GHENT UNIVERSITY

HONOURS PROGRAMME IN LIFE SCIENCES 2016-2017

FACULTY OF PHARMACEUTICAL SCIENCES | FACULTY OF MEDICINE AND HEALTH SCIENCES | FACULTY OF VETERINARY MEDICINE accredited by EAEVE

Computational protein discovery: can the elimination of ORF length in coding potential prediction improve the detection of micropeptides?

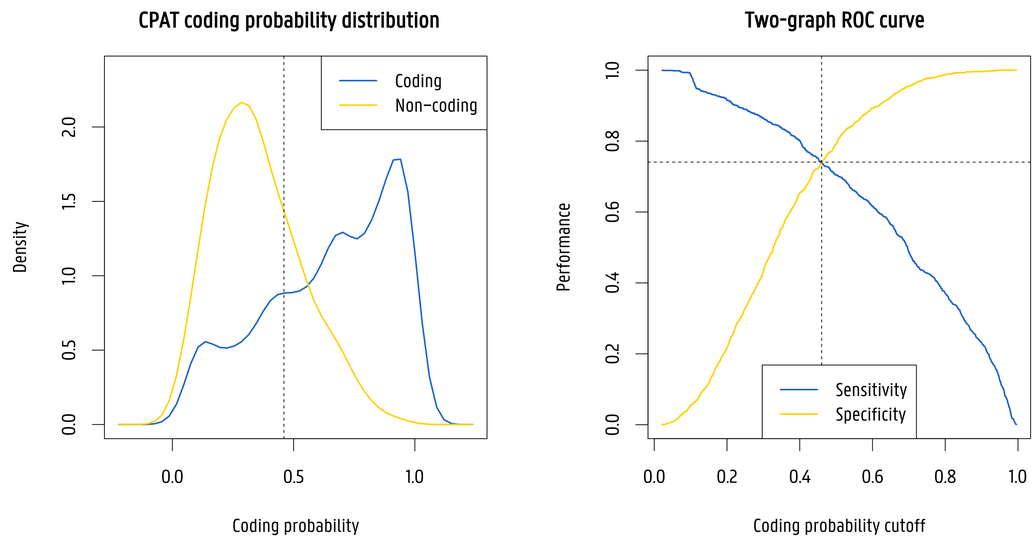
Basiel Weyers | Promotor: Dr. Pieter-Jan Volders | Copromotor: Prof. Dr. Lennart Martens | 20th of November 2017

1. ANALYSIS OF THE PROBLEM AND OBJECTIVE

- Many algorithms to distinguish between coding and non-coding RNA are available
- Training data: based on current annotations → these show a strong bias towards proteins larger than 100 amino acids (300 nucleotides) with high evolutionary conservation
- (Relative) Open Reading Frame (ORF) size: often the most powerful discriminator on typically used benchmarking sets, due to a bias in current annotation
- Our analysis shows poor performance of these algorithms on ORFs ≤ 100AA (data not shown)
- Objective: re-train one of these algorithms (CPAT or Coding-Potential Assessment Tool) to eliminate ORF size as a parameter and compare its performance to the original algorithm

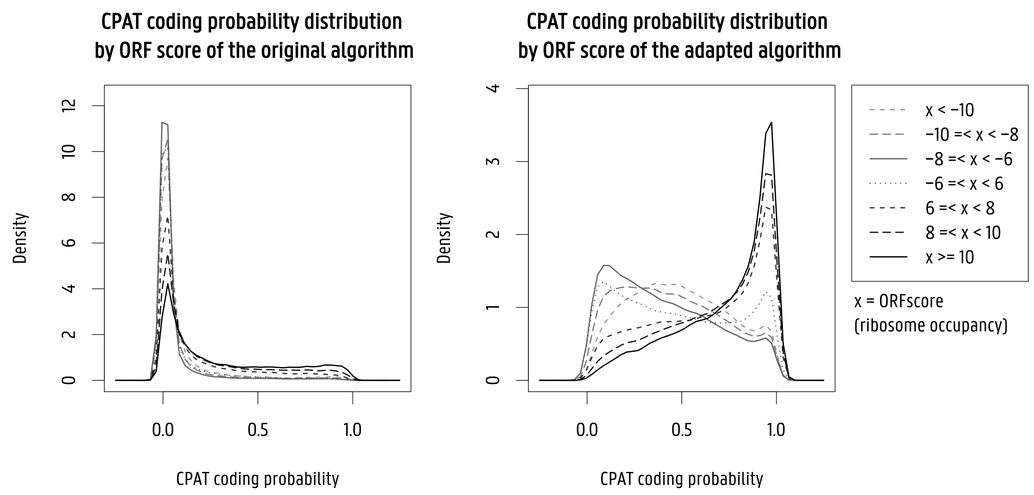
2. RE-TRAINING THE ALGORITHM: IMPROVED PERFORMANCE ON SMALL ORFS

We adapted CPAT so that it evaluates every ORF, instead of only the longest one. Afterwards, we trained a new logistic regression model for CPAT, using a dataset of 695 manually annotated proteins smaller than 100 amino acids (300 nucleotides), along with a dataset of non-coding sequences of the exact same length. Doing so, we completely eliminated ORF size as a predictor. The performance of the newly built model was tested using a 10-fold cross validation on the same dataset.



- Dotted lines: new optimal cut-off value of 0.460 at a sensitivity and specificity of 0.741
- Area Under the Curve (AUC) is now 0.806, being slightly better than the original algorithm on the same data (0.780)

Ribosome profiling is an RNA sequencing based technique to study ribosome occupancy in the transcriptome. We used a database of 756,359 small ORFs (www.sorfs.org) scored with ribosome profiling as an independent dataset to evaluate our algorithms' ability to discover new small ORFs.

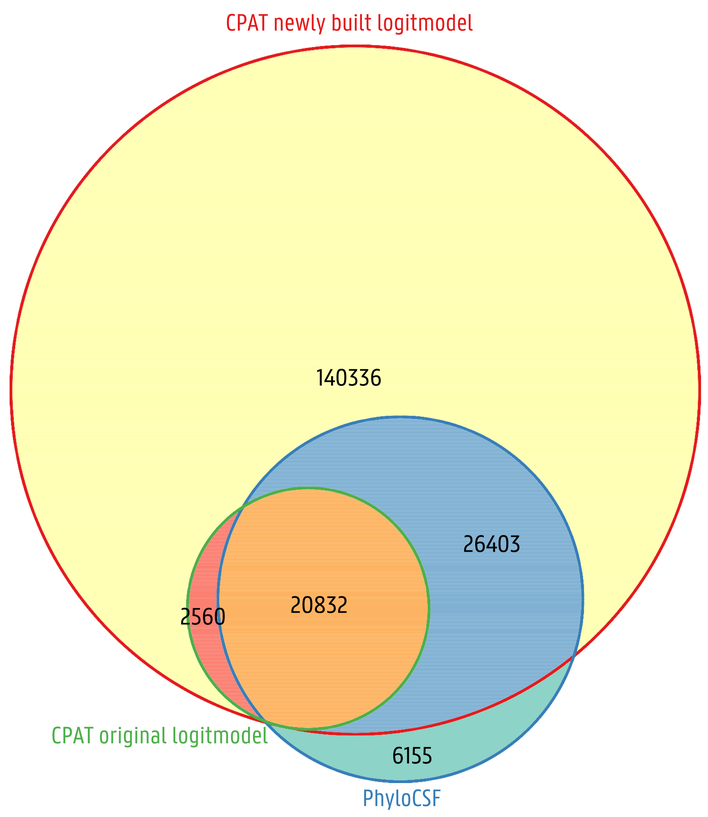


With the adapted algorithm, higher ribosome occupancy also corresponds to a higher CPAT coding probability, which is not the case for the original algorithm. Thus, our modifications resulted in an improved performance on small ORFs. Nonetheless, a decrease in specificity was observed.

3. SELECTED USE CASE

As a use case, we used the PANcancer transcriptome dataset, which is built from large-scale RNA sequencing efforts in the research group of prof. Jo Vandesompele and prof. Pieter Mestdagh at the Center for Medical Genetics Ghent. We compared the results of CPAT (both the original and adapted algorithm) with the results of PhyloCSF, another prediction program.

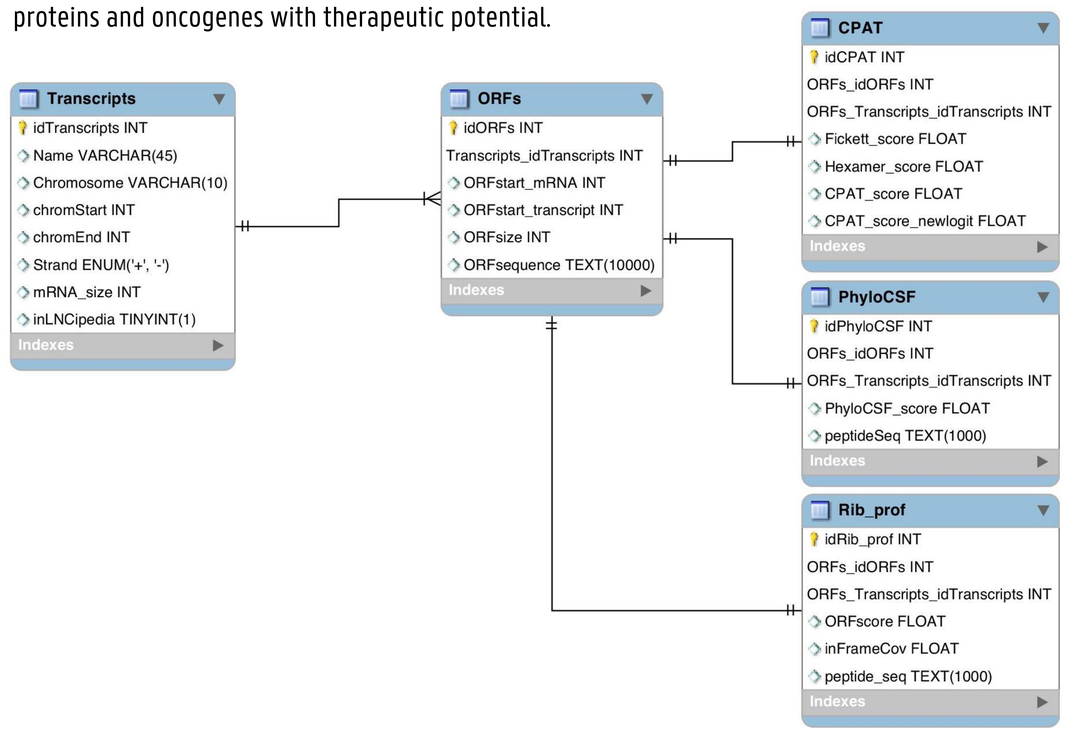
When we look at all ORFs regardless of length, the original algorithm outperforms the adapted one. However, when we only consider ORFs shorter than 300 nucleotides, the performance of the adapted algorithm is much better than that of the original one. As expected, the performance of the adapted algorithm is completely unaffected by the limit in ORF size.



When we examine the number of ORFs that are considered coding by each algorithm, it becomes clear how many more ORFs can be found when eliminating ORF length as a parameter to predict coding probability, that would have been overlooked otherwise. Given the excellent specificity and sensitivity of the model, it is unlikely that this can be attributed to false positives alone.

4. DATABASE WITH RESULTS

We created two MySQL databases to store the transcripts and the resulting scores: one for the small ORFs and one for PANcancer. These databases can be an interesting resource to discover new proteins and oncogenes with therapeutic potential.



CONCLUSION

By eliminating ORF length as a parameter to predict coding probability, we can make a more accurate analysis of the coding probability of small ORFs. Thus, ORF size is not a good parameter to predict coding probability, confirming our hypothesis. Unfortunately, we lost some of the performance on bigger ORFs. This small loss in performance should not be a reason to continue using ORF size as a parameter, especially because it is likely based on a bias in current annotations.