Allerdictor supplementary document

Ha X. Dang[1] and Christopher B. Lawrence[1,2,*]

[1]Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Virginia, USA.
[2]Department of Biological Sciences, Virginia Tech, Blacksburg, Virginia, USA.

[*] to whom correspondence should be addressed

Table S1. Statistics on k-mers learned from three datasets A, B, and C in relation with 183 known IgE epitopes coming from 29 allergens.

Training data	Total k-mers learned	IgE epitope-matched k-mers	Matched IgE epitopes *	Allergens of matched IgE epitopes
Dataset A	9,082,690	1,238	174	25
Dataset B	5,654,846	1,222	174	25
Dataset C	4,561,099	1,068	170	25

* not all epitopes are matched with k-mers due to some variations in the collected sequences as well as the removal of some allergen sequences in data preparation.

Table S2. Default performance measures of current methods on test set X of 167 allergens and 1,663 non-allergens randomly drawn from dataset C.

Method	TP	FP	TN	FN	Recall	Precision	Accuracy	MCC
AllerHunter*	90	43	1620	70	0.56	0.68	0.94	0.58
AlgPred-c*	133	860	803	34	0.80	0.13	0.51	0.16
AlgPred-d	125	766	897	42	0.75	0.14	0.56	0.17
APPEL*	67	49	1614	93	0.42	0.58	0.92	0.45
EVALLER*	106	100	1560	61	0.63	0.51	0.91	0.52
SORTALLER	144	354	1323	23	0.86	0.29	0.80	0.42

TP - true positive, FP - false positive, TN - true negative, FN - false negative, MCC - Mathews correlation coefficient
* server returned error on a few (<10) sequences

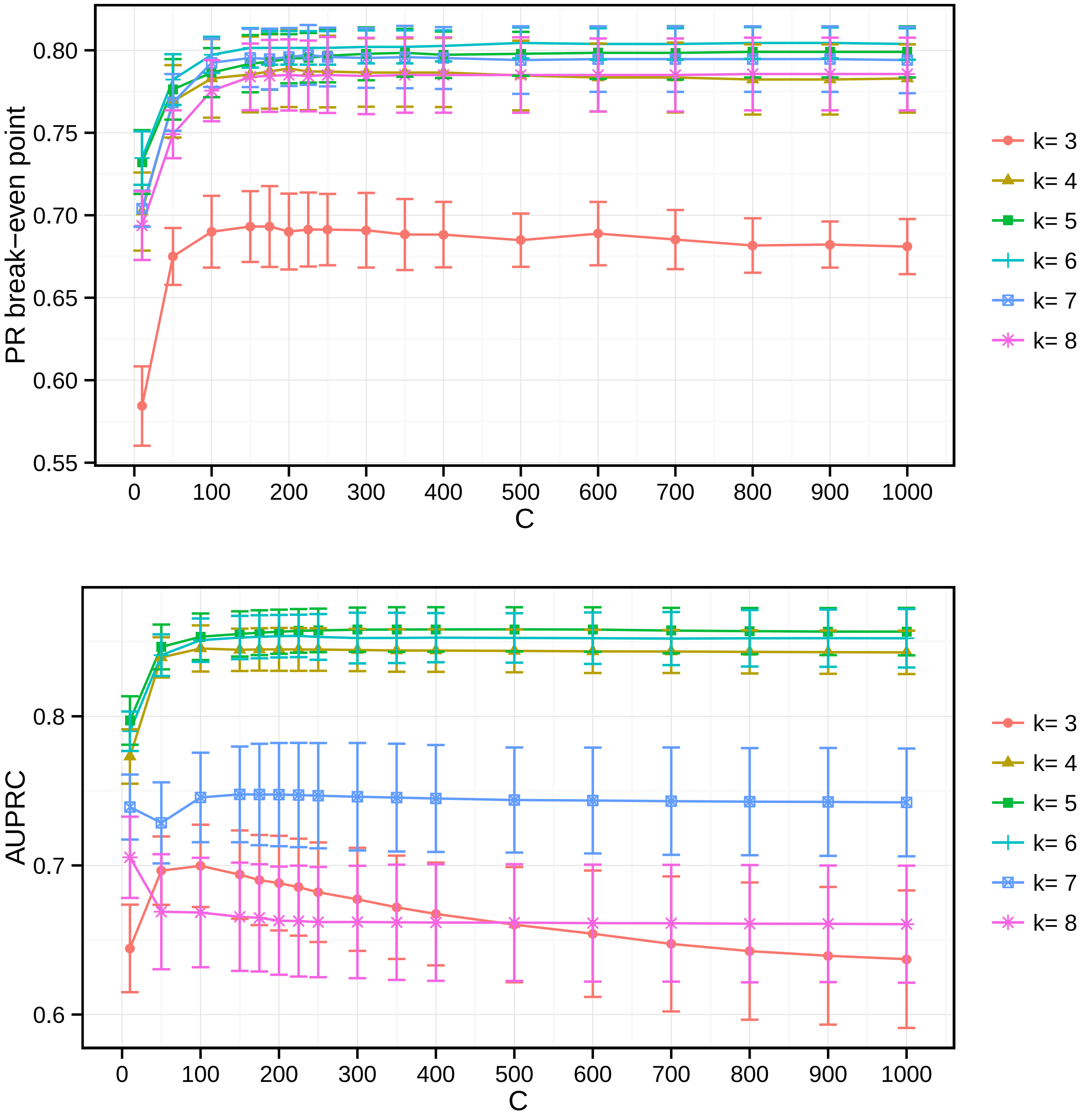

Fig. S1. AUPRCs and PR break-even points for Allerdictor 10-fold cross-validation on dataset C, with the different k-mer length (k) and regularization parameter (C) of the SVM model. The error bars show standard deviations of performance scores of 10 fold evaluation.

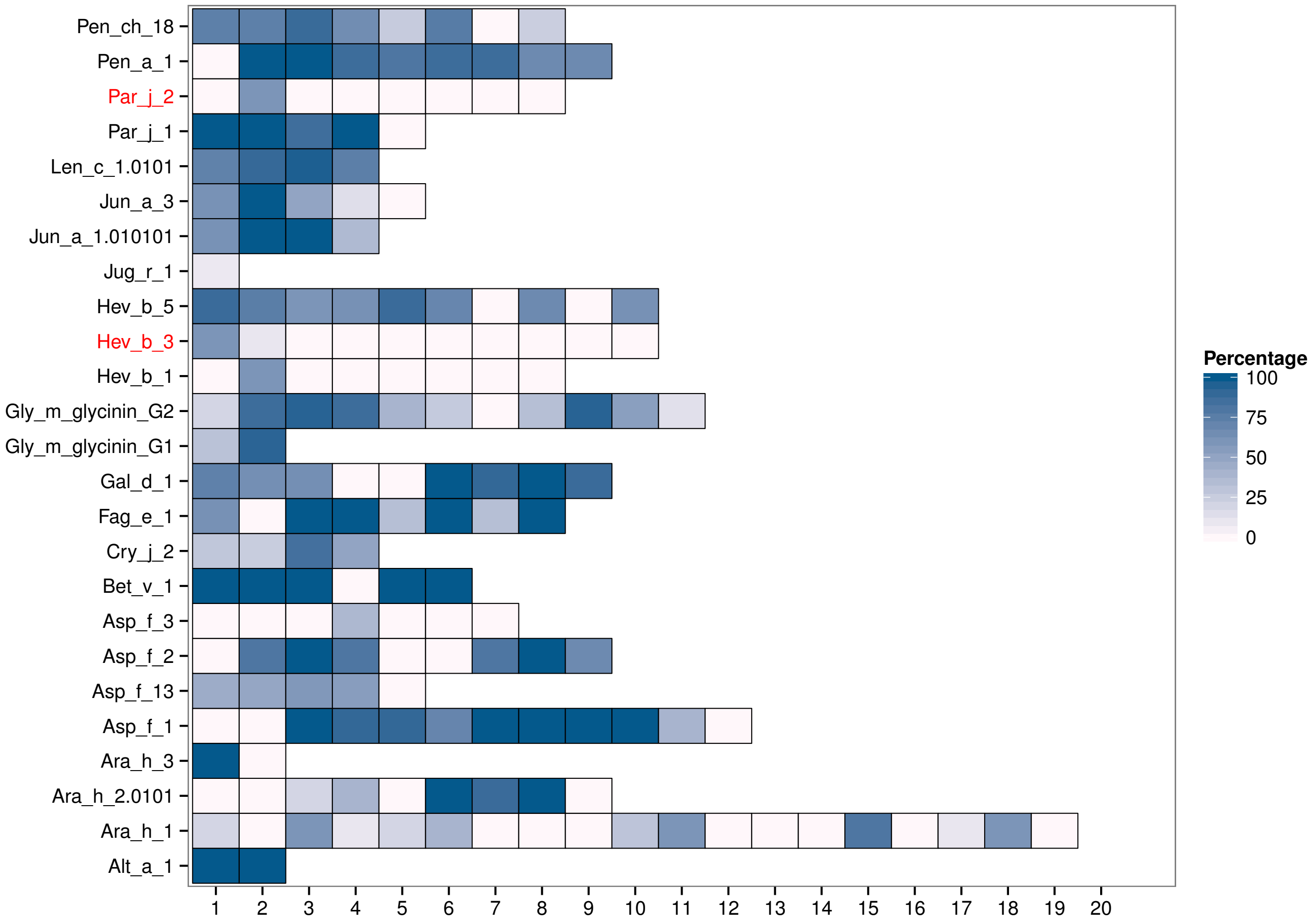

Fig. S2. Epitope coverage by high scoring regions for allergens with known IgE epitopes. High scoring regions were formed by merging consecutive k-mers that ranked in the top 25,000. Each box represents one epitope sequence (color density indicates percentage of an epitope covered by high scored regions). Sequences predicted to be allergen are labeled in black /non-allergen are in red. Training was performed on dataset A (known-epitope allergens and sequences that had a BLAST HSP of $\geq$99% identity with these allergens were removed from training data).

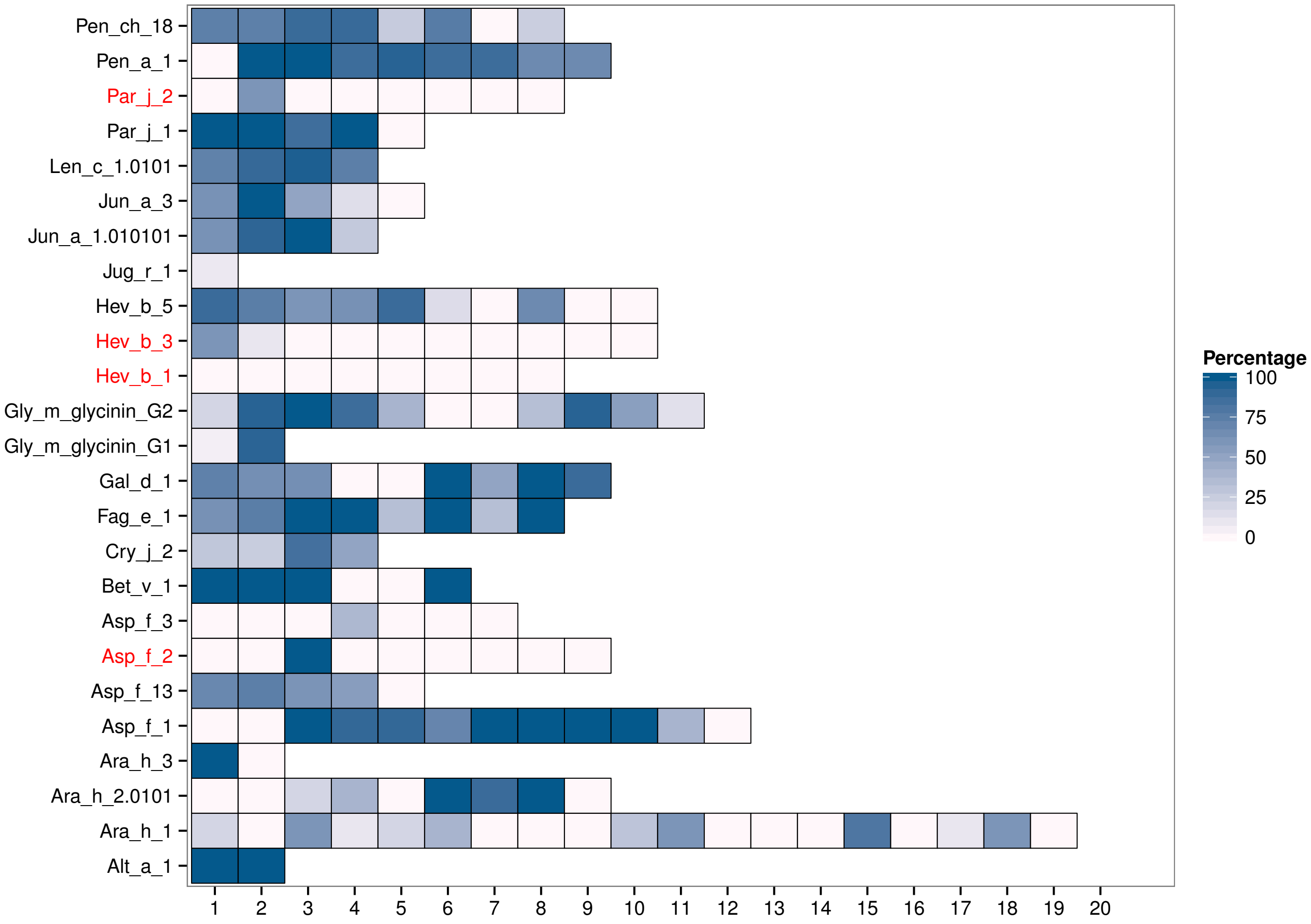

Fig. S3. Epitope coverage by high scoring regions for allergens with known IgE epitopes. High scoring regions were formed by merging consecutive k-mers that ranked in the top 25,000. Each box represents one epitope sequence (color density indicates percentage of an epitope covered by high scored regions). Sequences predicted to be allergen are labeled in black /non-allergen are in red. Training was performed on dataset B (known-epitope allergens and sequences that had a BLAST HSP of $\geq$99% identity with these allergens were removed from training data).

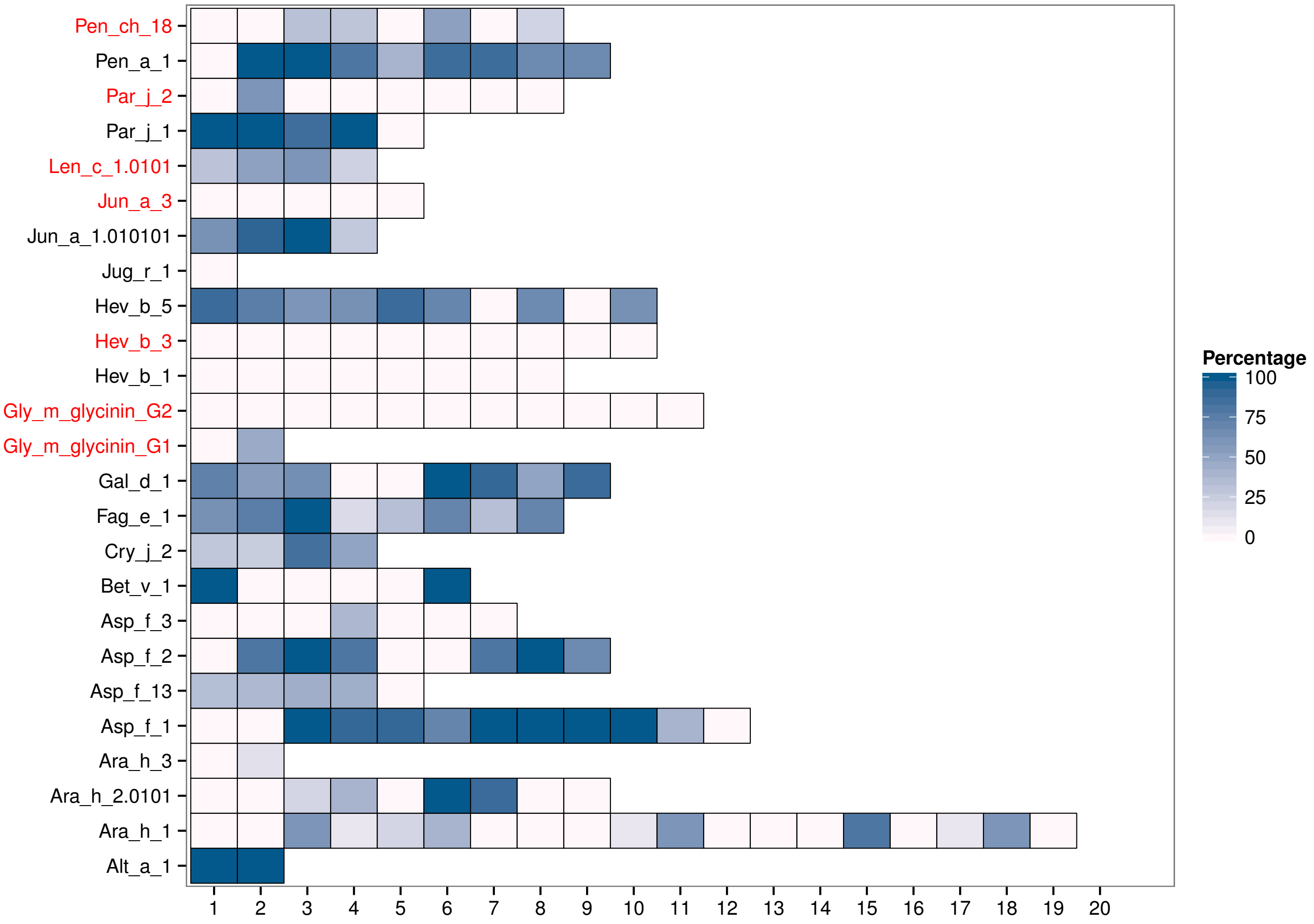

Fig. S4. Epitope coverage by high scoring regions for allergens with known IgE epitopes. High scoring regions were formed by merging consecutive k-mers that ranked in the top 25,000. Each box represents one epitope sequence (color density indicates percentage of an epitope covered by high scored regions). Sequences predicted to be allergen are labeled in black /non-allergen are in red. Training was performed on dataset C (known-epitope allergens and sequences that had a BLAST HSP of $\geq$99% identity with these allergens were removed from training data).

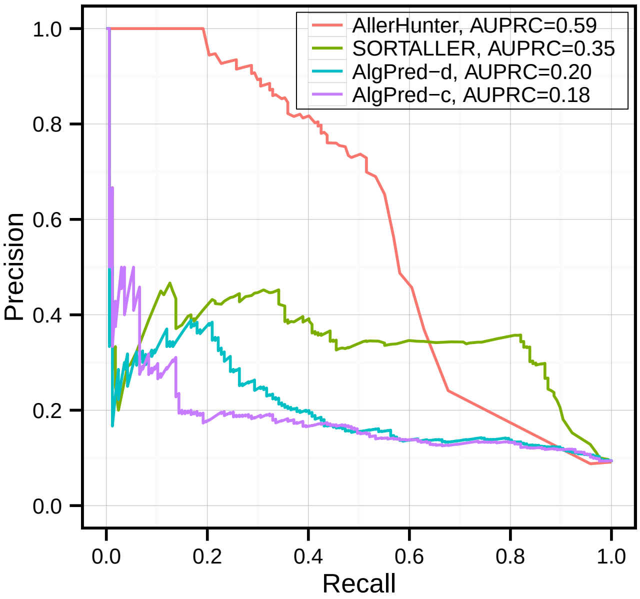

Fig. S5. PR curves for AllerHunter, AlgPred composition (AlgPred-c), AlgPred dipeptide (AlgPred-d), and SORTALLER on a test set of 167 allergens and 1,663 non-alergens randomly drawn from dataset C.

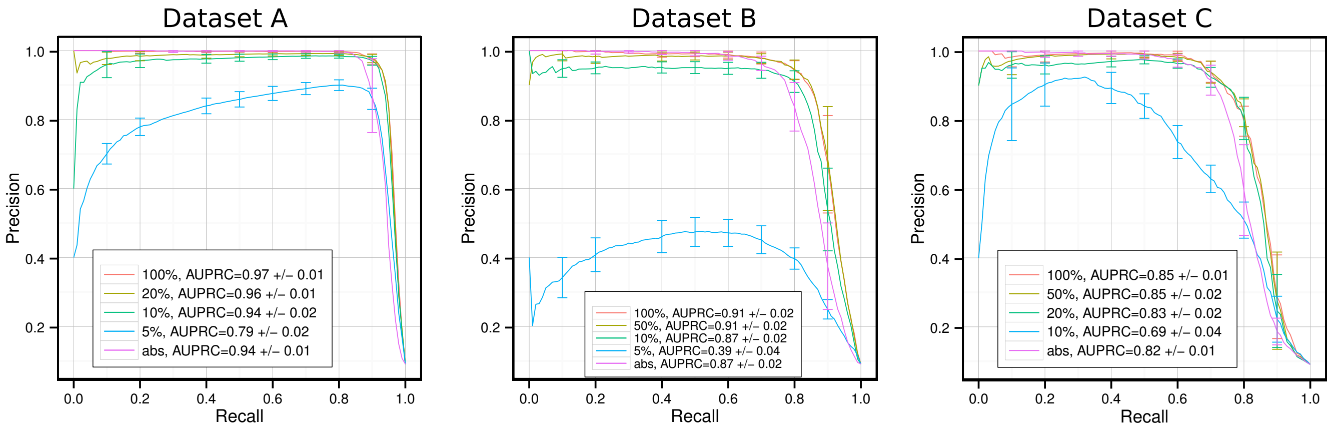

Fig. S6. PR curves for mutual information based feature selection (at 5%-100% top k-mers selected) and feature abstraction (abs) on datasets A, B, and C. The curves are average of 10-fold cross-validation with standard deviations as error bars.

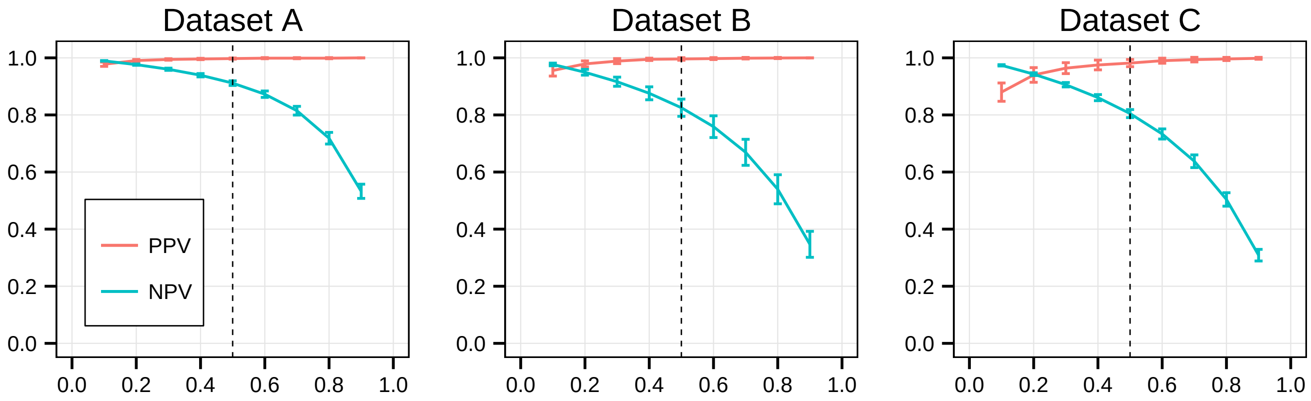

Fig. S7. Allerdictor positive predictive value (PPV) and negative predictive value (NPV) in relation to the ratio (prevalence) of allergens in test sets when trained and tested on datasets A, B, and C using nested 10-fold cross-validation. Error bars represent standard deviation of 10-fold nested cross-validation.

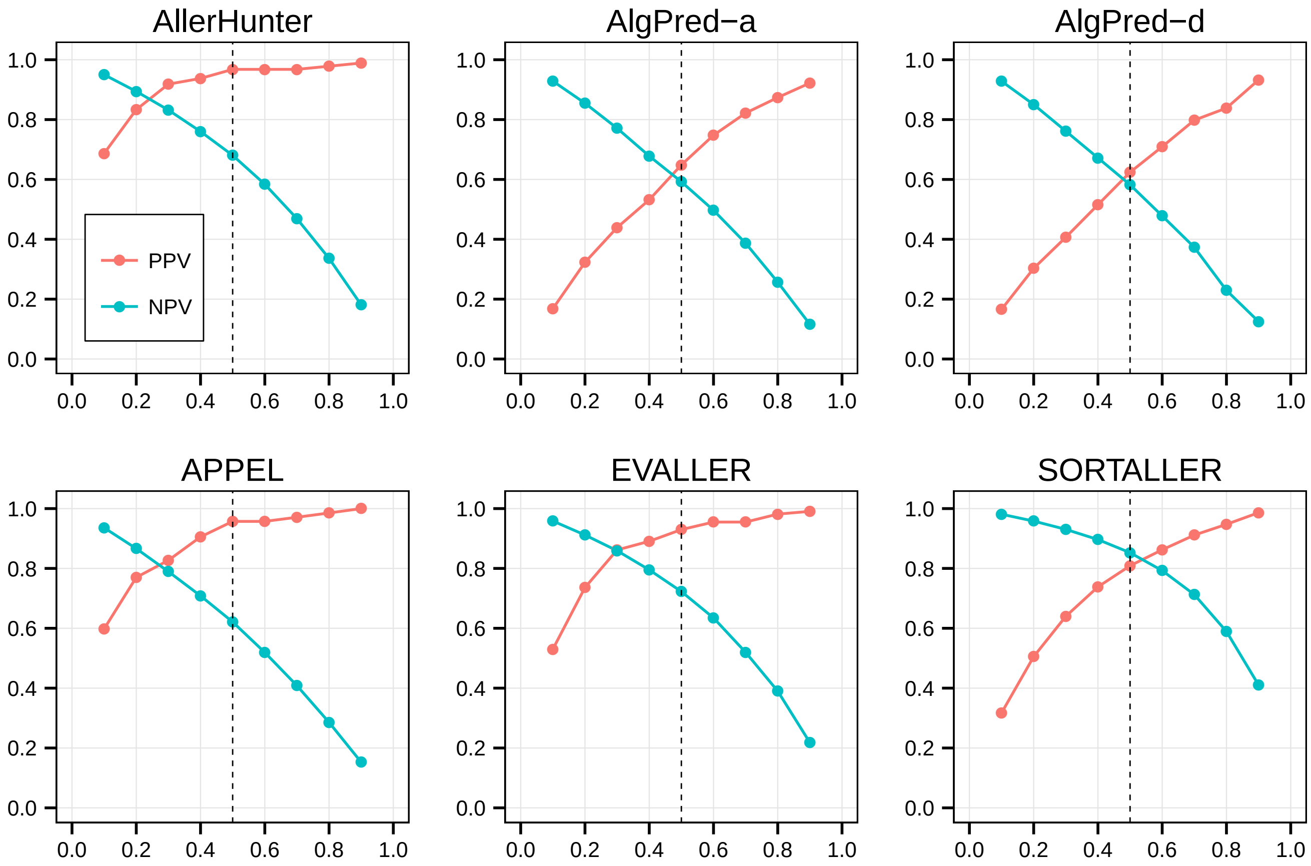

Fig. S8. Positive predictive value (PPV) and negative predictive value (NPV) of current allergen prediction tools in relation to the ratio (prevalence) of allergens in test sets. The tools were evaluated using a test set of 167 allergens and 1,663 non-allergens randomly drawn from dataset C (test set X used in section 3.5).

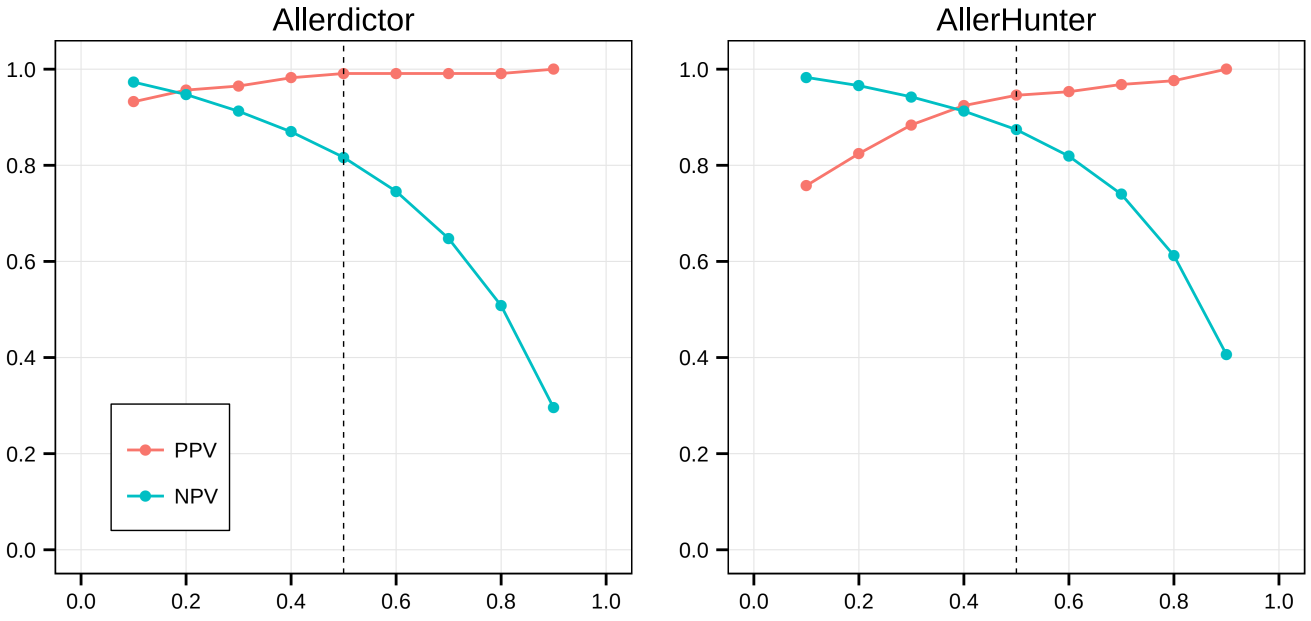

Fig. S9. Positive predictive value (PPV) and negative predictive value (NPV) of Allerdictor and AllerHunter in relation to the ratio (prevalence) of allergens in test sets. Both methods were trained using AllerHunter training set and evaluated using the revised AllerHunter test set.