Instructions for the constest participants

November 29, 2013

Contents

1 Introduction **1**

2 Dataset structure **1**
2.1 Folders structure 1
2.2 Input graphs 3
2.2.1 Graph file format 3
2.2.2 Example of graph file 4
2.3 Ground truth format 4

3 Executable Specifications **5**
3.1 Executable invocation 5
3.2 Input provided 6
3.3 Expected output 6

1 Introduction

The aim of this document is to explain how to use the databases that have been provided to take part to the *ICPR2014 Contest on Graph Matching Algorithms for Pattern Search in Biological Databases*. The document is composed of two main sections:

Dataset Structure In this section we will describe the folders structure common to each dataset and the file format for reading both the graphs and the ground truth.

Executable Specifications In this section we will provide the requirements for an executable that will be submitted to the contest.

2 Dataset structure

2.1 Folders structure

The whole dataset is composed of three graphs databases:

1. Molecules
2. Proteins and Backbones
3. Contact Maps

Every graph dataset has been stored inside a directory that is composed of three subdirectories, as shown in figure 2.1:

target This subdirectory includes all the target graphs

query This subdirectory includes all the query graphs

ground_truth This subdirectory includes the ground truth files

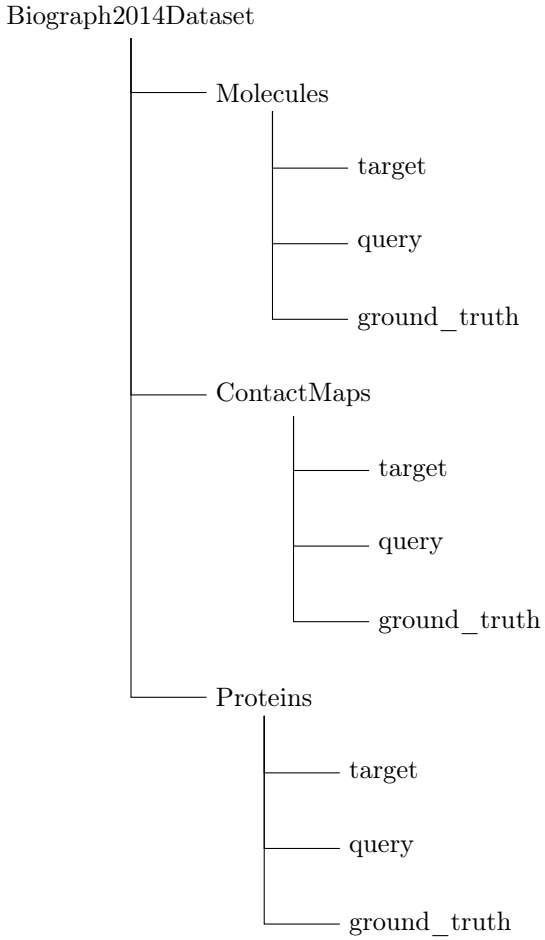


Figure 1: Structure of the directories composing the dataset

Each graph in the dataset has been stored into a file with the .grf extension, named according to the following convention:

- Target graph files have the name of the biological structure file from which have been extraced
- Query graph file names have this structure: [Target Graph Name].[Query Size].sub.grf; for instance, a query of eight nodes extracted from the target file *human_ 2JRI.grf* will be named *human_ 2JRI.8.sub.grf*.

The ground truth subdirectory contains a file for each query size; these files have the .grf extension and are named as follows: [Directory Dataset Name].[Query Size].gtr.

For instance, the ground truth file of the molecules dataset for eight nodes queries has the name: *Molecules.8.gtr*. Every file stores the expected results of the matching process for all the queries, with a given size, over the whole set of target graphs in the dataset.

2.2 Input graphs

The dataset has been built with graphs having the following features:

- Every graph is directed and weakly connected.
- The structure does not contains self loops or multiedges.
- A label has been attached to each node. The labels are short strings (max 5 characters).
- The edges do not have any kind of labels.

2.2.1 Graph file format

Graph files have the following stucture:

- The file may contain blank lines (i.e. lines containing only spaces), and comment lines starting with the "number" sign ('#'); these lines have to be ignored; the following will apply only to non-blank and non-comment lines.
- The first line contains the number of nodes in the graph.
- Afterwards, there is a line for each node, containing the node id (an integer in the range $0...N-1$, where N is the number of nodes) and the node attribute, separated by spaces.
- After these lines, for each node in the graph there will be:
  - a line with the number of edges coming out from the node
  - for each of such edges, a line containing the node ids of the edge ends, separated by spaces.

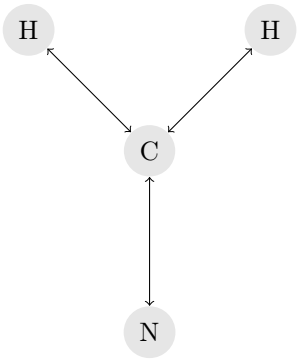
2.2.2 Example of graph file

In the following there is an example of a complete graph file, for a 4-nodes graph:

```
#Number of nodes
4

#Node Attributes
0 H
1 C
2 N
3 H

#Edges coming out from 0
1
0 1
#Edges coming out from 1
3
1 0
1 2
1 3
#Edges coming out from 2
1
2 1
#Edges coming out from 3
1
3 1
```



2.3 Ground truth format

The ground truth file contains a list of solution records separated with a blank line. A solution record has been formatted in the following manner:

```
T:[Target Graph File Name]
P:[Pattern Graph File Name]
N:[Number of Solutions]
[Solutions List]
```

Target and pattern file names do not contain the directory path. The solutions list contains one line for each expected solution. Each solution line has this format:

```
S:[Number of Pairs]:[Solution Pairs]
```

where the [Solution Pairs] is the list of matched node pairs that compose the solution. Pairs are separated by semicolons (';'); each pair contains the id of the pattern node and the id of the target node, separated by a comma (','). For instance, if a solutions is composed of eight pairs an expected solution line could be:

```
S:8:0,3;2,1;1,4;5,0;7,6;3,2;6,5;4,7
```

The following is a complete example of a solution record:

```
T:1MUC_cm_all_h.gfr
P:1STA_cm_all.8.sub.gfr
N:24
S:8:0,4;1,265;2,207;3,282;4,258;5,234;6,200;7,351
S:8:0,14;1,278;2,34;3,18;4,24;5,343;6,303;7,305
S:8:0,114;1,291;2,34;3,18;4,206;5,343;6,150;7,305
S:8:0,114;1,291;2,34;3,18;4,277;5,343;6,150;7,305
S:8:0,114;1,291;2,34;3,18;4,277;5,343;6,195;7,305
S:8:0,190;1,327;2,207;3,90;4,44;5,32;6,343;7,234
S:8:0,190;1,327;2,207;3,282;4,44;5,32;6,343;7,351
S:8:0,190;1,327;2,207;3,90;4,292;5,32;6,343;7,234
S:8:0,190;1,327;2,207;3,282;4,292;5,32;6,343;7,351
S:8:0,196;1,327;2,207;3,90;4,44;5,351;6,343;7,234
S:8:0,196;1,327;2,207;3,90;4,292;5,351;6,343;7,234
S:8:0,217;1,278;2,34;3,18;4,24;5,343;6,303;7,305
S:8:0,221;1,265;2,207;3,282;4,258;5,234;6,200;7,351
S:8:0,221;1,353;2,207;3,282;4,262;5,234;6,29;7,351
S:8:0,225;1,68;2,32;3,71;4,152;5,34;6,303;7,207
S:8:0,225;1,68;2,32;3,299;4,152;5,34;6,303;7,207
S:8:0,225;1,68;2,305;3,95;4,152;5,34;6,303;7,283
S:8:0,276;1,68;2,32;3,71;4,44;5,305;6,343;7,207
S:8:0,276;1,68;2,32;3,299;4,44;5,305;6,343;7,207
S:8:0,276;1,68;2,32;3,71;4,152;5,305;6,303;7,207
S:8:0,276;1,68;2,32;3,299;4,152;5,305;6,303;7,207
S:8:0,276;1,68;2,32;3,71;4,292;5,305;6,343;7,207
S:8:0,276;1,68;2,32;3,299;4,292;5,305;6,343;7,207
S:8:0,298;1,265;2,207;3,282;4,258;5,234;6,200;7,351
```

3 Executable Specifications

3.1 Executable invocation

Contest participants must provide their executable as a command line application, taking its parameters from the command line arguments and writing its results on the standard output. Please note that the executable will be run by a script; hence conformance with this specification is fundamental. The executable can be submitted for the following platforms:

- Windows XP or later
- Linux on X86 or X86-64

- Mac OS X 10.3 or later

Please note that, besides the executable, the participants must also provide any other non-standard libraries (i.e. DLLs) or runtime support files needed to run the program

3.2 Input provided

The command line parameters passed to the executable will be structured as follow:

```
[Run Mode] [Pattern Graph Path] [Target Graph Path]
```

The executable has to provide two run modes:

First Solution If the program is invoked in run mode "-f" it must provide only the first solution.

All Solutions If the program is invoked in run mode "-a" it must provide all the solutions.

For instance, in order to run the matching process to obtain all the solutions between the query graph 1STA_cm_all.8.sub.gfr and the target graph 1MUC_cm_all_h.gfr, which are contained in the contact maps dataset, then the command line will be:

```
./myExec -a ContactMaps/query/1STA_cm_all.8.sub.grf ContactMaps/
    target/1MUC_cm_all_h.grf > myMatchingResult.mch
```

3.3 Expected output

The algorithm has to provide an output coherent with the structure of the ground truth that has been described beforehand. The expected output depends on the selected run mode flag.

- If the executable will be run using the *all solutions flag* -a, then the output has to be structured as follows:

  ```
  T:[Target Graph File Name]
  P:[Pattern Graph File Name]
  N:[Number of Solutions]
  [Solutions List]
  ```

  For an example, please see the section 2.3.

- If the executable will be run using the *first solution flag* -f, then the output has to be structured as below:

  ```
  T:[Target Graph File Name]
  P:[Pattern Graph File Name]
  N:1
  S:[Number of Pairs]:[Solution Pairs]
  ```

An instance of the first solution output is the following:

```
T:1MUC_cm_all_h.gfr
P:1STA_cm_all.8.sub.gfr
N:1
S:8:0,4;1,265;2,207;3,282;4,258;5,234;6,200;7,351
```