

The parser has been written in Perl scripting language, and has been tested in Linux OS. It should work in any Operating System that support Perl scripting.

Requirements:

- Perl scripting language installed on the PC
- The Perl's module LWP::Simple installed (it can be installed manually or using CPAN)
- To be ran in a PC with direct internet access to the site www.unicarbkb.org

The Perl scripting language is installed by default on Linux and MacOS, while in Windows it should be manually installed (i.e. from <http://www.perl.org>).

Once Perl is installed and ready to use, the module LWP::Simple should be manually installed using the Perl CPAN management tool.

How to run the scripts:

OPTION 1: Obtain latest data from the GlycoSuite DB

- 1) Make each query manually on the <http://www.unicarbkb.org/builder> website, and save each resultant page as query1.html, query2.html
- 2) Delete the content of "proteins" and "structures" folders in order to clean the cached data (see OPTION 2)
- 3) Run the scripts (output will be the information related to progress and errors on the database such as missing information, among others) as follows:

```
./get_structures.pl query1.html "Man(a1-3)Man"
```

```
./get_structures.pl query2.html "Man(a1-2)Man"
```

```
./get_structures.pl query3.html "Man(a1-2)Man(a1-2)Man"
```

```
./get_structures.pl query4.html "Man(a1-2)Man(a1-3)[Man(a1-2)Man(a1-6)]Man"
```

```
./get_structures.pl query5.html "Man(a1-3)[Man(a1-6)]Man"
```

```
./get_structures.pl query6.html "Man(a1-2)Man(a1-6)[Man(a1-3)]Man"
```

```
./get_structures.pl query7.html "NeuAc(a2-3)Gal(b1-3)[NeuAc(a2-6)]GalNAc"
```

```
./get_structures.pl query8.html "Gal(b1-4)GlcNAc(b1-2)Man(a1-3)[Man(a1-6)]Man"
```

```
./parse_structures.pl 1
```

(output will be the relations between glycans and glycoproteins, along with the corresponding sources on the website for further verification if necessary)

OPTION 2: Process data from cache to reproduce the paper's results

It is assumed that both folders "structures" and "proteins" exists, and they contain data from a previous processing (i.e. from Option 1)

1) Run the scripts as follows:

```
./parse_structures.pl
```

(output will be the relations between glycans and glycoproteins, along with the corresponding sources on the website for further verification if necessary)