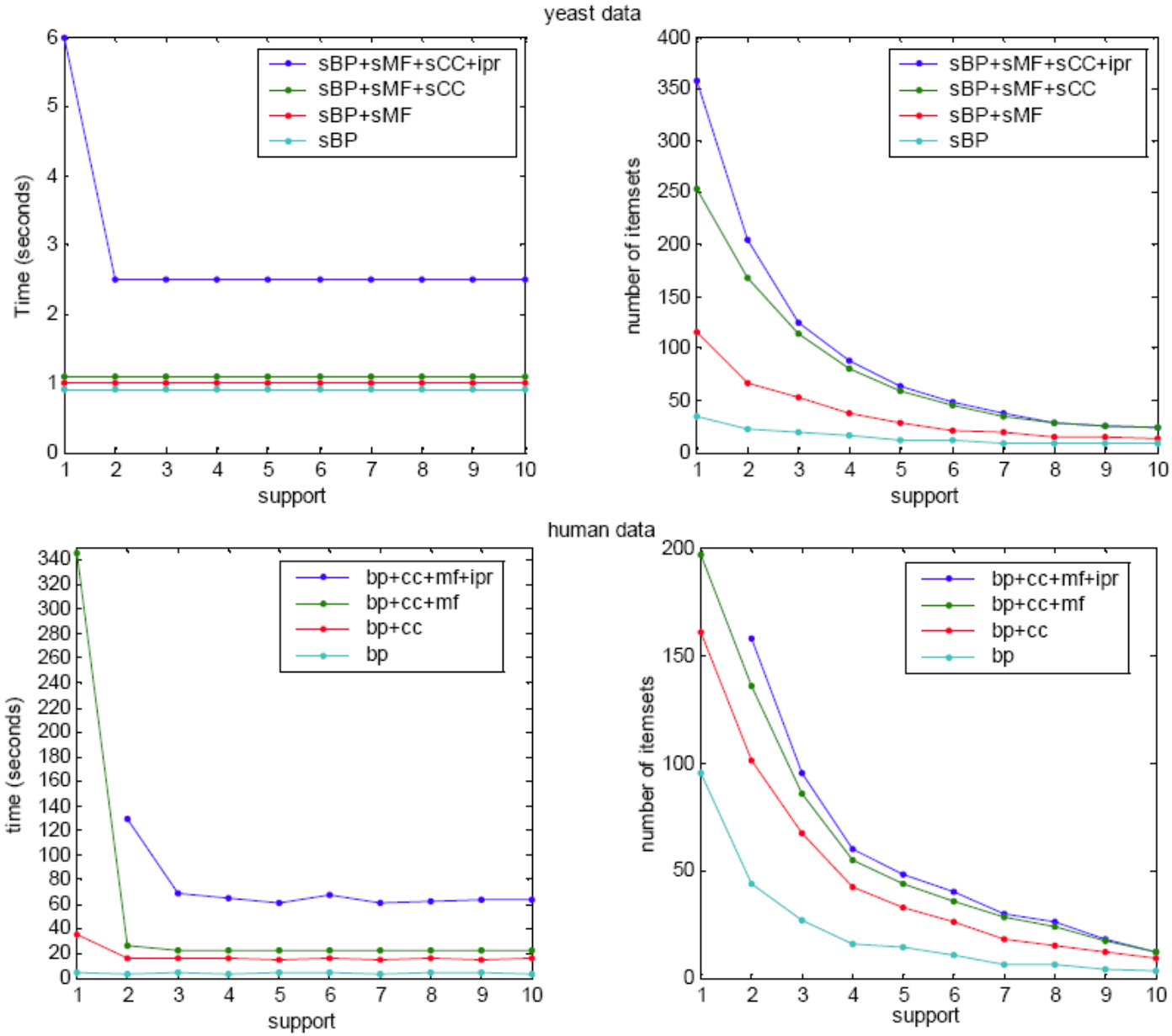
Other methods

Tool	Statistical Model (MTP)	Annotations supported	Organisms	Scope	Term co-occurrences
FatiGO+ (Al-Shahrour *et al.*, 2005)	Fisher's exact test (Step-down minP, FDR)	Gene Ontology, KEGG pathways, Interpro Motifs, SwisProt keywords, Transcription factors, cis-regulatory elements	*A. thaliana* *C. elegans* *D. melanogaster* *G. gallus* *H. sapiens* *M. musculus* *R. norvegicus* *S. cerevisiae* *S. coelicolor*	Multiple categories	No
Onto-Express (Khatri *et al.*, 2002)	Hypergeometric, Binomial, Fisher's exact test, χ^2 (Sidák, Holm, Bonferroni, FDR)	Gene Ontology, KEGG pathways, chromosome regions	more than 20 organisms	Multiple categories	No
GeneMerge* (Castillo-Davis *et al.*, 2003)	Hypergometric (Bonferroni)	Gene Ontology, KEGG pathways, Chromosomal Location	20 different organisms	One category	No
DAVID 2006 (Dennis *et al.*, 2003)	Fisher's exact test (None)	Gene Ontology, Protein Domains, Pathways, General Annotations, Functional Categories, Functional Interaction, Literature Diseases	more than 20 organisms	Multiple categories	No
WebGestalt** (Zhang *et al.*, 2005)	Hypergometric, Fisher's exact test (None)	Gene Ontology, KEGG pathways, BioCarta pathways, Protein Domains	*H. sapiens* *M. musculus*	Multiple Categories	No
GENECODIS	Hypergometric, χ^2 (simulation, FDR)	Gene Ontology, KEGG pathways, Interpro Motifs, SwisProt keywords	*A. thaliana* *B. taurus* *C. elegans* *D. melanogaster* *D. rerio* *G. gallus* *H. sapiens* *M. musculus* *R. norvegicus* *S. cerevisiae* *S. Pombe*	Multiple categories	Yes

Conclusion

- Importance of ontological analysis of such genelists is provent.
- Most current methods generate statistical scores for single annotations.
- GENECODIS provides statistical scores also for combinations of annotations.
- There is no other like this!

Execution time for real data



Minimum support value vs maximum length of combination