

An example for using GoPipe, inputs and outputs

[Version 1.00 example](#)

[New features of Version 1.21](#)

Version 1.00

1. Make sure your current directory is “../GoPipe_1.00/bin” directory
2. Input file formats:

Files of two formats can be used as inputs: XML output files from InterProScan and generic Blast output files against SWISS-PROT and Trembl databases.

InterProScan output files can be gained by feeding InterProScan with your sequences, both online (<http://www.ebi.ac.uk/InterProScan>) or locally (<ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/>).

Blast output files can be gained by feeding your sequences to a local Blast program against SWISS-PROT and Trembl databases. SWISS-PROT and Trembl sequences of Fasta format should be download at ftp://ftp.ebi.ac.uk/pub/databases/sp_tr_nrdb/fasta.
3. There are two example files in the “input” directory for trying: blast_result is a blast result file against SWISS-PROT and Trembl database; ipr_result is a joint of several InterProScan result files of XML format.

We use “blast_result” as an input and set the output folder name as “blast_out”.

Then type:

```
./gopipe.pl -i blast_result -o blast_out -n 6 -e 0.01
```

Here “-i” is for the names of input files, “-o” for the folder name of output files, “-e” for the E value cut-off, and “-n” is for maximum number of blast hits for searching Go-Ids.

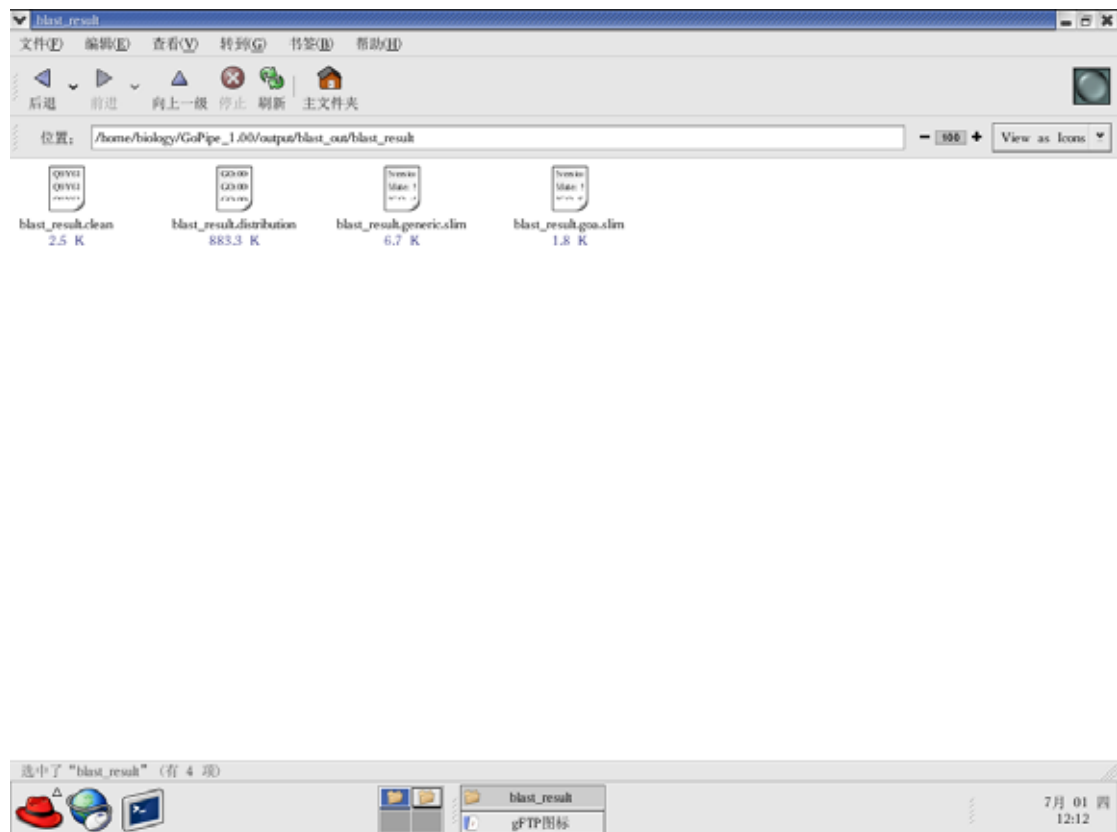
```
biology@localhost:~/GoPipe_1.00/bin
[biology@localhost bin]$ pwd
/home/biology/GoPipe_1.00/bin
[biology@localhost bin]$ ./gopipe.pl -i blast_result -o blast_out -n 6 -e 0.01
GoPipe version 1.0

N_Terms_Blast: 6
E_Value: 0.01

Parsing Blast and/or InterProScan results
.....
(parsing Blast results ... [done]) [done]
21 seconds

Removing redundancies
.....
[done]
3 seconds
```

The output files are at “../GoPipe_1.00/output/blast_out/blast_result” (each of them will be introduced below)



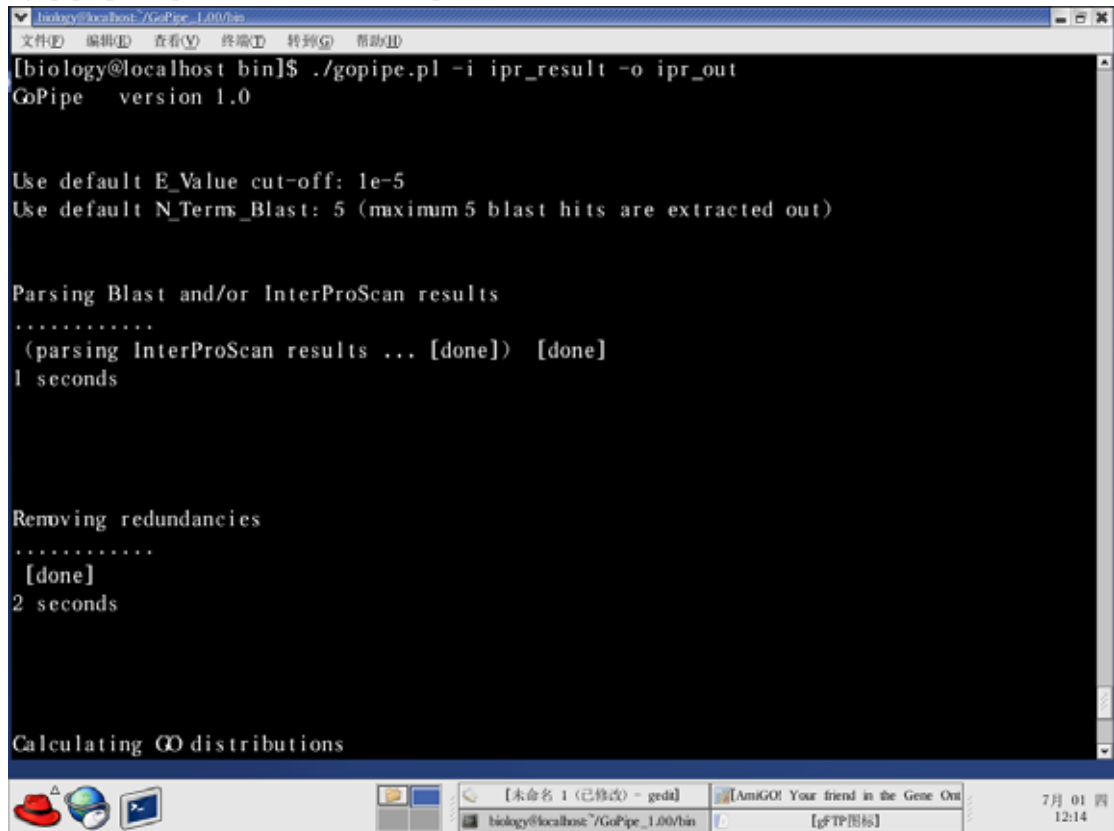
4. Then we use “ipr_result” as an input and set the output folder name as “ipr_out”.

Then type:

```
./gopipe.pl -i ipr_result -o ipr_out
```

(The number of input files are unlimited, and GoPipe will analyze all the input files. You can type such as:

```
./gopipe.pl -i ipr_result blast_out -o ipr_blast_out)
```



```
biology@localhost: /GoPipe_1.00/bin
[biology@localhost bin]$ ./gopipe.pl -i ipr_result -o ipr_out
GoPipe version 1.0

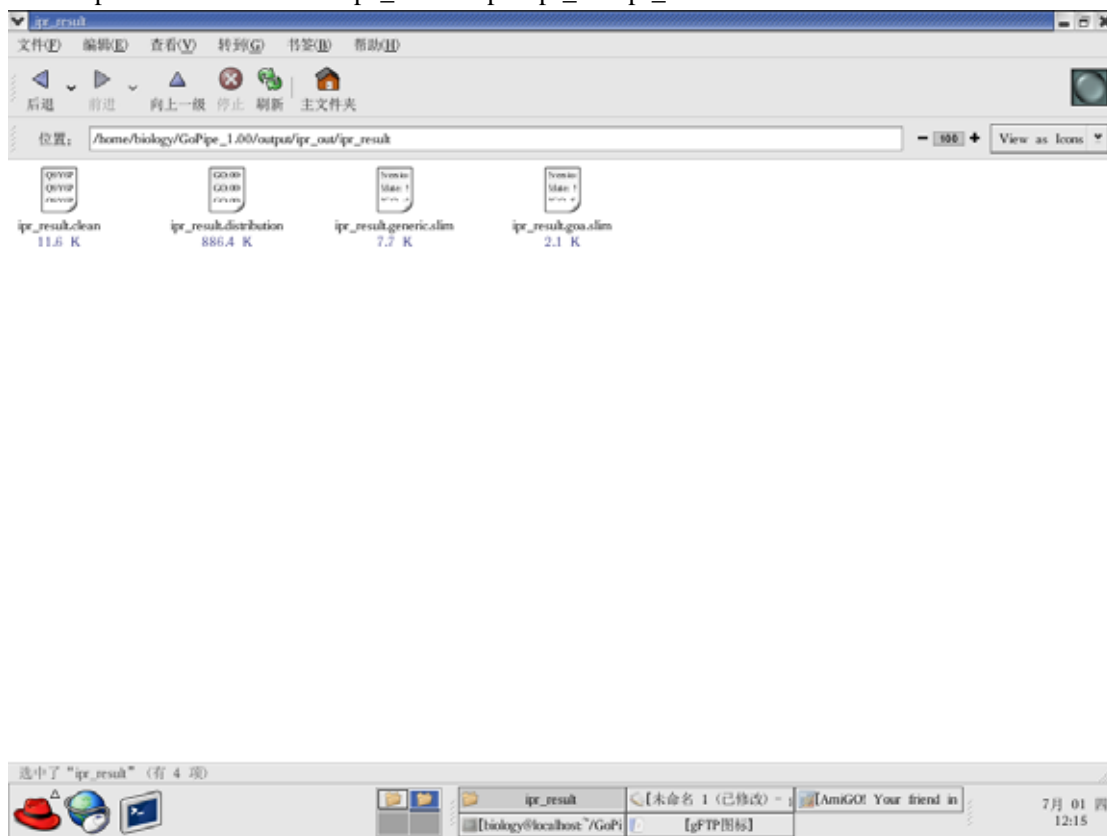
Use default E_Value cut-off: 1e-5
Use default N_Terms_Blast: 5 (maximum 5 blast hits are extracted out)

Parsing Blast and/or InterProScan results
.....
(parsing InterProScan results ... [done]) [done]
1 seconds

Removing redundancies
.....
[done]
2 seconds

Calculating GO distributions
```

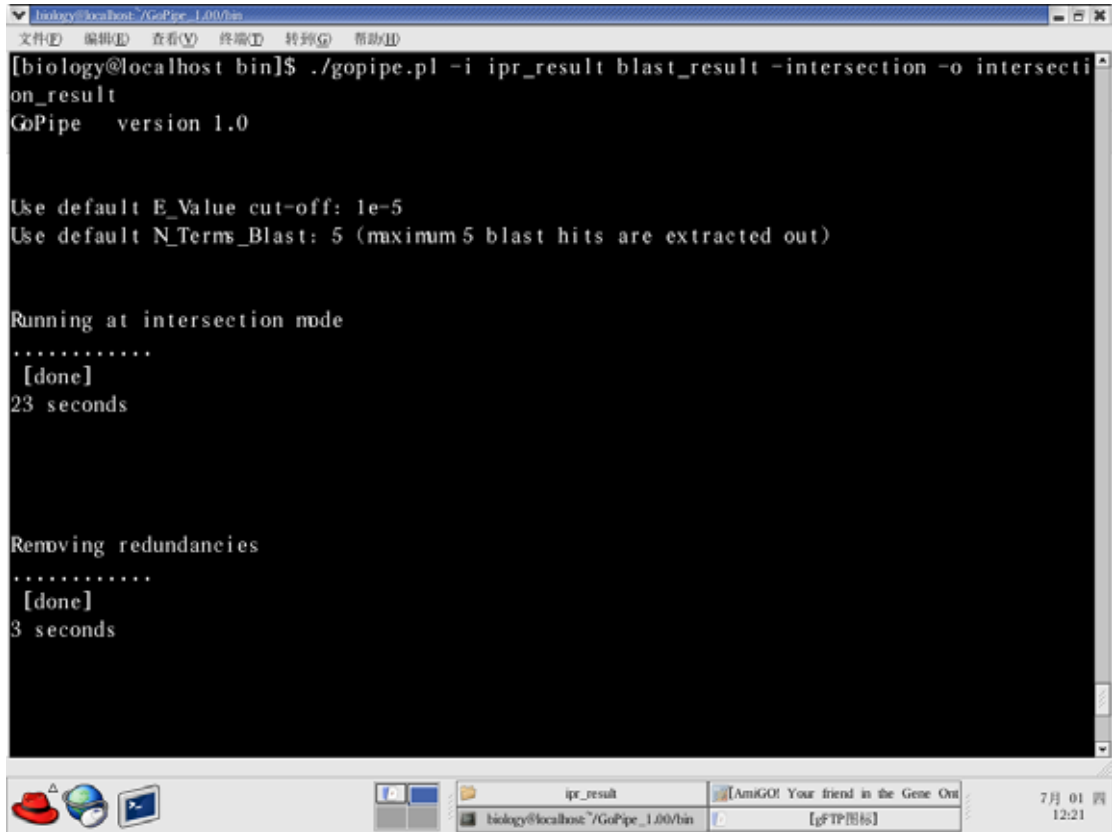
The output files are at “./GoPipe_1.00/output/ipr_out/ipr_result”



5. Run GoPipe at intersection mode to get high accuracy GO annotation

```
./gopipe.pl -i ipr_result blast_result -intersection -o intersection_result
```

Two files should be fed into GoPipe at intersection mode, one is an InterProScan output file, and the other should be a Blast output file



```
biology@localhost:~/GoPipe_1.00/bin
[biology@localhost bin]$ ./gopipe.pl -i ipr_result blast_result -intersection -o intersection_result
GoPipe version 1.0

Use default E_Value cut-off: 1e-5
Use default N_Terms_Blast: 5 (maximum 5 blast hits are extracted out)

Running at intersection mode
.....
[done]
23 seconds

Removing redundancies
.....
[done]
3 seconds
```

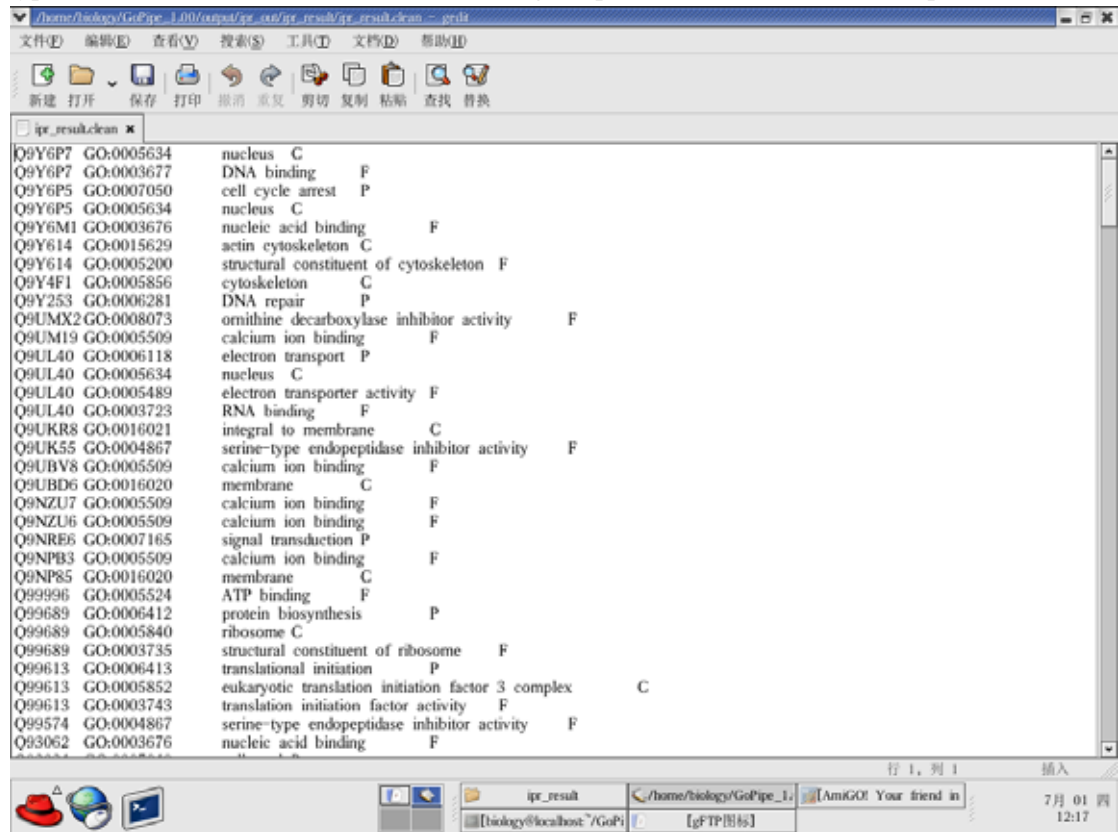
The output files are at
“../GoPipe_1.00/output/intersection_result/ipr_result”



6. The format of the four output files (The output files are in same formats in spite of different inputs)

a) Primary result of GoPipe

Columns from left to right are: sequence name, Go-Id, description of Go-Id, Go-Id type("F" for molecular function, "P" for biological process, "C" for cellular component)



b) Distributions for each GO-Id

Columns from left to right are:

Go-Id, the number of sequences that have been annotated to that GO term, the proportion of these sequences to all the sequences that have at least GO annotation, description for that Go-Id, type of GO

Go-Id	Count	Proportion	Description	Type
GO:0000068	0	0	chromosome condensation	P
GO:0000069	0	0	centromere/kinetochore complex maturation	P
GO:0000070	0	0	mitotic chromosome segregation	P
GO:0000071	0	0	mitotic spindle assembly (sensu Saccharomyces)	P
GO:0000072	0	0	M-phase specific microtubule process	P
GO:0000073	0	0	spindle pole body separation (sensu Saccharomyces)	P
GO:0000074	2	0.0111111111111111	regulation of cell cycle	P
GO:0000075	0	0	cell cycle checkpoint	P
GO:0000076	0	0	DNA replication checkpoint	P
GO:0000077	0	0	DNA damage response, signal transduction resulting in cell cycle arrest	P
GO:0000078	0	0	cell morphogenesis checkpoint	P
GO:0000079	0	0	regulation of CDK activity	P
GO:0000080	0	0	G1 phase of mitotic cell cycle	P
GO:0000082	0	0	G1/S transition of mitotic cell cycle	P
GO:0000083	0	0	G1/S-specific transcription in mitotic cell cycle	P
GO:0000084	1	0.0055555555555556	S phase of mitotic cell cycle	P
GO:0000085	0	0	G2 phase of mitotic cell cycle	P
GO:0000086	0	0	G2/M transition of mitotic cell cycle	P
GO:0000087	0	0	M phase of mitotic cell cycle	P
GO:0000088	0	0	mitotic prophase	P
GO:0000089	0	0	mitotic metaphase	P
GO:0000090	0	0	mitotic anaphase	P
GO:0000091	0	0	mitotic anaphase AP	P
GO:0000092	0	0	mitotic anaphase BP	P
GO:0000093	0	0	mitotic telophase	P
GO:0000094	0	0	septin assembly and septum formation	P
GO:0000095	0	0	S-adenosylmethionine transporter activity	F
GO:0000096	0	0	sulfur amino acid metabolism	P
GO:0000097	0	0	sulfur amino acid biosynthesis	P
GO:0000098	0	0	sulfur amino acid catabolism	P
GO:0000099	0	0	sulfur amino acid transporter activity	F
GO:000100	0	0	S-methylmethionine transporter activity	F
GO:000101	0	0	sulfur amino acid transport	P

c) Distributions for GO Slim (generic GO Slim) terms

```

/home/biology/GoPipe_1.00/output/06/ipr_result/ipr_result_generic_slim - .prdt
文件(F) 编辑(E) 查看(V) 搜索(S) 工具(T) 文档(D) 帮助(H)
新建 打开 保存 打印 取消 重复 剪切 复制 粘贴 查找 替换
ipr_result_generic_slim x
!version: $Revision: 1.7 $
!date: $Date: 2003/12/03 12:07:37 $
!GO_slim_name: generic go_slim
!GO_slim_version: 1.0
!GO_slim_date: 20020826
!GO_slim_authors: Suparna Mundodi and Amelia Ireland
!GO_slim_contact: smundodi@acoma.stanford.edu
!type: % is_a is_a
!type: < part_of Part of
$Gene_Ontology : GO:0003673      180      1
%biological_process : GO:0008150  52      0.288888888888889
%behavior : GO:0007610  0      0
%biological_process unknown : GO:0000004  0      0
%cell communication : GO:000715413  0.07222222222222222
%cell recognition : GO:0008037  0      0
%cell-cell signaling : GO:0007267  1      0.00555555555555556
%host-pathogen interaction : GO:0030383  0      0
%response to endogenous stimulus : GO:0009719  6      0.03333333333333333
%response to external stimulus : GO:0009605  7      0.03888888888888889
%response to abiotic stimulus : GO:0009628  1      0.00555555555555556
%response to biotic stimulus : GO:0009607  5      0.02777777777777778
%signal transduction : GO:0007165  12     0.06666666666666667
%cell growth and/or maintenance : GO:0008151  12     0.06666666666666667
%cell cycle : GO:0007049  4      0.02222222222222222
%cell growth : GO:0016049  0      0
%cell organization and biogenesis : GO:0016043  2      0.01111111111111111
%cytoplasm organization and biogenesis : GO:0007028  2      0.01111111111111111
%organelle organization and biogenesis : GO:0006996  2      0.01111111111111111
%mitochondrion organization and biogenesis : GO:0007005  0      0
%cytoskeleton organization and biogenesis : GO:0007010  2      0.01111111111111111
%cell proliferation : GO:0008283  4      0.02222222222222222
%chemi-mechanical coupling : GO:0006943  0      0 ; synonym:mechanochemical coupling
%cell homeostasis : GO:0019725  1      0.00555555555555556

```

d) Distributions for GO Slim (GOA GO Slim) terms

```

/home/biology/GoPipe_1.00/output/go_slim/ipr_result/go_result.go.slim - goSlim
文件(F) 编辑(E) 查看(V) 搜索(S) 工具(T) 文档(D) 帮助(H)
新建 打开 保存 打印 撤消 重复 剪切 复制 粘贴 查找 替换
ipr_result.go.slim x
!version: $Revision: 1.4 $
!date: $Date: 2003/12/03 12:07:37 $
!GO_Slim_name:GOA and whole proteome analysis
!GO_Slim_version: 1.0
!GO_Slim_date: 20021127
!GO_Slim_authors: N.Mulder, M.Pruess
!GO_Slim_author_contact: goa@ebi.ac.uk
!GO_Slim_reference: Brief Bioinform.3:285-295
!type: % is_a is a
!type: < part_of Part of
$Gene_Ontology : GO:0003673      180      1
%biological_process : GO:0008150  52      0.288888888888889
%biological_process unknown : GO:0000004  0      0
%cell communication : GO:0007154 13      0.072222222222222
%cell growth and/or maintenance : GO:0008151 12      0.066666666666667
%cell cycle : GO:0007049      4      0.022222222222222 ; synonym:cell-division cycle
%cell motility : GO:0006928    0      0
%metabolism : GO:0008152     26      0.144444444444444
%response to stress : GO:0006950 6      0.033333333333333
%transport : GO:00068105     0.027777777777778
%death : GO:0016265          0      0
%development : GO:0007275    0      0
%physiological processes : GO:0007582  44      0.244444444444444
%cellular_component : GO:0005575  80      0.444444444444444
%cell : GO:0005623          71      0.394444444444444
%cellular_component unknown : GO:0008372  0      0
%external encapsulating structure : GO:0030312  0      0
%extracellular : GO:0005576   10      0.055555555555556
%unlocalized : GO:0005941    0      0
%molecular_function : GO:0003674  128     0.711111111111111
%cell adhesion molecule activity : GO:0005194  0      0
%chaperone activity : GO:0003754, 2  0.011111111111111 GO:0003757, 0      0 GO:0003758, 0      0 GO:0003760,
0      0 GO:0003761 0      0
    
```

行 1, 列 1 插入

ipr_result /home/biology/GoPipe_1.00 [AmiGO! Your friend in 7月 01 四 12:19
[biology@localhost:~/GoP [gFTP图标]

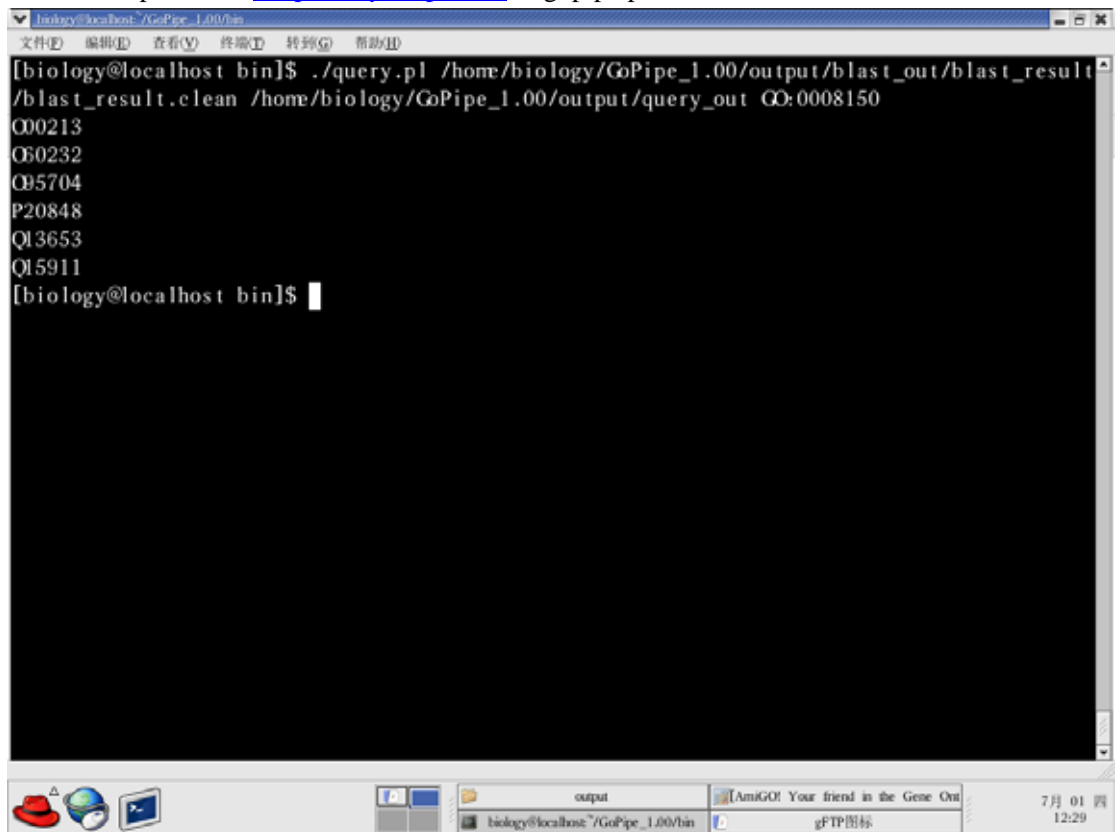
7. Additional tools

- a) query.pl is provided to search sequences annotated to a Go-Id of interest in the primary result file.

```
./query.pl /home/biology/GoPipe_1.00/output/blast_out/blast_result/blast_result.clean  
/home/biology/GoPipe_1.00/output/query_out GO:0008150
```

Three parameters are needed. They are the absolute path of the input file, output file and the Go-Id of interest respectively, from left to right.

Here the input file is [the primary output file](#) of gopipe.pl.



```
biology@localhost:~/GoPipe_1.00/bin
文件(F) 编辑(E) 查看(V) 终端(T) 转到(G) 帮助(H)
[biology@localhost bin]$ ./query.pl /home/biology/GoPipe_1.00/output/blast_out/blast_result/blast_result.clean /home/biology/GoPipe_1.00/output/query_out GO:0008150
GO:00213
GO:0232
GO:05704
P20848
Q13653
Q15911
[biology@localhost bin]$
```

- b) comparecr.pl is provided to compare two sets of GO annotation to reveal GO terms that are differentially distributed

```
./comparecr.pl
```

```
/home/biology/GoPipe_1.00/output/blast_out/blast_result/blast_result.distribution
```

```
/home/biology/GoPipe_1.00/output/ipr_out/ipr_result/ipr_result.distribution
```

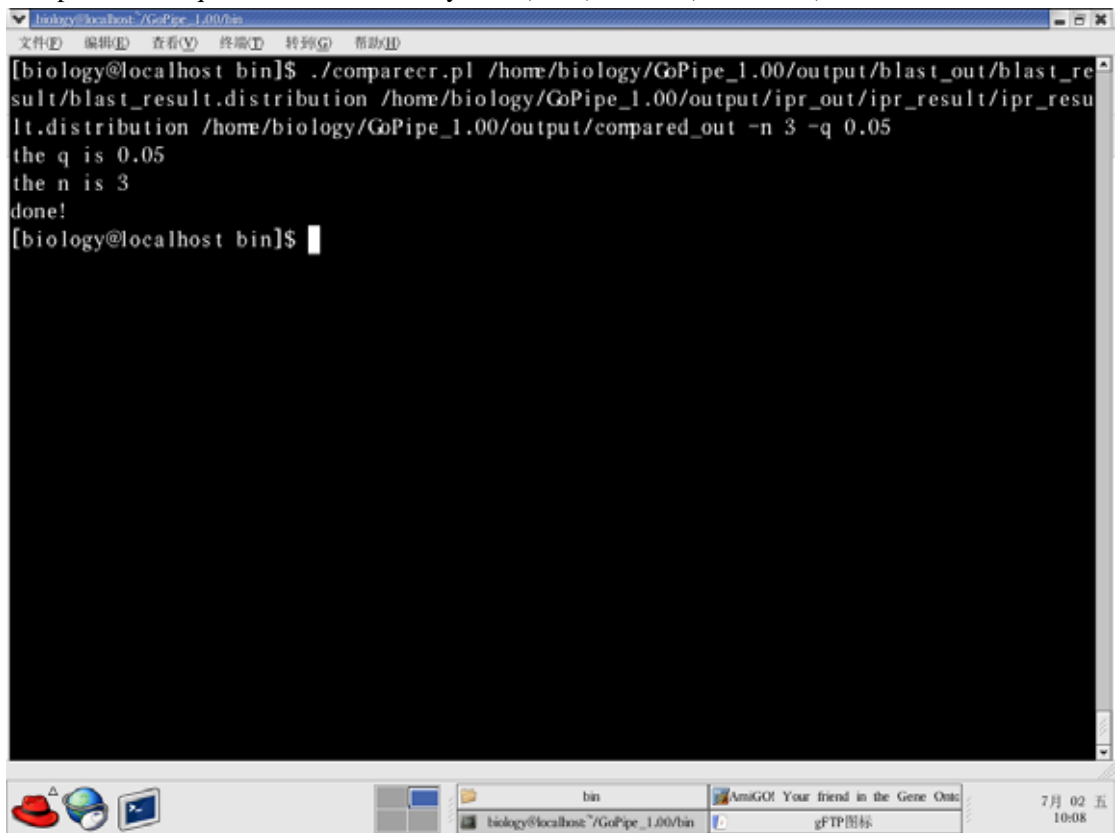
```
/home/biology/GoPipe_1.00/output/blast_out/compared_out -n 3 -q 0.05
```

The first and second parameters are two input files to be compared. They are [the second output files](#) derived by two different inputs of gopipe.pl.

The third parameter is the absolute path of the output.

The parameter -n is the cut-off of the multiple of two GO proportion (the second column in [the second output file](#), the first divided by the second, default ">2 or <1/2")

The parameter -q is the "false discovery rate"(FDR) cut-off (default 0.1)



```
biology@localhost:~/GoPipe_1.00/bin
[biology@localhost bin]$ ./comparecr.pl /home/biology/GoPipe_1.00/output/blast_out/blast_result/blast_result.distribution /home/biology/GoPipe_1.00/output/ipr_out/ipr_result/ipr_result.distribution /home/biology/GoPipe_1.00/output/compared_out -n 3 -q 0.05
the q is 0.05
the n is 3
done!
[biology@localhost bin]$
```

The result file:

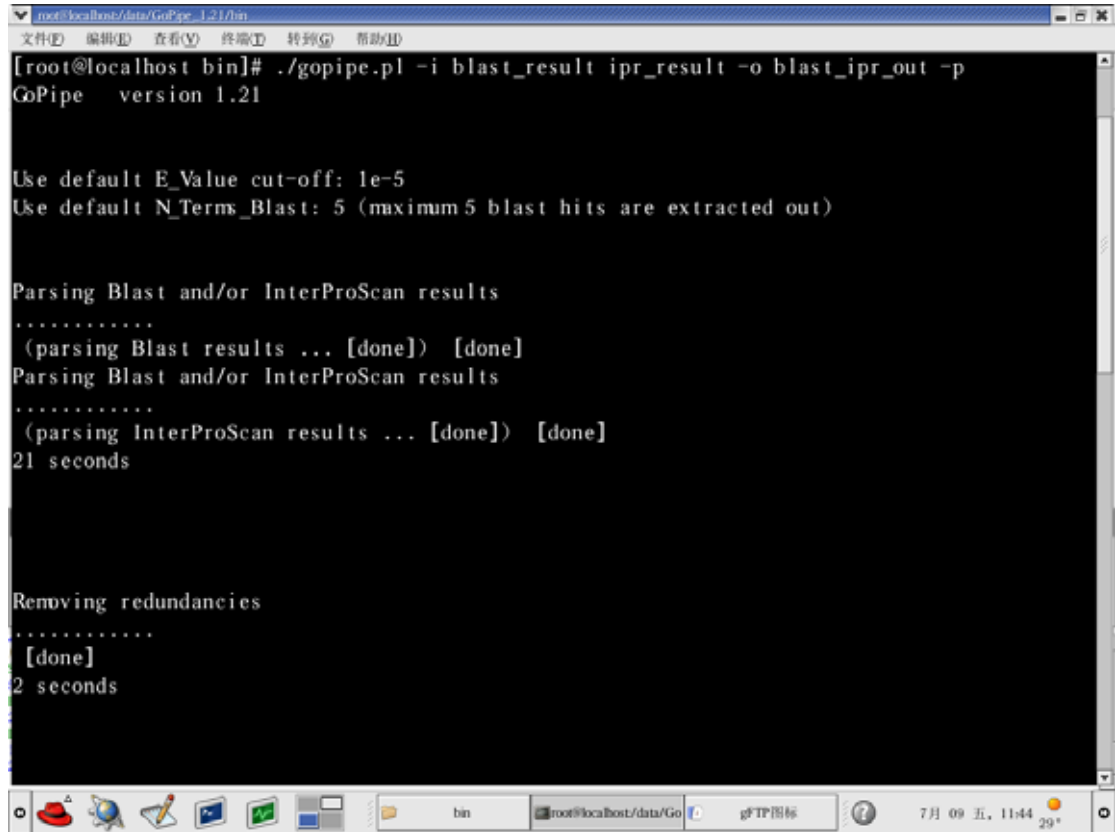
The screenshot shows a Windows Explorer window titled 'compared_out' with the address bar showing the path '/home/biology/GoPipe_1.00/output/compared_out'. The main content area displays a list of GO terms with columns for Go-Id, sequence number(set1), proportion of these sequences(set1), sequence number(set2), proportion of these sequences(set2), P value, corrected P value, description of GO, and GO type. The table is sorted by P value in ascending order.

Go-Id	sequence number(set1)	proportion of these sequences(set1)	sequence number(set2)	proportion of these sequences(set2)	P value	corrected P value	description of GO	GO type	
GO:0007275	4	0.5	0	0	0	1.38875345787211e-06	development	P	
GO:0005886	3	0.375	0	0	0	5.13838779412682e-05	plasma membrane	C	
GO:0006350	3	0.375	0	0	0	5.13838779412682e-05	transcription	P	
GO:0009653	3	0.375	0	0	0	5.13838779412682e-05	morphogenesis	P	
GO:0009887	3	0.375	0	0	0	5.13838779412682e-05	organogenesis	P	
GO:0019222	3	0.375	0	0	0	5.13838779412682e-05	regulation of metabolism	P	
GO:0045449	3	0.375	0	0	0	5.13838779412682e-05	regulation of transcription	P	
GO:0000278	3	0.375	1	0.00555555555555556	0.0148148148148148	0.000201369251391457	0.00694723917300527	mitotic cell cycle	P
GO:0001540	2	0.25	0	0	0	0.00159290021617932	beta-amyloid binding	F	
GO:0001558	2	0.25	0	0	0	0.00159290021617932	regulation of cell growth	P	
GO:0006417	2	0.25	0	0	0	0.00159290021617932	regulation of protein biosynthesis	P	
GO:0007090	2	0.25	0	0	0	0.00159290021617932	regulation of S phase of mitotic cell cycle	P	
GO:0007346	2	0.25	0	0	0	0.00159290021617932	regulation of mitotic cell cycle	P	
GO:0007399	2	0.25	0	0	0	0.00159290021617932	neurogenesis	P	
GO:0007409	2	0.25	0	0	0	0.00159290021617932	axonogenesis	P	
GO:0008134	2	0.25	0	0	0	0.00159290021617932	transcription factor binding	F	
GO:0008372	2	0.25	0	0	0	0.00159290021617932	cellular component unknown	C	
GO:0009889	2	0.25	0	0	0	0.00159290021617932	regulation of biosynthesis	P	
GO:0009890	2	0.25	0	0	0	0.00159290021617932	negative regulation of biosynthesis	P	
GO:0009892	2	0.25	0	0	0	0.00159290021617932	negative regulation of metabolism	P	
GO:0016049	2	0.25	0	0	0	0.00159290021617932	cell growth	P	
GO:0017028	2	0.25	0	0	0	0.00159290021617932	protein stabilization activity	F	
GO:0017148	2	0.25	0	0	0	0.00159290021617932	negative regulation of protein biosynthesis	P	
GO:0019899	2	0.25	0	0	0	0.00159290021617932	enzyme binding	F	
GO:0030027	2	0.25	0	0	0	0.00159290021617932	lamellipodium	C	
GO:0030029	2	0.25	0	0	0	0.00159290021617932	actin filament-based process	P	
GO:0030048	2	0.25	0	0	0	0.00159290021617932	actin filament-based movement	P	
GO:0030308	2	0.25	0	0	0	0.00159290021617932	negative regulation of cell growth	P	
GO:0030426	2	0.25	0	0	0	0.00159290021617932	growth cone	C	
GO:0030427	2	0.25	0	0	0	0.00159290021617932	site of polarized growth	C	
GO:0035055	2	0.25	0	0	0	0.00159290021617932	histone acetyltransferase binding	F	
GO:0045202	2	0.25	0	0	0	0.00159290021617932	synapse	C	
GO:0045749	2	0.25	0	0	0	0.00159290021617932	negative regulation of S phase of mitotic cell cycle	P	
GO:0045786	2	0.25	0	0	0	0.00159290021617932	negative regulation of cell cycle	P	

Columns from left to right are: Go-Id, sequence number(set1) corresponding to that Go-Id, proportion of these sequences(set1), sequence number(set2), proportion of these sequences(set2), P value, corrected P value, description of GO, GO type

New features for Version 1.21

1. A new sub-program draw.pl is provided for plotting graphs for GO Slim distributions
If your GD libraries have been installed in your system, by adding “-p” to the command line, six graphs will be plotted.



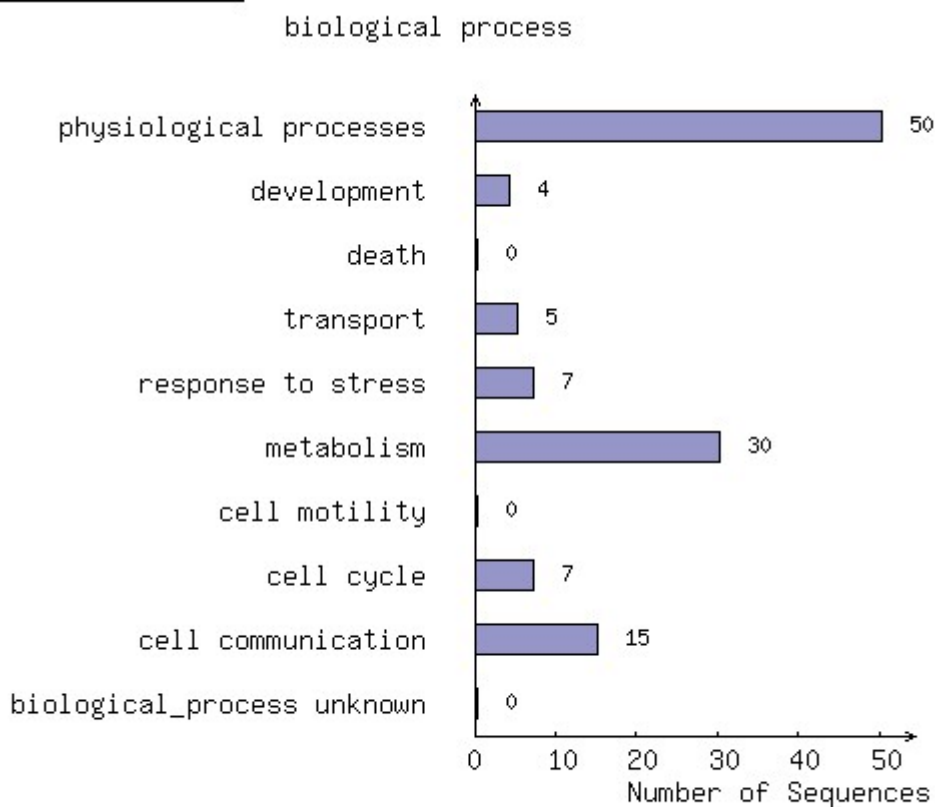
```
root@localhost/data/GoPipe_1.21/bin
[root@localhost bin]# ./gopipe.pl -i blast_result ipr_result -o blast_ipr_out -p
GoPipe version 1.21

Use default E_Value cut-off: 1e-5
Use default N_Term_Blast: 5 (maximum 5 blast hits are extracted out)

Parsing Blast and/or InterProScan results
.....
(parsing Blast results ... [done]) [done]
Parsing Blast and/or InterProScan results
.....
(parsing InterProScan results ... [done]) [done]
21 seconds

Removing redundancies
.....
[done]
2 seconds
```

One of the graphs:



Three modules are required for plotting graphs:

libjpeg : ../install_modules/jpegsrc.v6b.tar.gz

gd lib : ../install_modules/gd-2.0.22.tar.gz

GD.PM : please go to <http://search.cpan.org/~lds/GD-2.12/> for GD.PM lib

If you need not plotting function or you do not want to install these modules, just use GoPipe_1.21 as version 1.00 without “-p” in the command line.

2. Data files update

Term	GO Consortium	07/08/2004
SPTR_GO	GOA	June 4 2004 GOA UniProt 18.0 Released
GO Slims	GO database	2003/12/03