GCB 2016 notification for paper 29

1 wiadomość

GCB 2016 <gcb2016@easychair.org> 6 lipca 2016 13:36
Do: Michal Burdukiewicz <michalburdukiewicz@gmail.com>

Dear Michal Burdukiewicz Burdukiewicz

thank you submitting your work

Prediction of amyloidogenicity based on the n-gram analysis

to GCB 2016.

We are pleased to let you know that your paper has been accepted for presentation at GCB 2016 in Berlin. Attached you will find the reviewers comments. Please address them and submit the final version before August 7th.

Yours,

Bernhard Renard, Knut Reinert, Joachim Selbig

---------- REVIEW 1 ----------
PAPER: 29
TITLE: Prediction of amyloidogenicity based on the n-gram analysis
AUTHORS: Michal Burdukiewicz, Piotr Sobczyk, Stefan Rödiger, Anna Duda-Madej, Paweł Mackiewicz and Malgorzata Kotulska

OVERALL EVALUATION: 2 (accept)
REVIEWER'S CONFIDENCE: 4 (high)

---------- Review ----------
The authors use a machine learning method to detect amyloidogenic peptides. Specifically, they select n-grams of encoded peptides (by physicochemical properties) and subsequently train a predictor using random forests. Prediction accuracy is then assessed and compared with other methods.

General:
The work fits well into the scope of the GCB'16 and should be accepted for the proceedings once the following revisions have been made.
1) There are many grammatical mistakes and essential descriptions of variables and formulas are missing. Normally this would be a minor issue, but the list is quite long (see at the end of the document).
2) The threshold for the classification as amyloidogenic is currently 0.5 as far as I understood. First, this is nowhere mentioned. Secondly, this may lead to many false positives/false negatives. You may think to use more conservative thresholds, i.e. akin to common statistics, a peptide would be predicted to be amylogenic (with >95% confidence) if the probability of not being amyloidogenic is < 0.05. Vice versa, a peptide may not be amyloidogenic with 95% confidence if the probability of being amyloidogenic is < 0.05. This leaves you with a range, where you cannot make conclusive statements, but this is ok. Please comment, or change. Also, multiple test correction (e.g. Benjamini-Hochberg Method) could be useful in the aforementioned context.
3) Quick permutation test (page 4): Please explain which hypothesis the p-value actually assesses/how it is computed.
4) The authors state that amylogram creates a highly interpretable outcome, however they fail to interpret it sufficiently. This part should be extended. The introduction and sections 3.1-3.3 can be shortened to accommodate for this additional discussion.

Minor:

a) The tool doesn't work with data provided on github (e.g. /data/amyloid_neg_full.fasta). error: "n-gram too long". I guess the removal of the short sequences is not implemented.
b) Page 5 under the contingency table: Shouldn't it be the product of two binomials, if x & y are independent, not multinomial?
c) The classification and the training data is currently binary (amyloidogenic vs. non amyloidogenic). Which always creates a bias with regard to the actual processes being modelled. May it be useful, if not more accurate, to use a

continuous measure instead? E.g. rate/propensity of aggregation initiated by the amyloidogenic peptide? Is such data available (see also comment 2))? Please comment/discuss.

Grammatical errors, missing annotation:

- Page 5: ‘p’ and ‘n’ need to be explained
- Page 5: “…QuiPT is heuristics…” -> “…QuiPT is a heuristic…”
- Page 5: During the learning stage, random forest… -> During the learning stage, a random forest…
- Page 5/6: section 2.5-2.7: There are lots of grammatical mistakes (in particular missing articles and prepositions) in this section. Please revise the entire sections.
- The AUC computation should be explained (formula?).
- Page 8: “able to outperformed the other published methods” -> “able to outperform published methods”
- Page 10: “did not tested” -> “did not test”
- Page 10: “Thanks to the reduction …” is not the way to start a sentence in a scientific paper.
- Page 10 “assumed approach” -> “pursued approach”, or simply “this approach”
- The citation on page 3, first paragraph: Wozniak and Kotulska (2014) seems to have a different format than all other citations.
- The sentence (middle of page 2) “The aim of our study is opposite to choose out of thousands of created hot spot models the most appropriate one and from its analysis gain a new insight into the mechanism of amyloidogenicity.” should be re-phrased for clarity.
- Consider re-phrasing the half-sentence “…thanks to the technological advancements.” (page 3)
- The sentence (page 3) “Based on that, we created 524,284 encodings with different levels of amino acid alphabet reduction, from three to six groups of amino acid using Ward’s clusterization (Joe H. Ward Jr, 1963), which was performed on all combinations of the normalized values of 17 selected physicochemical properties.” Is a bit too long and hard to understand.
- Page 6: “It results most probably from the pattern homogeneity of the short peptides.” And the paragraph around needs to be revised. It is hard to follow your line of thoughts here… Maybe shorten this section and focus only on the message you want to transmit.
- Table 1: The abbreviations used in the classifiers developed by this group (row 5 & 6 in Table 1) need to be explained

---------- REVIEW 2 ----------
PAPER: 29
TITLE: Prediction of amyloidogenicity based on the n-gram analysis
AUTHORS: Michal Burdukiewicz, Piotr Sobczyk, Stefan Rödiger, Anna Duda-Madej, Paweł Mackiewicz and Malgorzata Kotulska

OVERALL EVALUATION: 1 (weak accept)
REVIEWER'S CONFIDENCE: 3 (medium)

---------- Review ----------
The authors developed a method to train predictor in order to predict the amyloidogenicy of amino acid sequences. This predictor, called AmyloGram, is available as a free web application and the source are available on github.The method which has been used to train a random forest-based classifier is based on the reduction of complexity by using a reduced amino acid alphabet (based on physico-chemical properties) and using short k-mers extracted from overlapping hexamers as features for the classifier. After feature selection performed by an own developed approach in order to extract the most informative k-mers the random forest-based classifier was trained. Known amyloidogenic and non-amyloidogenic peptides from the AmyLoad-database published by the last author of this work were used as training data. Using an independent benchmark data set the predictor was compared with alternative tools.

This approach may be interesting not only for researches interested in sequences of amyloids but also for general protein sequence research. The results show a good performance in comparison with alternative tools. A quick test of the implemented web application (the full sequence of the Amyloid beta A4 protein was tested) was successful (the predictor voted for "yes"). The results regarding the best performing alternative amino acid encoding are interesting.

However, there are some remarks regarding this paper:

minor remarks:
1. In the last sentence of the introduction I suggest to write "a external data set" instead "the external data set" because the benchmark data was mentioned only in the abstract before. Alternatively, you should mention the pep424-data in the Introduction.

2. The numbers of sequences shown in table 2 are very confusing. In the reduced AmyLoad-data there are 1033 non-amyloidogenic and 397 amyloidogenic sequences. In table 2 the numbers for the test set sum to 841+123+28+41=1033 and 247+65+30+55=397. But it is unclear why the data for the training set do not and where are these numbers coming from. So, please explain this table in more detail in the caption or in the text. Alternatively you may design an alternative depiction.

3. In figure 2, in each box plot there are two blue squares for the two standard encodings. Please use different symbols for standard encoding 1 and standard encoding 2. This is confusing for the readers. Additionally there may be readers who want to know the performance of a particular standard encoding.

4. In table 3, it would be valuable to show also the results for AmyloGram versions trained with peptides with different lengths (i.e., AmyloGram(6), AmyloGram(6-10) and Amylogram(6-15)) and using the best encoding, in order to compare them directly to the full alphabet-results.

5. The alternative methods are described not sufficiently detailed. E.g., the reader wants to know whether the tested tools are also based on random forests or alternative approaches. Are the alternative methods using the full alphabet or an encoding (which one?). These descriptions may reveal additional aspects for the discussion of the results of the tool comparison.

6. Generally, the discussion of the results is very short. E.g., the advantage of using the encoding-based version of AmyloGram is shown in table 3 but described in the text very briefly. It would be valuable to discuss more detailed why the reduced alphabet outperforms the full alphabet.

major remarks:
none

There is general interest in this manuscript. However, the issues mentioned above
should be revised to make it partly less confusing. The decision is "weak accept".

---------- REVIEW 3 ----------
PAPER: 29
TITLE: Prediction of amyloidogenicity based on the n-gram analysis
AUTHORS: Michal Burdukiewicz, Piotr Sobczyk, Stefan Rödiger, Anna Duda-Madej, Paweł Mackiewicz and Malgorzata Kotulska

OVERALL EVALUATION: 3 (strong accept)
REVIEWER'S CONFIDENCE: 4 (high)

---------- Review ----------
The authors present a methodology of feature selection for the classification of (non-)amyloid peptides encoded by different alphabets based on physical properties. They developed a new algorithm and provide all used data. The manuscript is well written, the results are clearly described and discussed. Only two remarks:

(1) The introduction (especially the biomedical part) is missing some references and sources.
(2) From the original 1457 (418+1039) peptides, 35 (8+27) were removed because of the sequence length. So, the final data set should be consist of 1422 and not 1430 as written in the paper. Can you comment this?