Prediction of RNA and DNA binding sites: weekly report

Pandu Raharja [1,2] Julian Schmidt [1,2]

[1]Technische Universität München

[2]Ludwig-Maximilians-Universität München

December 8, 2016

Outline

- Who's presenting what:
  - Julian: XNA binding prediction
  - Pandu: XNA binding site prediction
- Problems.

As always:

- Scripts and results could be found in: https://github.com/raharjaliu/BIFers

Part I: XNA binding prediction

Check your ids...

```
sp|Q86XP6|GKN2_HUMAN Gastrokine-2 OS=Homo sapiens GN=GKN2 PE=1 S
!=
>sp|Q86XP6|GKN2_HUMAN
```

```
Use of uninitialized value in concatenation (.) or string at
/usr/bin/profkernel-workflow line 1090 (#1)
    (W uninitialized) An undefined value was used as if it were already
    defined.  It was interpreted as a "" or a 0, but maybe it was a mistake.
    To suppress this warning assign a defined value to your variables.

    To help you figure out what was undefined, perl will try to tell you
    the name of the variable (if any) that was undefined.  In some cases
    it cannot do this, so it also tells you what operation you used the
    undefined value in.  Note, however, that perl optimizes your program
    anid the operation displayed in the warning may not necessarily appear
    literally in your program.  For example, "that $foo" is usually
    optimized into "that " . $foo, and the warning will refer to the
    concatenation (.) operator, even though there is no . in
    your program.
```

Can't build model for split.0

```
...
Warning: sequence Q8NFD4 has an empty profile
Warning: sequence Q3SY05 has an empty profile
Warning: sequence Q3Y452 has an empty profile
Warning: sequence Q495D7 has an empty profile
Warning: sequence O60756 has an empty profile
Warning: diag 0 at 604
Warning: diag 0 at 605
Normalizing Matrix
Illegal division by zero at /usr/bin/profkernel-workflow line 250, <GRAM> line
1 (#1)
    (F) You tried to divide a number by 0.  Either something was wrong in
    your logic, or you need to put a conditional in to guard against
    meaningless input.

Uncaught exception from user code:
Illegal division by zero at /usr/bin/profkernel-workflow line 250, <GRAM> line 1.
profkernel-workflow -f /mnt/project/pp2_1617/xna_raharjaschmidt/split.1.dna.fasta -p /mnt/project/pp2_1617
```

Current state

1. Train profkernel: finished (for split 1 and 2)
2. Predict binding: finished (for existing models)
3. Evaluate predictions for different k and s: Work in progress

Part II: XNA binding site prediction

Intro

- We're a bit late onto the game due to illness and conference.
- Extracted features could be found in `/mnt/project/pp2_1516/xrna_raharjaschmidt/machine_learning/{dna_big.arff,rna_big.arff}`.

Extraction of Features on ppDNA2

- `query.disis not found.`
- `Q84ZU4:  no significant pfam hit, using default values...`
- `P03206` **(same warning)**

Extraction of Features on ppRNA2

- `query.disis not found.`
- `Processing P17574... ParseError: "It seems that we have a situation now. The expected amount of columns is 22, found: %s!" % len(tokens)`
- `P24264` **(same ParseError)**
- `P67876: no significant pfam hit, using default values...`
- `P0C206` **(same warning)**
- `P0C8P8` **(same warning)**
- `P07243` **(same warning)**
- `Q57817` **(same warning)**
- `P04891` **(same warning)**
- `P18683` **(same warning)**

Machine Learning Sorcery

Some considerations:

- Stack choice: not a big fan of Java but features are contained in ARFF already:
  $\rightarrow$ Weka binding on Python?
- Easy model (SVN, DT/RF etc) vs. sophisticated model (deep representation learning and all the new fancy things coming out of NIPS/ICML 201X):
  $\rightarrow$ Implementation **will** be constrained by Weka (and time).
- Single model vs. ensemble learning (with boosting etc).
- Time constraint: exams and works.

Some notes regarding Weka on Py

- Seems to be possible:
  http://www.cs.waikato.ac.nz/~eibe/WEKA_Ecosystem.pdf
- Requires python-weka-wrapper and cos's:
  `$sudo apt-get install python-pip python-numpy python-dev python-imaging python-matplotlib python-pygraphviz imagemagick`
  `$sudo pip install javabridge python-weka-wrapper`
  $\rightarrow$ are all these installed on the server?
- Actually requires **starting-up of JVM on Python**.

Example (1)

- Starting the JVM from Python:

```
import weka.core.jvm as jvm
jvm.start()
```

- Getting help:

```
help(jvm.start)
```

- Loading and printing some data in ARFF format:

```
from weka.core.converters import Loader
l = Loader("weka.core.converters.ArffLoader")
d = l.load_file("weka-3-7-11/data/iris.arff")
d.set_class_index(d.num_attributes() - 1)
print(d)
```

Example (2)

- Building and printing a decision tree:

```
from weka.classifiers import Classifier
c = Classifier("weka.classifiers.trees.J48")
c.build_classifier(d)
print(c)
```

- Evaluating classifier using cross-validation:

```
from weka.classifiers import Evaluation
from weka.core.classes import Random
e = Evaluation(d)
e.crossvalidate_model(c, d, 10, Random(1))
print(e.percent_correct())
print(e.to_summary())
print(e.to_class_details())
```

Some notes regarding prediction

- What constitutes a good predictor?
  - which scoring function to optimize:
    $\rightarrow F_\beta$, recall, coverage?
- Relaxed or rigid definition of binding residue?
  −!!!− vs −!−!− vs −−!!− vs −!!!!
- Binary vs continuous (very wrong, wrong, kinda correct, correct) classification value.
- etc (what is the meaning of life?)

References