ampir: an R package for fast genome-wide prediction of antimicrobial peptides

Legana C H W Fingerhut ✉, David J Miller, Jan M Strugnell, Norelle L Daly, Ira R Cooke ✉

Bioinformatics, Volume 36, Issue 21, 1 November 2020, Pages 5262–5263, https://doi.org/10.1093/bioinformatics/btaa653

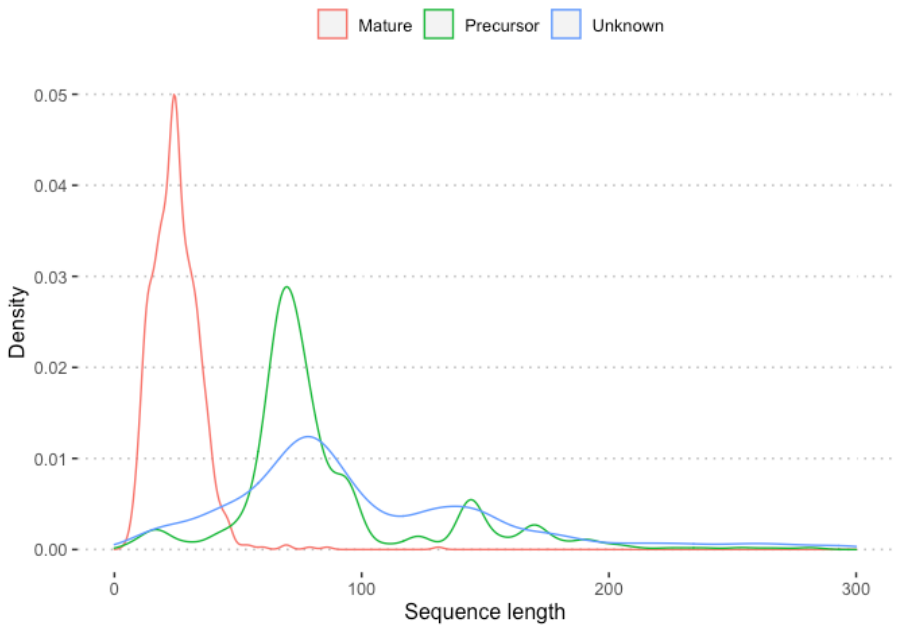
Published: 19 July 2020 **Article history** ▾

Shortcomings of other AMP predictors

- Designed to handle relatively small numbers of sequences
- Most lack a programming interface
- Trained on data not resembling whole genome data

ampir: precursor and mature models

- Mature: comparable with previous approaches
- Precursor: trained on full-length precursors

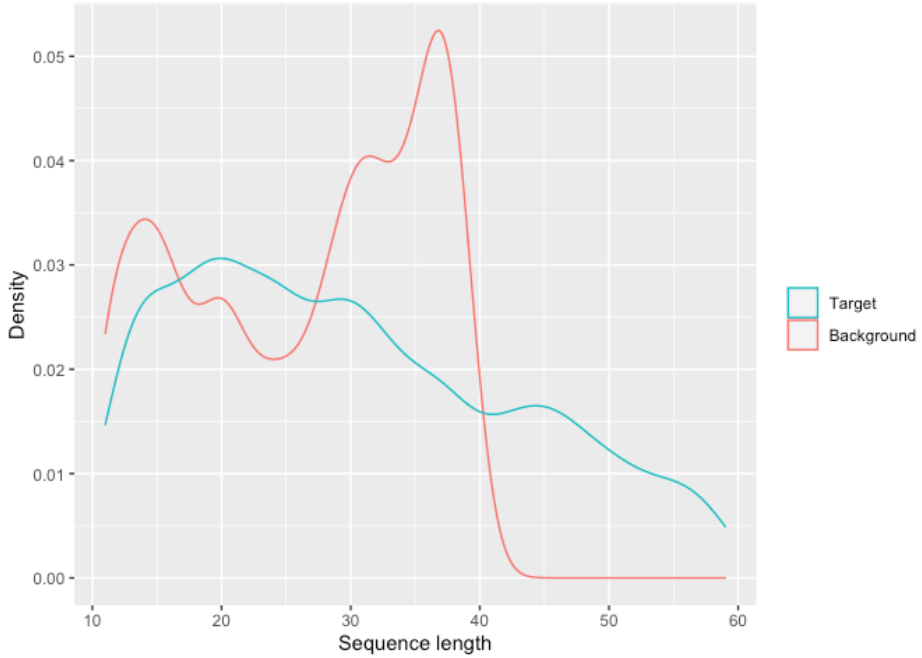


AMP datasets

- Sequences from APD, dbAMP, DRAMP and SwissProt
- Precursor:
  - UniProt proteins annotated as ‘antimicrobial’
  - Exclude all mature and unreviewed (unless also present in APD, DRAMP or dbAMP)
  - Remove proteins shorter than 50 aa and longer than 500 aa
  - Remove identical sequences and those containing nonstandard aa
  - CD-HIT 90%
- Mature
  - All AMPs from APD, DRAMP, dbAMP and mature peptides from SwissProt with lengths 11-59
  - Remove identical sequences and containing nonstandard aa
  - CD-HIT 90%

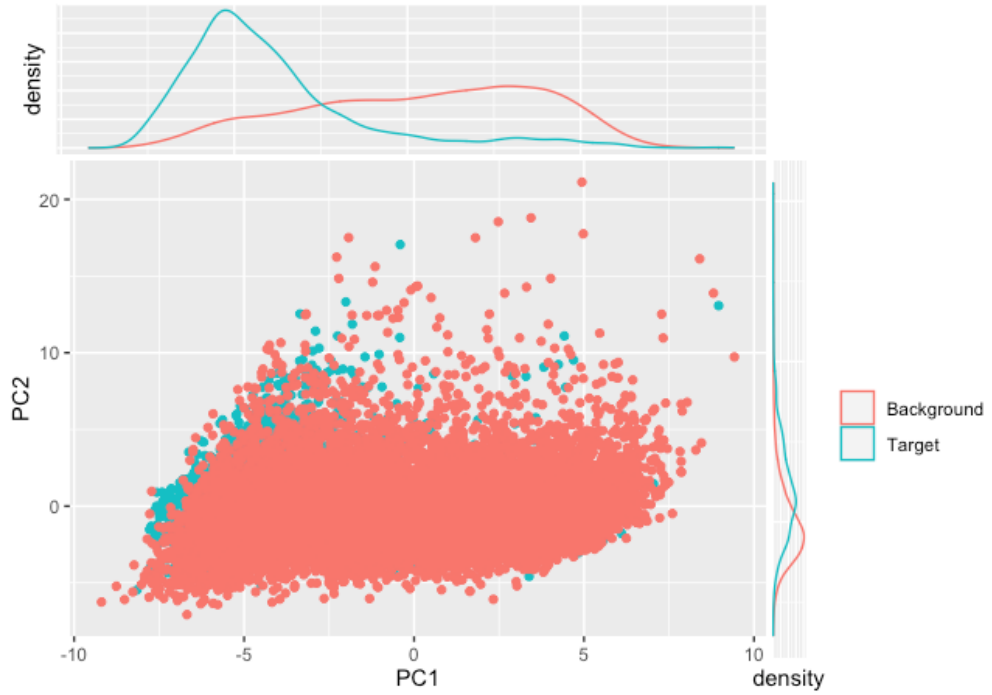
Negative datasets

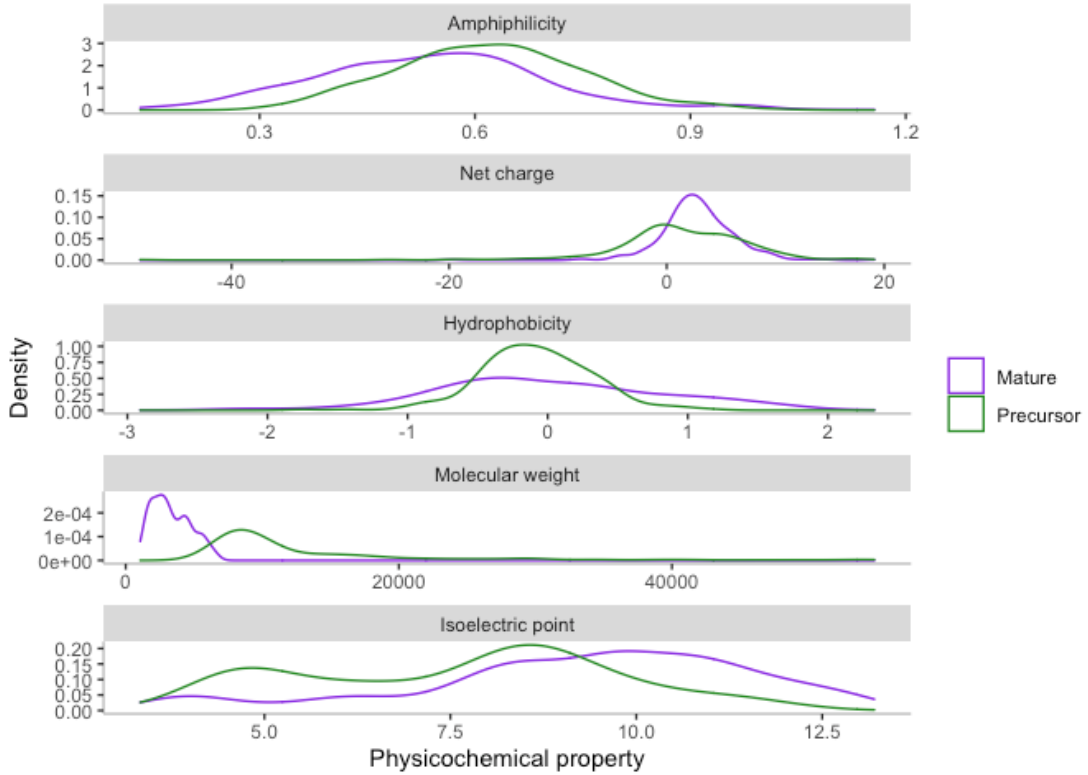
- Precursor:
  - UniProt proteins after CD-HIT 90%
  - Remove annotated as ‘antimicrobial’ and with nonstandard aa
  - Remove proteins shorter than 50 aa and longer than 500 aa
  - Sample data so the AMP:nonAMP ratio is 1:10
- Mature
  - UniProt proteins after CD-HIT 90%
  - Remove annotated as ‘antimicrobial’ and with nonstandard aa
  - Remove proteins shorter than 10 aa and longer than 40



Feature selection

- Physicochemical properties
- Pseudo-amino acid composition

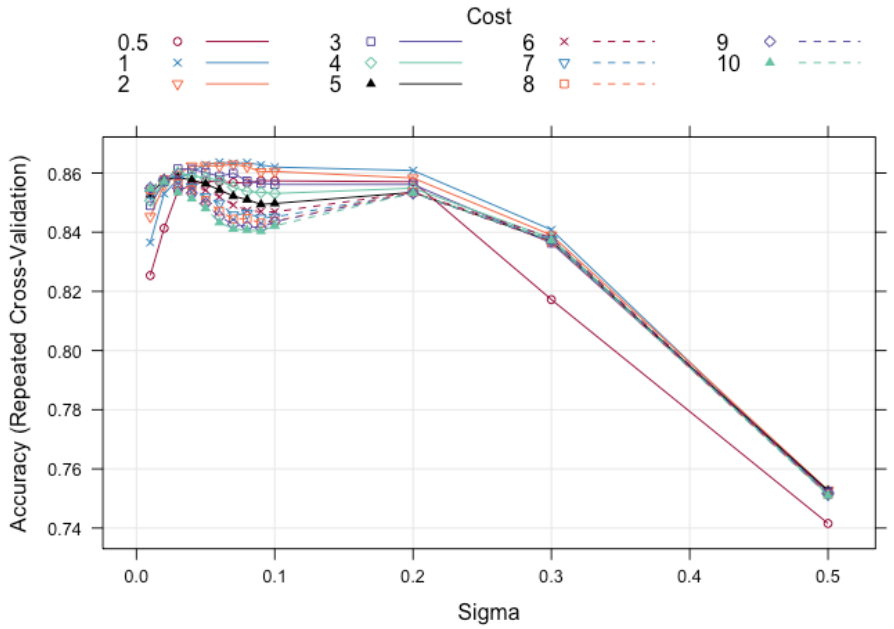




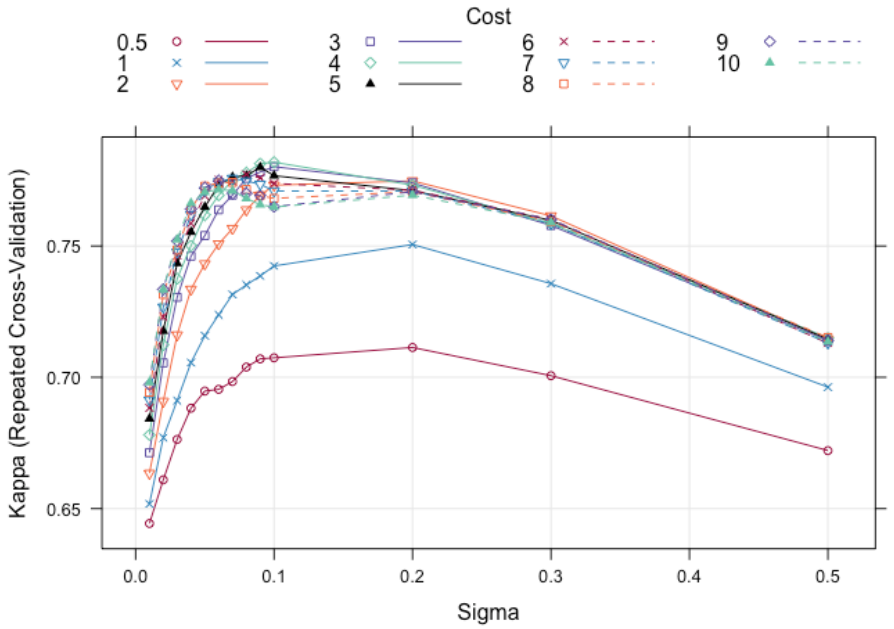
Model architecture

- SVMs with radial kernel
- Sigma and c parameters were tuned

Mature: sigma = 0.06, c = 1



Precursor: sigma = 0.1, c = 4



Benchmark

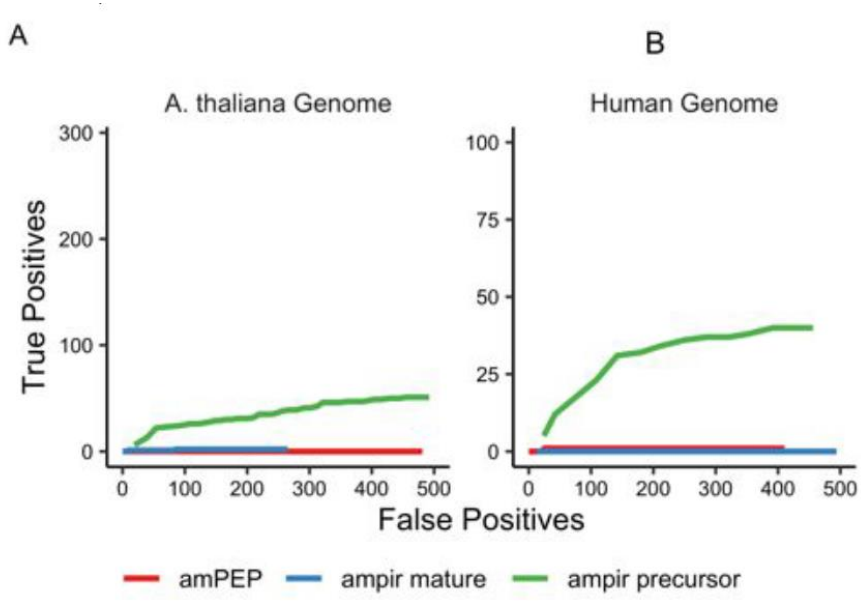
- 20% hold-out and Xiao et al. 2013 benchmark datasets were used for evaluation

Metric	ampir_mature	ampir_precursor	ampscanner v2	amPEP	iAMPpred
Xiao et al. 2013 benchmark set					
Accuracy	0.97	0.61	0.83	**1.00**	0.64
F1 score	0.97	0.44	0.85	**1.00**	0.73
AUROC	0.99	0.80	0.94	**1.00**	0.86
ampir_mature test set					
Accuracy	**0.91**	0.58	0.73	0.76	0.70
F1 score	**0.91**	0.39	0.56	0.58	0.53
AUROC	**0.97**	0.66	0.79	0.90	0.74
ampir_precursor test set					
Accuracy	0.50	**0.94**	0.70	0.46	0.47
F1 score	0.17	**0.92**	0.26	0.06	0.16
AUROC	0.84	**1.00**	0.82	0.52	0.50

Performance on whole proteome data

Numbers of known AMPs in proteomes:

- Arabidopsis: 291
- Human: 101



Summary

- ampir is a fast tool for genome-wide search of AMPs
- Model based on precursor proteins achieves the best performance on whole proteome data
- The AMP prediction in whole genomes is still imperfect