Challenge LLL
convertor User's Manual

Julien Gobeill

5th march 2005

1.INTRODUCTION

In order to compete in the Challenge LLL, the teams have to send a predictions-file, that contains all the interactions automatically found by their system. The format of this file is very strict ; particularly, agents and targets have to be done in their canonical form, not with an ID. As we know that a lot of your systems will output interactions in terms of ID, and to avoid that each team implements a software that convert interactions in terms of ID to canonical forms, the Organization committee lets to your disposition a standard convertor. The purpose of this document is to briefly explain how using the convertor.

The convertor software for Challenge LLL is called giec-eval_convertor.pl ; to be downloaded from the web page of the Challenge (http://genome.jouy.inra.fr/texte/LLLchallenge/). The software is implemented in Perl.

2.USAGE

The software needs a Perl interpreter to run. The syntax of the command line for running it is the following one:

```
perl giec-eval_convertor.pl -r <raw-predictions-file> -d <dictionary-file>
```

The software requires two parameters:

- r <raw-predictions-file>: the path to the raw-predictions-file. The raw-predictions-file is the file containing the genic interactions as extracted by the competing system. More details about this file and its format are given in section 3.1.

- d <dictionary-file> : the path of the dictionary-file. The dictionary-file is the file containing all supported canonical forms of protein and genes names and their synonyms. It is downloadable from the web page of the Challenge. More details about this file and its format are given in section 3.2 .

An example of call to the software is:

```
perl giec-eval_convertor.pl -r exempleOfRawPredictionFile.txt -d dictionary_data.txt
```

3.INPUT FILE FORMATS

3.1 raw-predictions-file

Warning : this raw-predictions-file is different to the predictions-file that can be found in input of the format-validator, or of the evaluator. On the contrary, the goal of the convertor is to help the teams to convert the file that their system had automatically outputted, that could be in terms of IDs, into the asked format for the Challenge. Example :

A competing system automatically outputs this :

```
ID              10515909-4
sentence        Expression of the sigma(K)-dependent cwlH gene depended on gerE.
words           word(0,'Expression',0,9)    word(1,'of',11,12)          word(2,'the',14,16)         word
(3,'sigma(K)',18,25)        word(4,'dependent',27,35)               word(5,'cwlH',37,40)        word
(6,'gene',42,45)            word(7,'depended',47,54)                word(8,'on',56,57)          word
(9,'gerE',59,62)
agents          agent(3)        agent(9)
targets         target(5)
genic_interactions              genic_interaction(3,5)      genic_interaction(9,5)
```

In order to take these predictions in the Challenge format, the teams have to convert this file in this following format. The convertor does it automatically :

```
ID              10515909-4
agents          agent('sigK') agent('gerE')
targets         target('cwlH')
genic_interactions              genic_interaction('sigK','cwlH')            genic_interaction( 'gerE','cwlH')
```

The format of the raw-predictions-file has to be the same than the training datas, as defined by the web page of the Challenge. Field names and literals are separated by tabulations. People are free to insert blank lines or comments between sets, but not inside a given set. The processing of a set is triggered by the keyword ID at the beginning of a line.

Please notice that *there is no control of coherence (verifying that agents and targets are well used after being declared) in that part* ; this will be the task of the format-validator. For each sentence, even if there are no agents, no targets and no interactions to found, the field names “agents”, “targets” and “genic_interactions” must be present.

3.2 Dictionary file

The dictionary-file is the named-entity dictionary downloadable from the web page of the challenge. Participants are free to modify this file, or create their own ones. Synonyms have to be separated by tabulations.

4. OUTPUT

The format checking program stops at the first error met and warns the user with a message containing the name of the file, the number of the line, and the nature of the error. *It does not analyze the rest of the file after encountering an error.* Errors have then to be detected and corrected one by one. When the file is free of error, the program outputs – in the standard output – the file in the converted format.