Presentation of the Conflict Resolution Task

Julien Jourde, Wednesday 3rd June 2009

Transys – Training Session

B. subtilis gene names : some figures

- We try to integrate data from **7** different data soures

- There are many variations between them :

Databases	BSU	BG	Names	rRNA	tRNA	Releases
BSORF	4106	4106	5559	0	0	2006-01-18

B. subtilis gene names : some figures

- We try to integrate data from **7** different data soures

- There are many variations between them :

Databases	BSU	BG	Names	rRNA	tRNA	Releases
BSORF	4106	4106	5559	0	0	2006-01-18
genetic map	911	911	1218	0	0	2001

B. subtilis gene names : some figures

- We try to integrate data from **7** different data soures

- There are many variations between them :

Databases	BSU	BG	Names	rRNA	tRNA	Releases
BSORF	4106	4106	5559	0	0	2006-01-18
genetic map	911	911	1218	0	0	2001
EloAnn	4106	4106	5799	0	0	2006-11

B. subtilis gene names : some figures

- We try to integrate data from **7** different data soures

- There are many variations between them :

Databases	BSU	BG	Names	rRNA	tRNA	Releases
BSORF	4106	4106	5559	0	0	2006-01-18
genetic map	911	911	1218	0	0	2001
EloAnn	4106	4106	5799	0	0	2006-11
GenBank	4106	4106	5301	0	0	Genbank 170 // 2009-03-03

B. subtilis gene names : some figures

- We try to integrate data from **7** different data soures

- There are many variations between them :

Databases	BSU	BG	Names	rRNA	tRNA	Releases
BSORF	4106	4106	5559	0	0	2006-01-18
genetic map	911	911	1218	0	0	2001
EloAnn	4106	4106	5799	0	0	2006-11
GenBank	4106	4106	5301	0	0	Genbank 170 // 2009-03-03
Genome Reviews	4107	4106	5567	30	86	Rel 103.0 //2009-03-03

B. subtilis gene names : some figures

- We try to integrate data from **7** different data soures

- There are many variations between them :

Databases	BSU	BG	Names	rRNA	tRNA	Releases
BSORF	4106	4106	5559	0	0	2006-01-18
genetic map	911	911	1218	0	0	2001
EloAnn	4106	4106	5799	0	0	2006-11
GenBank	4106	4106	5301	0	0	Genbank 170 // 2009-03-03
Genome Reviews	4107	4106	5567	30	86	Rel 103.0 //2009-03-03
GenoList	4244	4036	6151	30	86	R17.1 // 2009-04-10

B. subtilis gene names : some figures

- We try to integrate data from **7** different data soures

- There are many variations between them :

Databases	BSU	BG	Names	rRNA	tRNA	Releases
BSORF	4106	4106	5559	0	0	2006-01-18
genetic map	911	911	1218	0	0	2001
EloAnn	4106	4106	5799	0	0	2006-11
GenBank	4106	4106	5301	0	0	Genbank 170 // 2009-03-03
Genome Reviews	4107	4106	5567	30	86	Rel 103.0 //2009-03-03
GenoList	4244	4036	6151	30	86	R17.1 // 2009-04-10
Swissprot	4107	4106	5778	0	0	UniProt 15.0 // 2009-03-24 (Swiss-Prot 57.0 + Trembl 40.0)

B. subtilis gene names : some figures

- We try to integrate data from **7** different data soures

- There are many variations between them :

Databases	BSU	BG	Names	rRNA	tRNA	Releases
BSORF	4106	4106	5559	0	0	2006-01-18
genetic map	911	911	1218	0	0	2001
EloAnn	4106	4106	5799	0	0	2006-11
GenBank	4106	4106	5301	0	0	Genbank 170 // 2009-03-03
Genome Reviews	**4107**	4106	5567	**30**	**86**	Rel 103.0 //2009-03-03
GenoList	**4244**	**4036**	6151	**30**	**86**	R17.1 // 2009-04-10
Swissprot	**4107**	4106	5778	0	0	UniProt 15.0 // 2009-03-24 (Swiss-Prot 57.0 + Trembl 40.0)

B. subtilis gene names : some figures

- We try to integrate data from **7** different data soures

- There are many variations between them :

Databases	BSU	BG	Names	rRNA	tRNA	Releases
BSORF	4106	4106	**5559**	0	0	2006-01-18
genetic map	911	911	**1218**	0	0	2001
EloAnn	4106	4106	**5799**	0	0	2006-11
GenBank	4106	4106	**5301**	0	0	Genbank 170 // 2009-03-03
Genome Reviews	4107	4106	**5567**	30	86	Rel 103.0 //2009-03-03
GenoList	4244	4036	**6151**	30	86	R17.1 // 2009-04-10
Swissprot	4107	4106	**5778**	0	0	UniProt 15.0 // 2009-03-24 (Swiss-Prot 57.0 + Trembl 40.0)

Stability between databases ?

• There are only **193** couples (bsu,names) existing at the same time in all databases. (All databases represent 7787 names)

• Few informations about tRNA and rRNA are recent and in only 2 databases.

• **Swissprot** and **Genome Reviews** try to update similar data (the new BSU is the same one).

• **GenoList** is completely different. It contains the last annotation of the genome :

- **70** BSU/BG ids have been deleted.
- **208** BSU (without BG ids) have been added.
- Pseudogenes are now annotated and have several BSU ids :

BSU06073	ydzW	putative phosphomannomutase; C-terminal part of YdzW
BSU06074	ydzW	putative phosphomannomutase; internal part of YdzW
BSU06076	ydzW	putative phosphomannomutase; internal part of YdzW
BSU06077	ydzW	putative phosphomannomutase; internal part of YdzW
BSU06078	ydzW	putative phosphomannomutase; internal part of YdzW
BSU06079	ydzW	putative phosphomannomutase; internal part of YdzW
BSU06083	ydzW	putative phosphomannomutase; N-terminal part of YdzW

Conflicts between databases !

- Characterization of the different conflicts ?

  - One name missing in only one base
  - One name missing in N bases
  - One name missing or not and reused for others BSU
  - One name missing or not, used in different BSU in different bases
  - ...

Many possibilities. **One of your task will be to characterize many conflicts as you can**.

Strategy used to cure the base

- Database files :

  - **BSU00010** BG10065 **dnaA** **dnaH** **dnaJ** **dnaK**
  - **BSU00020** BG10066 **dnaN** **dnaG** **dnaK**
  - **BSU00030** BG10067 **yaaA**

- Strategies ?

  - Find current name AND synonyms at the same time.

  - Cure in two times :

    1. Find all synonyms for each BSU (with **dates of renaming** !)
    2. Determine which name is the current one

Where to find decisive evidences ?

Another task for you today will be to design an effective method to find evidences.

You can use databases :

- BSORF
- GenBank
- KEGG
- GenoList
- Swissprot
- Gene lists
- ...

You can use tools :

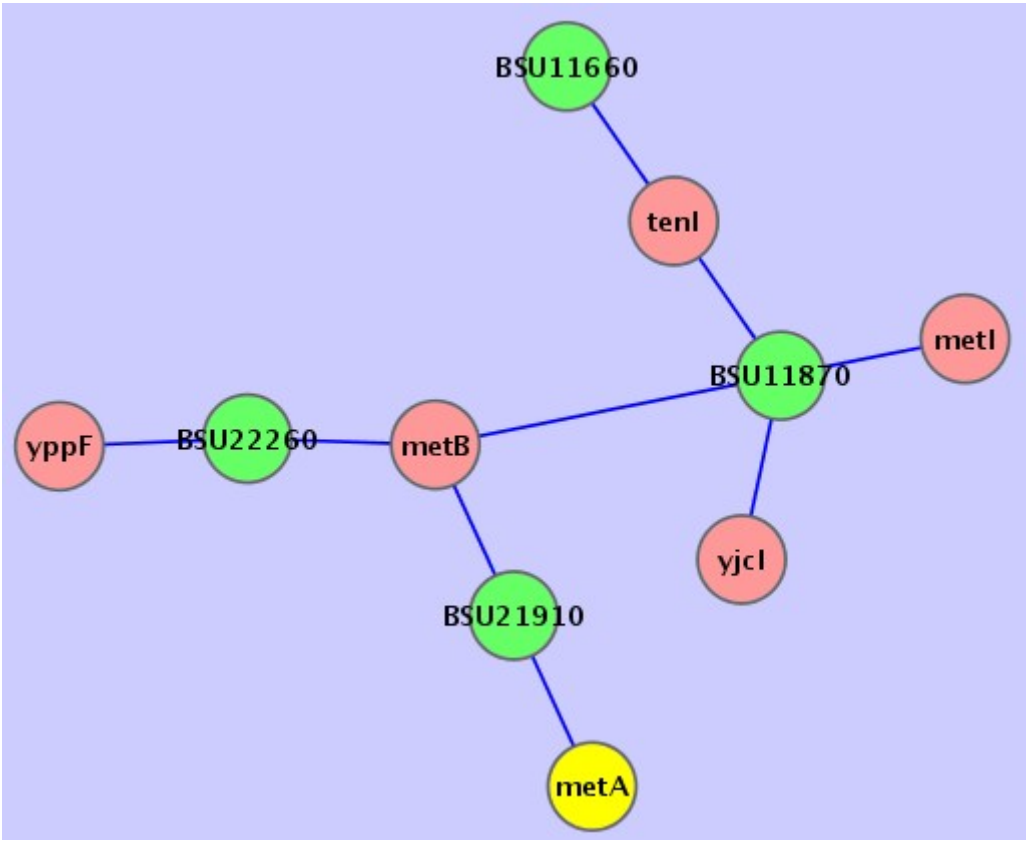
- PubMed
- Google Scholar
- Blast
- ...

A simple case to start

We selected **3** names which are not reused and are missing in **BSORF** only :

- **BSU15600** cysC
- **BSU34990** gerF
- **BSU21910** metA

When it will be done, you will be able to extend your work to the other synonyms of those BSU. It's the only way to do : go step-by-step

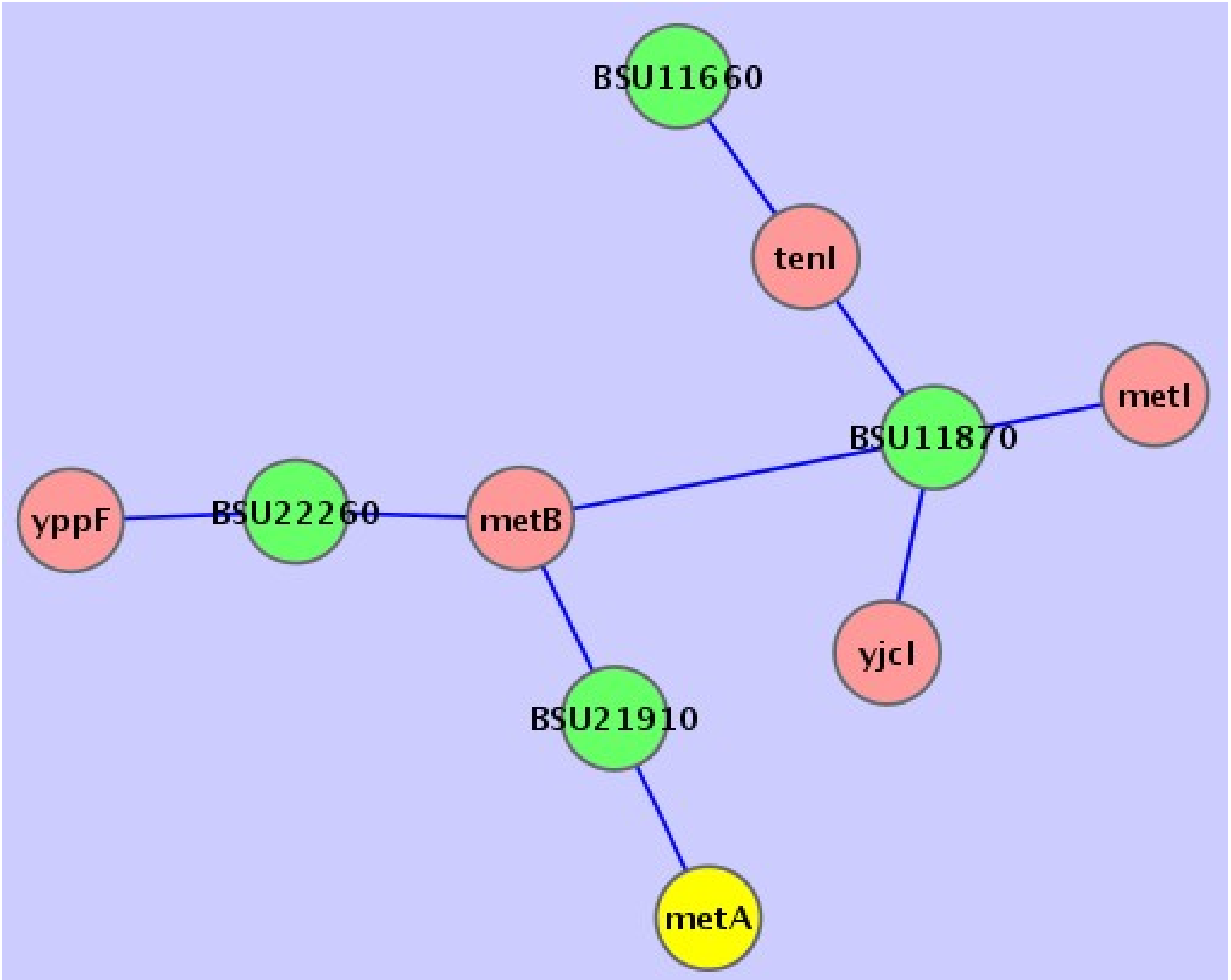


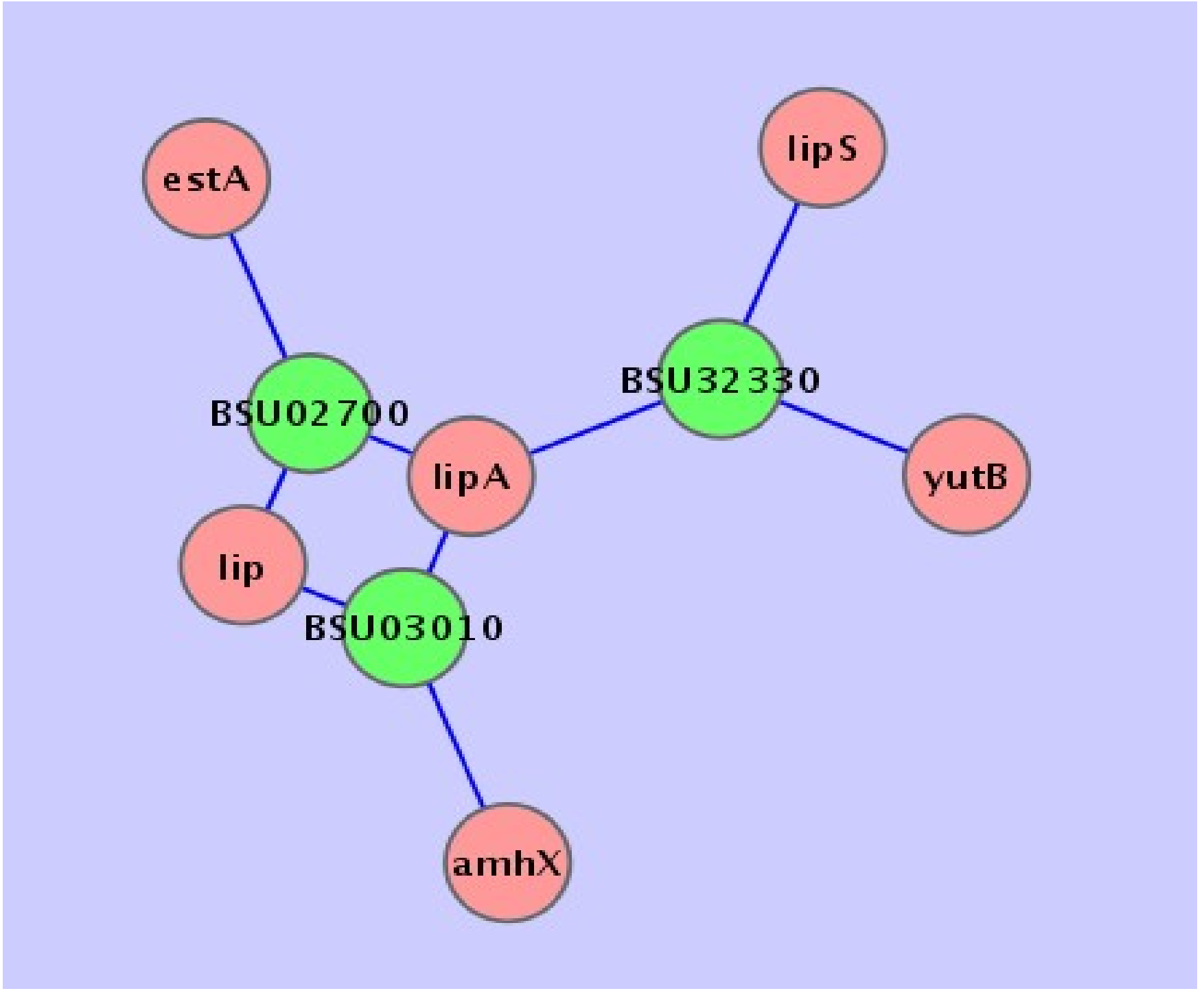
Exercise 1

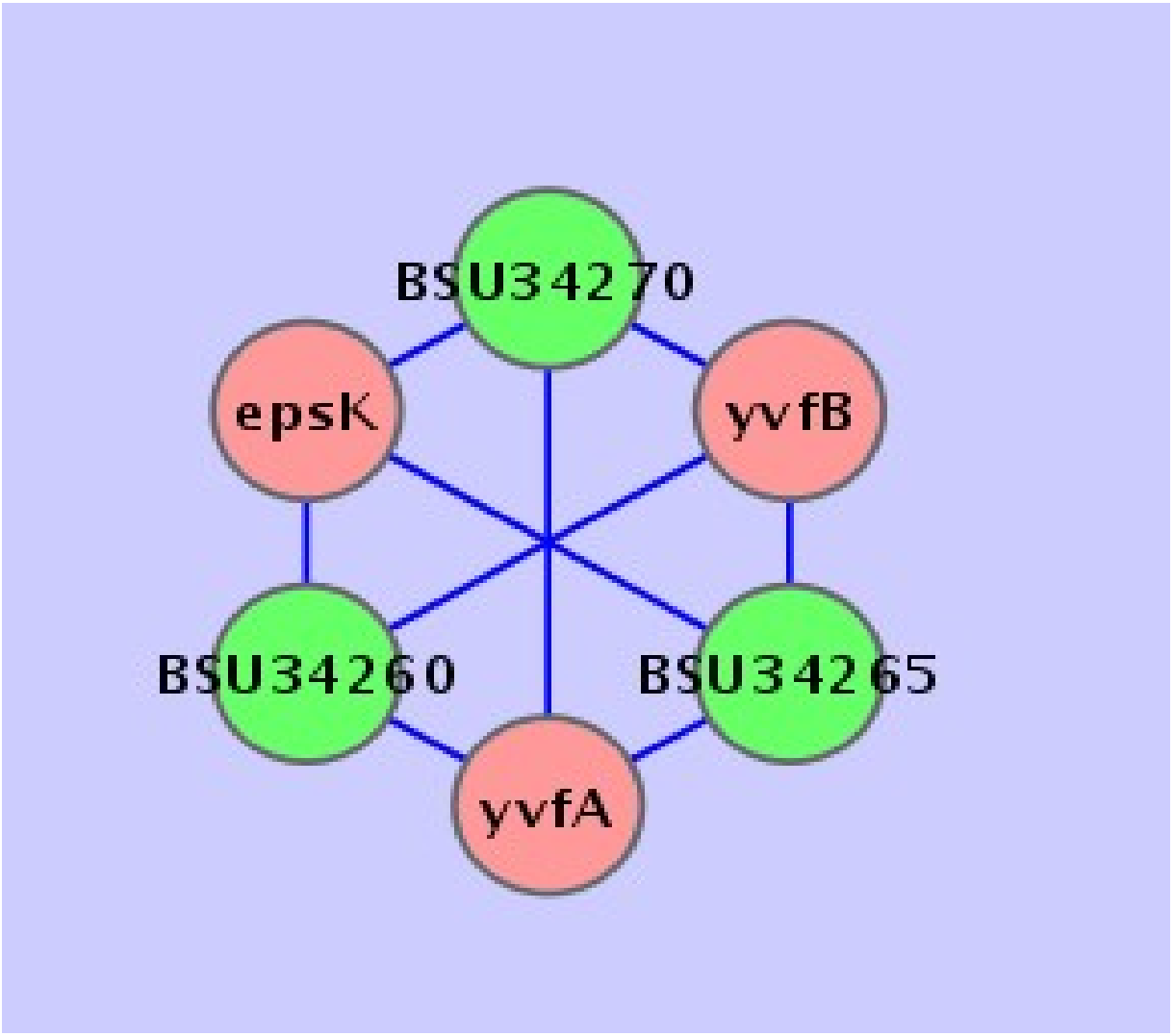
BSORF	BSU15600	BG13379	cycC	ylnC	
Genetic Map	BSU15600	BG13379	pyrE	pyrX	**cysC**
EloAnn	BSU15600	BG13379	**cysC**	ylnC	
GenBank	BSU15600	BG13379	**cysC**	ylnC	
Genome Rev	BSU15600	BG13379	**cysC**	ylnC	
GenoList	BSU15600	BG13379	**cysC**	ylnC	
Swissprot	BSU15600	BG13379	**cysC**	ylnC	

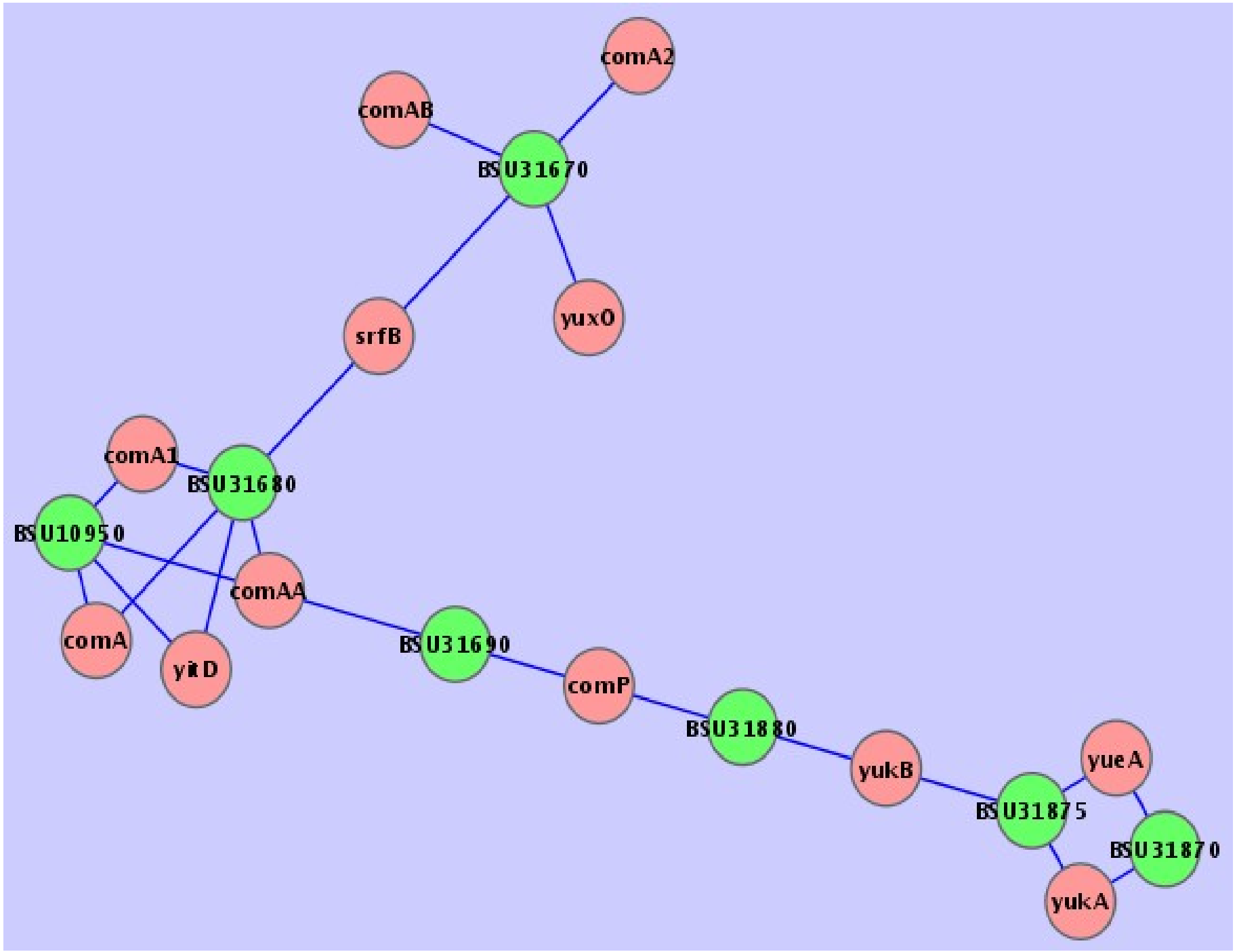
BSORF	BSU34990	BG12611	lgt		
Genetic Map	BSU34990	BG12611	**gerF**		
EloAnn	BSU34990	BG12611	lgt	**gerF**	yvoC
GenBank	BSU34990	BG12611	lgt	**gerF**	yvoC
Genome Rev	BSU34990	BG12611	lgt	**gerF**	yvoC
GenoList	BSU34990	BG12611	lgt	**gerF**	
Swissprot	BSU34990	BG12611	lgt	**gerF**	

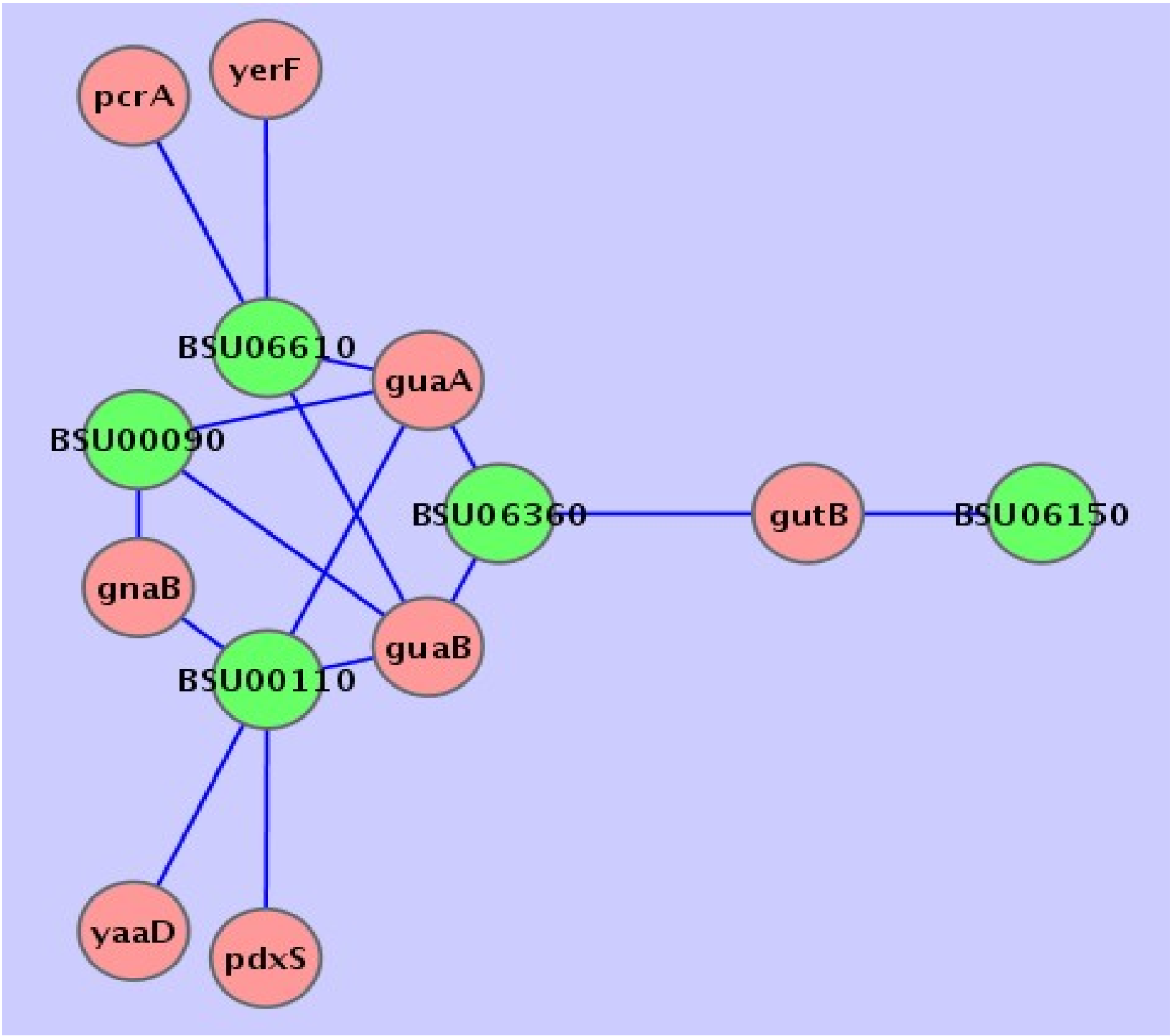
BSORF	BSU21910	BG11534	metB	
Genetic Map	BSU21910	BG11534	**metA**	
EloAnn	BSU21910	BG11534	**metA**	metB
GenBank	BSU21910	BG11534	**metA**	metB
Genome Rev	BSU21910	BG11534	**metA**	metB
GenoList	BSU21910	BG11534	**metA**	metB
Swissprot	BSU21910	BG11534	**metA**	metB

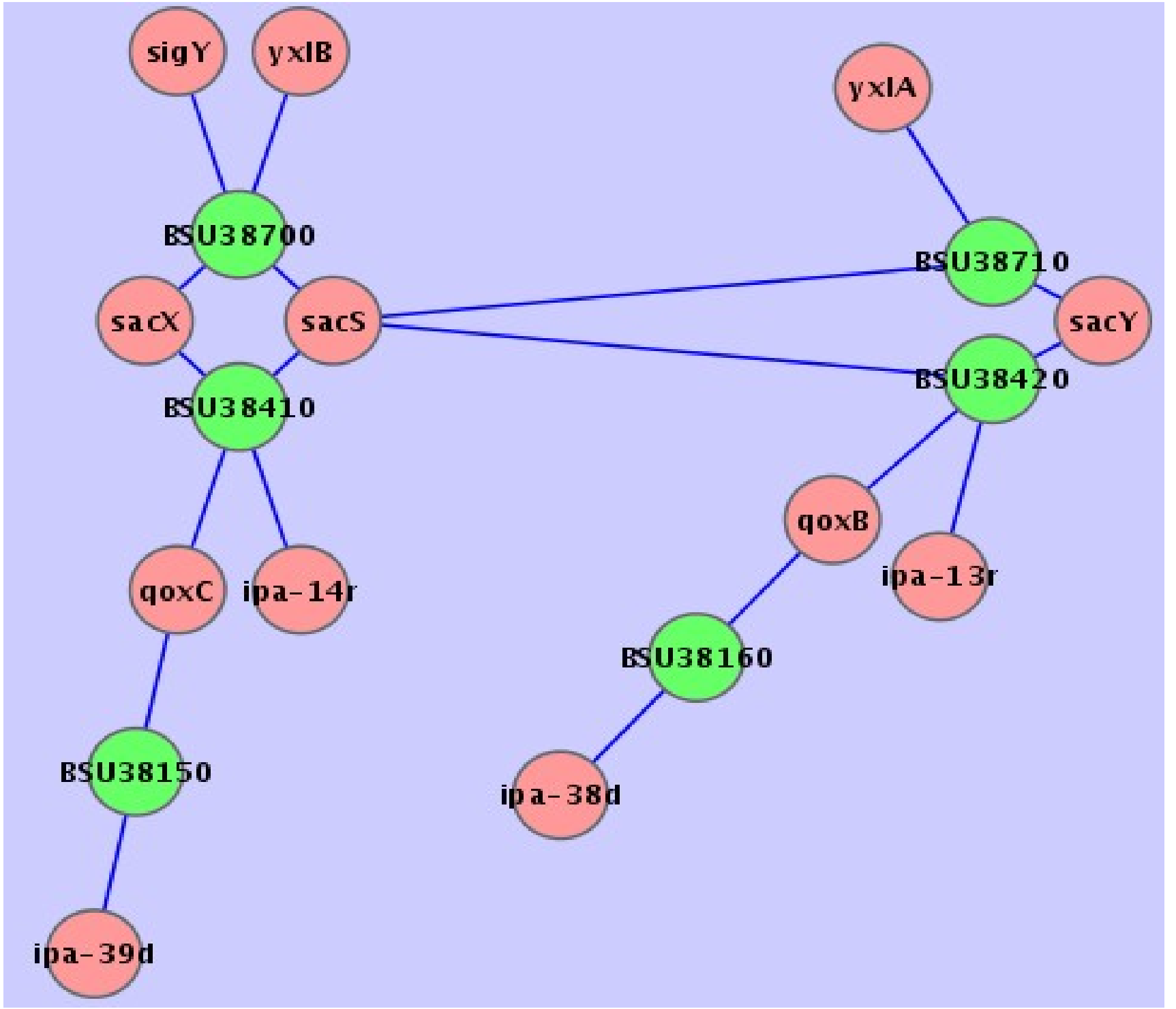

BSU11660
tenI
metI
BSU11870
yppF
BSU22260
metB
yjcI
BSU21910
metA


estA
lipS
BSU02700
BSU32330
lipA
yutB
lip
BSU03010
amhX


BSU34270
epsK
yvfB
BSU34260
BSU34265
yvfA


comA2
comAB
BSU31670
srfB
yuxO
comA1
BSU31680
BSU10950
comAA
comA
ykD
BSU31690
comP
BSU31880
yukB
yueA
BSU31875
BSU31870
yukA


pcrA
yerF
BSU06610
guaA
BSU00090
BSU06360
gutB
BSU06150
gnaB
guaB
BSU00110
yaaD
pdxS


sigY
yxlB
yxlA
BSU38700
BSU38710
sacX
sacS
sacY
BSU38420
BSU38410
qoxB
qoxC
ipa-14r
ipa-13r
BSU38160
BSU38150
ipa-38d
ipa-39d