STRING database : protein interactions Mappings, Triple/Publication generation

Data Integration Activities

The following activities have to take place:

1. Create Mapping collection
2. Create Triples, with measures
3. Create Publications, with measures

The input file name is protein.actions.v10.txt.gz

The URL http://string-db.org/download/protein.actions.v10.txt.gz

Dataset / catalog level metadata

Dataset title: STRING

Dataset description:

Download URL: http://string-db.org/download/protein.actions.v10.txt.gz

Release/version: 10.0

Release issue date: current: since Apr 16, 2016

Download date/time: 01-11-2016

distribution format: text file TSV format

Data Record Metadata

The file is redundant. If the the action goes in the other direction, then this will be indicated at another line where the name identifiers are swapped between the 1st and the 2nd column.

#	name	description
1	item_id_a	identifier of protein A
2	item_id_b	identifier of protein B
3	mode	type of interaction (e.g. "reaction", "expression", "activation", "ptmod"(post-translational modifications), "binding", "catalysis")
4	action	the effect of the action ("inhibition", "activation")
5	a_is_acting	The directionality of the action if applicable (1 gives that item_id_a is acting upon item_id_b) . If the column a_is_acting is 1 (TRUE) then this means that protein_a is acting on protein_b. On the other hand, if it is 0 (FALSE) then the opposite is not necessarily true. In this case the zero can indicate that directionality of the interaction is not known or not applicable (e.g. binding).
6	score	Combined score of all interactions

Example entry

```
9606.ENSP00000000233    9606.ENSP00000332454    expression    inhibition    0    276
```

Create a mapping collection based on key=id

	key	M/O	Meaning	Syntax	example
1	id	M	counter	number	34
2	item_id_a	M	subject	String: tax_id.protein_id	9606.ENSP00000000233
3	item_id_b	M	object	String: tax_id.protein_id	9606.ENSP00000332454
4	mode	M	predicate	string	expression

5	action	O	predicate	string	inhibition
6	a_is_acting	O	-	number	0
7	score	M	evidence	number	276
Constant	date	M	File date	dd-mm-yyyy	01-11-2016
version	version	M	From file name	Substring: protein.actions.**v10**.txt	v10

Triple Generation

There are maximum 4 types of triple generated per record. Since various databases annotate interactions on gene level, eg most pathway databases, where other describe interactions on protein level, it makes sense to create pairwise interactions on both protein and gene level.

1. Item_id_a (type=protein) - <mode> - item_id_b (type=protein)
2. Item_id_a (type=protein) - <action> - item_id_b (type=protein)
3. Item_id_a (type=gene) - <mode> - item_id_b (type=gene)
4. Item_id_a (type=gene) - <action> - item_id_b (type=gene)

In order to determine the predicate, the following tables describe the mappings for Mode and for Action

Mapping key	string	predicate
mode	"reaction"	interacts with
mode	"expression"	controls expression of
mode	"activation"	stimulates
mode	"ptmod"	modifies
mode	"binding"	binds with
mode	"catalysis"	augments
action	“inhibition”	inhibits
action	“activation”	stimulates

Triple 1: Item_id_a (type=protein) - <mode> - item_id_b (type=protein)
For each document, create a triple:

Subject: Item_id_a

solr query: term: substring(item_id_a[protein_id]); semantictype: 116 ;

Predicate : <mode> mapping table

Object: item_id_b

solr query: term: substring(item_id_a[protein_id]); semantictype: 116 ;

Triple 2: Item_id_a (type=protein) - <action> - item_id_b (type=protein)
For each document where action is present, create a triple:

Subject: Item_id_a

solr query: term: substring(item_id_a[protein_id]); semantictype: 116 ;

Predicate : <action> mapping table

Object: item_id_b

solr query: term: substring(item_id_a[protein_id]); semantictype: 116 ;

Triple 3: Item_id_a (type=gene) - <mode> - item_id_b (type=gene)
For each document, create a triple:

Subject: Item_id_a

solr query: term: substring(item_id_a[protein_id]); semantictype: 28 ;

Predicate : <mode> mapping table

Object: item_id_b

solr query: term: substring(item_id_a[protein_id]); semantictype: 28 ;

Triple 4: Item_id_a (type=gene) - <action> - item_id_b (type=gene)

For each document where action is present, create a triple:

Subject: Item_id_a

solr query: term: substring(item_id_a[protein_id]); semantictype: 28 ;

Predicate : <action> mapping table

Object: item_id_b

solr query: term: substring(item_id_a[protein_id]); semantictype: 28 ;

Publication Generation

For all triples generated out of a mapping record (so per <id>) 1 publication is created with the following characteristics:

Measure: Publicationtype= database

Scientific value : mapping table based on score

Score (0-1000)	Scientific Value (1-7)
0 - 400	1
401 - 700	2
701 - 1000	4

Institution= STRING Consortium

Publicationtitle= STRING/<version>/item_id_a

Publication ID = STRING/item_id_a-item_id_b

Publicationdate= <date>

Publicationsource: EMBL

URL : http://string-db.org/

Source Mongo collection	**RD**	**RT**
STRING	0	112