flowCL: Semantic labelling of flow cytometric cell populations

Justin Meskas

May 16, 2018

jmeskas@bccrc.ca

Contents

1 Licensing

Under the Artistic License, you are free to use and redistribute this software.

2 Loading the Library

To install ***flowCL*** type *source("http://bioconductor.org/biocLite.R")* into R and then type *biocLite("flowCL")*. For more information on installation guidelines, see the Bioconductor and the CRAN websites. Once installed, to load the library, type the following into R:

```
> library("flowCL")
```

3 Running *flowCL*

3.1 Introduction

Given a cell type, one can use a cell ontology to discover which markers are used to uniquely identify that particular cell type. In the reverse situation, how can one find an appropriate cell type with a given phenotype (or list of markers)? ***flowCL*** matches a phenotype to a cell type from the cell ontology. If the match is not unique, then the best alternatives are returned. Markers are input followed with a +, **-**, **hi** or **lo**. The + stands for *has plasma membrane part*, the **-** stands for *lacks plasma membrane part*, the **hi** stands for *has high plasma membrane amount*, and the **lo** stands for *has low plasma membrane amount*. For example, CD8+ will search for cell types that *has plasma membrane part* of CD8. Synonyms of tags are also accepted: **low**, **dim** and **- -** for **lo**, and **high**, **bri**, **bright** and ++ as synonyms for **hi**.

flowCL executes queries against the Cell Ontology (CL), available at http://cellontology.org. The CL file is hosted on a triplestore, i.e., a database for storage and retrieval of Resource Description Framework (RDF) triples. The SPARQL endpoint at http://cell.ctde.net:8080/openrdf-sesame/repositories/CL is used to execute the SPARQL queries retrieving the correct matches from the CL. While other SPARQL endpoints can be used, users should be aware that in our case the CL file has been reasoned upon, and resulting extra inferred axioms have been added to the triplestore, providing a more complete result set.

3.2 Archive

A folder called "flowCL_results" will be created in the current directory. In this folder a file called "listPhenotypes.csv" is created. This file lists the results of ***flowCL***, and is the same as what is returned in the $Table section of the returned data. In addition, tree diagrams are created and saved as ".pdf" files and these show the cell dependency of the matched cell types. The code will slowly build an archive of information in "[current directory]/flowCL_results/" inside the folders of "parents", "parents_query", "results" and "reverse_query". Accessing this archive is much faster than querying every time, however, if the ontology gets updated the code will use the now old data from these folders. To make sure the code is as up to date as possible the archive should be reset every once in a while.

For the sake of time, a pre-loaded archive can be created from a data file in ***flowCL***. To load this archive, enter the following:

```
> flowCL("Archive")
```

This pre-loaded archive has the possibility of being outdated. This archive will be updated by the maintainer, along with the vignette, whenever the ontology is updated. More discussion of this is in the "Last Comments" section. The current version was updated last on January 4^{th} 2018. To check that the version is up to date run:

```
> flowCL("Date")
```

If the returned date matches or is prior to the one above, then the pre-loaded archive is the most up to date. Please note that the user can skip this section altogether and ***flowCL*** will always give the most current archive. The down side is the code will take time every time there is something new to query while building an archive.

3.3 Simple Example

The simplest example of ***flowCL*** is to enter in one phenotype:

```
> Res <- flowCL("CCR7+CD45RA+")
```

```
ExpMrkrLst was not defined. Defining ExpMrkrLst as MarkerList.
```

```
> Res$Table
```

```
                      [,1]
Short marker names    "CCR7+CD45RA+"
Ontology marker names "C-C chemokine receptor type 7, receptor-type tyrosine-protein"
                      "phosphatase C isoform CD45RA"
Experiment markers    "CCR7,CD45RA"
Ontology exper. names "C-C chemokine receptor type 7, receptor-type tyrosine-protein"
                      "phosphatase C isoform CD45RA"
Successful Match?     "No"
Marker ID             "1) PR_000001203, PR_000001015 2) PR_000001203, PR_000001015 3)"
                      "PR_000001203, PR_000001015 4) PR_000001203, PR_000001015 5)"
                      "PR_000001203, PR_000001015"
Marker Label          "1) C-C chemokine receptor type 7, receptor-type tyrosine-protein"
                      "phosphatase C isoform CD45RA 2) C-C chemokine receptor type 7,"
                      "receptor-type tyrosine-protein phosphatase C isoform CD45RA 3) C-C"
                      "chemokine receptor type 7, receptor-type tyrosine-protein phosphatase"
                      "C isoform CD45RA 4) C-C chemokine receptor type 7, receptor-type"
                      "tyrosine-protein phosphatase C isoform CD45RA 5) C-C chemokine"
                      "receptor type 7, receptor-type tyrosine-protein phosphatase C isoform"
                      "CD45RA"
Marker Key            "1) { } CD45RA+, CCR7+ ( ) [ interleukin-7 receptor subunit alpha+,"
                      "CD3+, CD44 moleculelo, L-selectinhi ] 2) { } CD45RA+, CCR7+ ( ) [ T"
                      "cell receptor co-receptor CD8-, interleukin-2 receptor subunit"
                      "alpha-, CD4 molecule+, alpha-beta T cell receptor complex+,"
                      "interleukin-7 receptor subunit alpha+, CD3+, CD44 moleculelo,"
                      "L-selectinhi ] 3) { } CD45RA+, CCR7+ ( ) [ CD4 molecule-,"
                      "interleukin-2 receptor subunit alpha-, T cell receptor co-receptor"
                      "CD8+, alpha-beta T cell receptor complex+, interleukin-7 receptor"
                      "subunit alpha+, CD3+, CD44 moleculelo, L-selectinhi ] 4) { } CD45RA+,"
```

```
                    "CCR7+ ( ) [ T cell receptor co-receptor CD8-, CD4 molecule+,"
                    "alpha-beta T cell receptor complex+, interleukin-2 receptor subunit"
                    "alpha+, interleukin-7 receptor subunit alphalo, CD3+, CD44"
                    "moleculelo, L-selectinhi ] 5) { } CD45RA+, CCR7+ ( ) [ T cell"
                    "receptor co-receptor CD8-, receptor-type tyrosine-protein phosphatase"
                    "C isoform CD45RO-, CD4 molecule+, alpha-beta T cell receptor"
                    "complex+, interleukin-2 receptor subunit alpha+, interleukin-7"
                    "receptor subunit alphalo, C-C chemokine receptor type 4+, CD3+, CD44"
                    "moleculelo, L-selectinhi ]"
Score (Out of 1)    "1) 0.333 2) 0.2 3) 0.2 4) 0.2 5) 0.167"
Cell ID             "1) CL_0000898 2) CL_0000895 3) CL_0000900 4) CL_0002677 5) CL_0001045"
Cell Label          "1) naive T cell 2) naive thymus-derived CD4-positive, alpha-beta T"
                    "cell 3) naive thymus-derived CD8-positive, alpha-beta T cell 4) naive"
                    "regulatory T cell 5) naive CCR4-positive regulatory T cell"
```

```
> tmp <- Res$'CCR7+CD45RA+'
> plot(tmp[[1]], nodeAttrs=tmp[[2]], edgeAttrs=tmp[[3]], attrs=tmp[[4]])
```

A list containing N + 5 elements was returned into Res, where N is the number of phenotypes queried, in this case N = 1. Each of these N elements contains information for plotting the results. The other five elements show the cell labels ($Cell_Labels), the matching markers in a list form ($Marker_Groups) and a bracket form ($Markers), ranking scores ($Ranking) and a table ($Table). The cell labels element lists the cell labels in order of highest score based on their ranking, which is in a form easily extracted and used by other R packages and functions. $Marker_Groups and $Markers list markers that were queried and markers that are used to define the certain cell types. In $Markers these markers are displayed in the form of { **A** } **B** (**C**) [**D**]. **A** and **B** together make up the markers input for the query. **B**, **C** and **D** together are the markers that make up the definition of the particular cell type. **C** lists the markers that were part of the experiment that were not part of the query, while **D** lists all other markers that make up the cell type that were not part of the experimental markers. **A** lists all the markers in the input for the query that were not required for that particular cell type.

The table from *Res*$*Table* shows many properties of the phenotype queried. The name of the markers in the ontology are shown, in this case "C-C chemokine receptor type 7" and "receptor-type tyrosine-protein phosphatase C isoform CD45RA" for CCR7 and CD45RA, respectively. The experiment markers are shown as well. These are a list of all the markers used in the experiment. ***flowCL*** will inform the user of which additional markers, which are in the experiment markers and are not in the input markers, that would better define a particular cell type. Hence, the user can refine her/his cell population to achieve a more accurate population. A successful match is also specified given that there are cell types that contain all the markers queried and only those markers. The 1) - 5) in the next six sections of the table represent the five cell types that are considered matches.

In Figure 1, a tree diagram shows the cell hierarchy that is dependent on both CCR7+ and CD45RA+. There are five cell types that contain both markers. The black arrows are defined as "is a" (ex. native cell "is a" cell). The coloured arrows are the inverse of *has/lacks plasma membrane part* or *has high/low plasma membrane amount* (ex. native regulatory T cell *has plasma membrane part* CD45RA). Each marker is associated with its own colour, which makes it easier for the user to tell which cell type contains which markers. The blue nodes are the *has plasma membrane part* markers, while the red nodes are the *lacks plasma membrane part* markers. The navy blue nodes are the *has high plasma membrane amount* markers and the grey blue nodes are the *has low plasma*

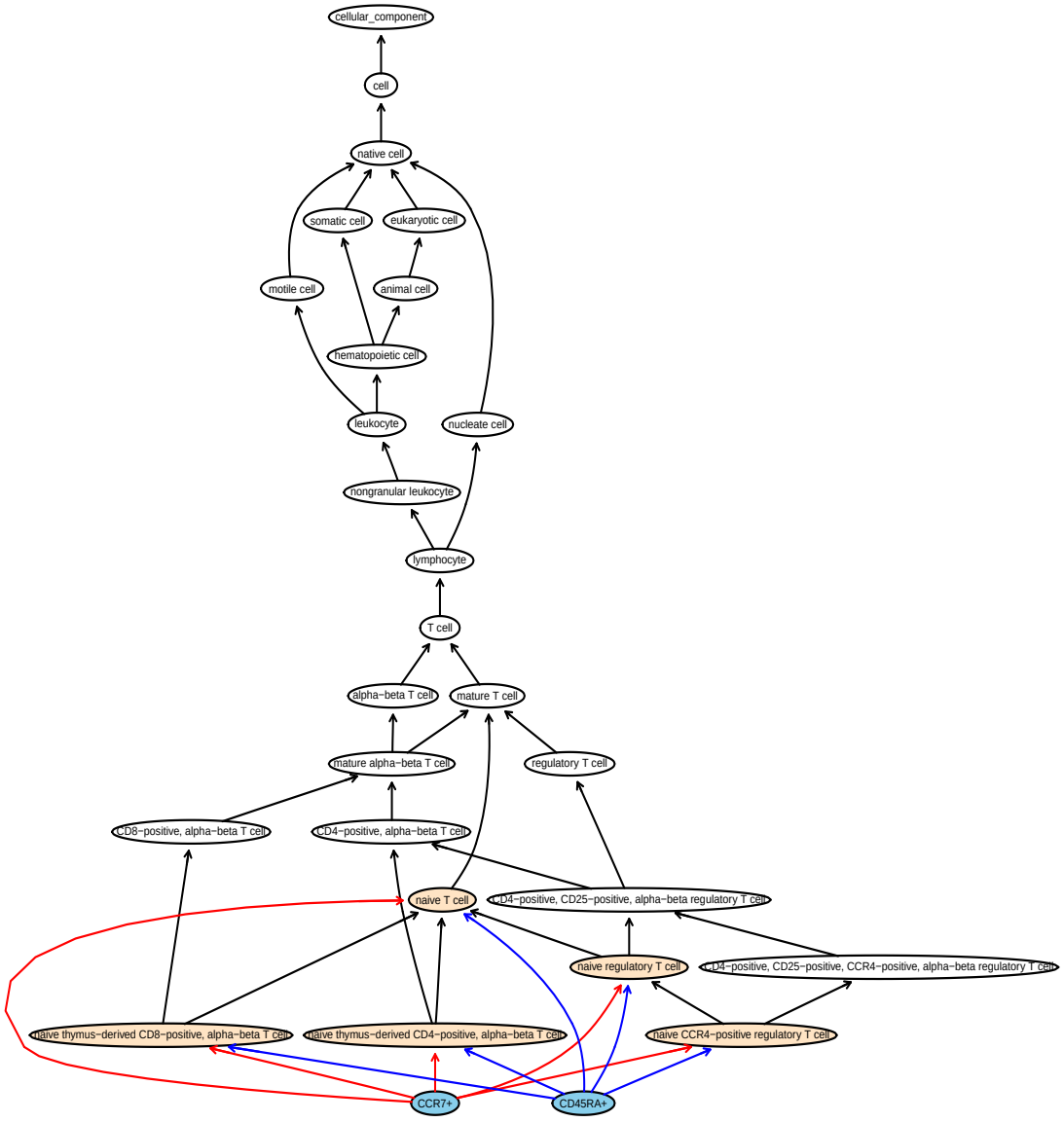

Figure 1: Tree diagram of the cell hierarchy when querying with phenotype: “CCR7+CD45RA+”

membrane amount markers. The green nodes are the exact matches, while the beige nodes are the partial matches.

3.4 Arguments Discussion

- The main argument, "MarkerList", is a single conjoined marker or a list of conjoined markers that make a phenotype or phenotypes. ***flowCL*** will attempt to find the best cell type that contains the highest possible amount of the input markers without having too many extra markers.
- If the "ExpMrkrLst" argument is blank then it is set to equal "MarkerList". The main use of "ExpMrkrLst" is to list all the markers in your experiment, and if there is one that is not in "MarkerList" and it is part of the cell type, then ***flowCL*** will inform the user.
- The "Indices" argument allows the user to not have to change the list in "MarkerList" if only a subset of the list is wanted.
- The "Verbose" option allows the user to view the computational updates while the code is in progress. The default for "Verbose" is FALSE.
- If the argument called "ResetArch" is set to TRUE, it will first delete the entire archive and start adding data to a new one every time something is queried. Therefore, the code is slower after the archive is reset, however, the code will become faster on average with every query.
- The argument called "KeepArch" allows the user to remove the archive at the end of the simulation if it is set to FALSE. This is useful when the user rather not have files stored on the hard drive.
- The "MaxHitsPht" argument controls how many cell types are returned if there are maybe closely ranked best matched. The default is set to 5.
- The "OntolNamesTD" option allows for the marker nodes to display their ontology names in the tree diagrams instead of their short names (ex. "T cell receptor co-receptor CD8" is displayed instead of "CD8"). The default for "OntolNamesTD" is FALSE.
- The "VisualSkip" argument in ***flowCL*** can be used if the user does not want the visual results. This will reduce the computational time.

3.5 Scoring

The ranking score is calculated by adding all the markers that were queried that are also part of the cell type, possibly subtracting a penalty, and dividing the number of markers that define that cell type. The penalty of -2 is given for every marker that was queried that does not define that cell type. In equation form (amount of each): (**B** - 2*__A__ - 2*__Z__) / (**B** + **C** + **D**), where **Z** is the number of markers that are queried with the wrong tag. Currently if Z is not 0 then the cell type is removed. However, if that is the only cell type left, it will not be removed and the **Z** penalty will apply both in **A** and **Z**. The ranking score of 1, the maximum, can only be obtained if every marker in the query was used in the definition of the cell type, i.e. an exact match (only **B** markers).

3.6 Cell Label Example

In many cases ***flowCL*** will only be used to extract the best possible cell label given a set of markers. The following code shows an example of this:

```
> x <-"CD3-CD19-CD20-CD14+"
> Res <- flowCL(x)

ExpMrkrLst was not defined. Defining ExpMrkrLst as MarkerList.

> Res$Cell_Label[[x]][[1]]

[1] "CD14-positive monocyte"
```

The user should note that the returned value could be one of many possible best cell labels. To access the other possible best cell labels the "1" in $Res\$Cell\_Label[[x]][[1]]$ can be changed to a higher index.

3.7 Last Comments

- If the user wants all possible HIPC results and all HIPC tree diagrams, then the command of "flowCL("")" or "flowCL("HIPC")" will achieve this. All the defaults are set to run the HIPC phenotypes and their individual markers. Running the full code with no archive can take up to 20 minutes.
- There are two packages that ***flowCL*** depends upon. The first is ***SPARQL***, which is used when retrieving cell ontology data. The other is ***Rgraphviz***, which is used when creating the tree diagrams.
- The user should note that since the ontology can be and will be updated, the R results shown in this vignette, both the figure and the static text, may not match a live run of ***flowCL***. Also the data files containing a pre-loaded archive can be outdated. The maintainer will try to keep the vignette and data files as up to date as possible. The data files as well as text and the figures in the examples from this vignette describe results from the cell ontology that was last updated on December 11^{th} 2017.