Challenge LLL
check-format User's Manual

Julien Gobeill

24rd march 2005

1. INTRODUCTION

In order to evaluate systems competing in the framework of the Challenge LLL, the Organization committee will compute some scores and measures with a scoring software. Since the evaluation will take place at MIG, with the files that competing teams will have sent, those files needs to be in the right format to be computed by the scorer. The purpose of this document is to briefly explain how using the format checking software, in order to ensure that the predictions files sent to MIG are in the right format.

The format checking software for Challenge LLL is called giec-eval_format-validator.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_check-format.pl -p <predictions-file> -d <dictionary-file>

The software requires two parameters:

- p <predictions-file>: the path to the predictions-file. The 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 to the dictionary-file. The dictionary-file is the file containing all supported canonical forms of proteins 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_check-format.pl -p exempleOfPredictionsFile.txt -d exempleOfDictionary_data.txt
```

3.INPUT FILE FORMATS

3.1 predictions-file

The predictions-file is computed by a competing software. It contains a list of sets of 4 lines, each set corresponding to one sentence. Here are an example of a set :

```
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 predictions-file is more formally defined by the web page of the Challenge. Field names and literals are separated by tabulations. *Agents and targets are identified by their biological names* (between simple quotes), not by ID. As announced, *only canonical forms of agents and targets are expected*. 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 *the control of the syntax is very strict*. No agents or targets must be declared without being used in the genic_interaction line, and no agents or targets must be used in the genic_interaction line without being declared. 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 software 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 says: “Format checking succeeded without error. Thanks.”