Supplementary Information

AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest

Pratiti Bhadra, Jielu Yan, Jinyan Li, Simon Fong, and Shirley W. I. Siu*

Department of Computer and Information Science, University of Macau
Avenida da Universidade, Taipa, Macau, China

*Corresponding author: shirleysiu@umac.mo

Last update: 14 Dec 2017
Supplementary Table S1. Pearson correlation coefficients (PCC) of AMP/non-AMP distributions using \( M_{model_{train}} \). A descriptor is named with its physiochemical property, class, and distribution (“first residue” is coded as 001, “25% residues” as 025, “50% residues” as 050, “75% residues” as 075, “last residue” as 100). Descriptors with PCC < 0.5 are shown with boldface; those with PCC < 0.5 also in the two other datasets (\( C_{\text{train}} \), \( C_{\text{test}} \)) are marked with asterisks.

| Descriptor                  | PCC | Descriptor                  | PCC | Descriptor                  | PCC |
|-----------------------------|-----|-----------------------------|-----|-----------------------------|-----|
| Charge_C1_001               | 0.690| Polarizability_C3_025       | 0.751| NormalizedVDWV_C2_075       | 0.632|
| **Charge_C2_001**           | **0.063**| SecondaryStr_C1_025         | 0.797| NormalizedVDWV_C3_075       | 0.675|
| **Charge_C3_001**           | **0.252**| SecondaryStr_C2_025         | 0.711| Polarity_C1_075             | 0.730|
| Hydrophobicity_C1_001       | 0.747| SecondaryStr_C3_025         | 0.571| Polarity_C2_075*            | **0.311**|
| Hydrophobicity_C2_001       | 0.549| SolventAccessibility_C1_025 | 0.643| Polarity_C3_075             | 0.544|
| Hydrophobicity_C3_001*      | **0.128**| SolventAccessibility_C2_025 | 0.729| Polarity_C1_075             | 0.525|
| NormalizedVDWV_C1_001       | 0.554| SolventAccessibility_C3_025 | 0.669| Polarity_C2_075             | 0.801|
| NormalizedVDWV_C2_001       | 0.625| Charge_C1_050               | 0.597| Polarity_C3_075             | 0.675|
| NormalizedVDWV_C3_001*      | 0.301| **Charge_C2_050**           | **0.288**| SecondaryStr_C1_075         | 0.747|
| **Polarity_C1_001**         | **0.130**| Charge_C3_050               | **0.268**| SecondaryStr_C2_075         | 0.807|
| Polarity_C2_001             | 0.512| Hydrophobicity_C1_050       | 0.679| SecondaryStr_C3_075         | 0.627|
| Polarity_C3_001             | 0.689| Hydrophobicity_C2_050       | 0.719| SolventAccessibility_C1_075 | 0.727|
| Polarizability_C1_001       | 0.584| Hydrophobicity_C3_050       | 0.683| SolventAccessibility_C2_075 | 0.609|
| Polarizability_C2_001       | 0.614| NormalizedVDWV_C1_050       | 0.628| SolventAccessibility_C3_075 | 0.613|
| **Polarizability_C3_001**   | **0.301**| NormalizedVDWV_C2_050       | 0.627| Charge_C1_100               | 0.571|
| **SecondaryStr_C1_001**     | **0.346**| NormalizedVDWV_C3_050       | 0.672| **Charge_C2_100**           | **0.484**|
| SecondaryStr_C2_001         | 0.657| Polarity_C1_050             | 0.710| **Charge_C3_100**           | **0.184**|
| SecondaryStr_C3_001         | 0.768| **Polarity_C2_050**         | **0.357**| Hydrophobicity_C1_100*      | **0.464**|
| SolventAccessibility_C1_001 | 0.432| Polarity_C3_050             | 0.648| Hydrophobicity_C2_100       | 0.559|
| SolventAccessibility_C2_001 | 0.747| Polarizability_C1_050       | 0.636| Hydrophobicity_C3_100       | 0.652|
| **SolventAccessibility_C3_001** | **0.296**| Polarizability_C2_050       | 0.684| NormalizedVDWV_C1_100       | 0.529|
| Charge_C1_025               | 0.588| Polarizability_C3_050       | 0.672| NormalizedVDWV_C2_100       | 0.558|
| Charge_C2_025               | 0.661| SecondaryStr_C1_050         | 0.720| NormalizedVDWV_C3_100       | 0.567|
| **Charge_C3_025**           | **0.237**| SecondaryStr_C2_050         | 0.739| Polarity_C1_100             | 0.645|
| Hydrophobicity_C1_025       | 0.729| SecondaryStr_C3_050         | 0.640| Polarity_C2_100             | 0.521|
| Feature                      | Value   | Feature                      | Value   | Feature                      | Value   |
|------------------------------|---------|------------------------------|---------|------------------------------|---------|
| Hydrophobicity_C2_025        | 0.782   | SolventAccessibility_C1_050 | 0.432   | Polarity_C3_100              | 0.570   |
| Hydrophobicity_C3_025        | 0.640   | SolventAccessibility_C2_050 | 0.679   | Polarity_C1_100              | 0.503   |
| NormalizedVDWV_C1_025        | 0.703   | SolventAccessibility_C3_050 | 0.629   | Polarity_C2_100              | 0.565   |
| NormalizedVDWV_C2_025        | 0.617   | Charge_C1_075                | 0.473   | Polarity_C3_100              | 0.567   |
| NormalizedVDWV_C3_025        | 0.751   | Charge_C2_075                | 0.782   | SecondaryStr_C1_100*         | 0.420   |
| Polarity_C1_025              | 0.636   | Charge_C3_075                | 0.231   | SecondaryStr_C2_100          | 0.723   |
| Polarity_C2_025*             | 0.315   | Hydrophobicity_C1_075        | 0.609   | SecondaryStr_C3_100          | 0.618   |
| Polarity_C3_025              | 0.657   | Hydrophobicity_C2_075        | 0.696   | SolventAccessibility_C1_100  | 0.546   |
| Polarizability_C1_025        | 0.705   | Hydrophobicity_C3_075        | 0.727   | SolventAccessibility_C2_100* | 0.464   |
| Polarizability_C2_025        | 0.731   | NormalizedVDWV_C1_075        | 0.638   | SolventAccessibility_C3_100  | 0.639   |
Supplementary Table S2. Datasets generated from $M_{\text{model_train}}$ for P:N ratio tests of AMP prediction. Size of the positive dataset is 3268.

| P:N ratio | Size of one non-AMP subset | Total number of non-AMP subsets |
|-----------|-----------------------------|---------------------------------|
| 1:1       | 3268                        | 51                              |
| 1:1.5     | 4902                        | 34                              |
| 1:2       | 6536                        | 26                              |
| 1:2.5     | 8170                        | 20                              |
| 1:3       | 9804                        | 17                              |
| 1:3.5     | 11438                       | 15                              |
| 1:4       | 13072                       | 13                              |
| 1:4.5     | 14706                       | 11                              |
| 1:5       | 16340                       | 10                              |
| 1:5.5     | 17974                       | 9                               |
| 1:6       | 19608                       | 9                               |
| 1:6.5     | 21242                       | 8                               |
| 1:7       | 22876                       | 7                               |
| 1:7.5     | 24510                       | 7                               |
| 1:8       | 26144                       | 6                               |
| 1:8.5     | 27778                       | 6                               |
| 1:9       | 29412                       | 6                               |
| 1:9.5     | 31046                       | 5                               |
| 1:10      | 32680                       | 5                               |
Supplementary Table S3. Performance of RF classifiers using different P:N ratios in 10-fold cross validation. Values shown are averages and standard deviations (in brackets) over all corresponding subsets. The optimal model based on C-measure is ratio 1:3.

| P:N ratio | Sn     | Sp     | Acc    | MCC     | AUC-ROC | AUC-PR | Kappa | C-measure |
|-----------|--------|--------|--------|---------|---------|--------|-------|-----------|
| 1:1       | 0.978  | 0.945  | 0.962  | 0.924   | 0.988   | 0.698  | 0.923 | 0.588     |
|           | (0.002)| (0.004)| (0.002)| (0.004) | (0.001)| (0.024)| (0.005)|          |
| 1:1.5     | 0.972  | 0.952  | 0.960  | 0.917   | 0.989   | 0.755  | 0.917 | 0.628     |
|           | (0.002)| (0.003)| (0.002)| (0.004) | (0.001)| (0.017)| (0.004)|          |
| 1:2       | 0.965  | 0.957  | 0.960  | 0.912   | 0.989   | 0.791  | 0.911 | 0.650     |
|           | (0.003)| (0.002)| (0.002)| (0.004) | (0.001)| (0.014)| (0.005)|          |
| 1:2.5     | 0.958  | 0.961  | 0.961  | 0.906   | 0.989   | 0.814  | 0.905 | 0.660     |
|           | (0.003)| (0.002)| (0.002)| (0.004) | (0.001)| (0.015)| (0.004)|          |
| 1:3       | 0.950  | 0.965  | 0.962  | 0.900   | 0.989   | 0.830  | 0.899 | 0.665     |
|           | (0.003)| (0.002)| (0.002)| (0.004) | (0.000)| (0.009)| (0.004)|          |
| 1:3.5     | 0.943  | 0.968  | 0.962  | 0.893   | 0.989   | 0.840  | 0.893 | 0.663     |
|           | (0.004)| (0.001)| (0.002)| (0.005) | (0.001)| (0.010)| (0.005)|          |
| 1:4       | 0.936  | 0.970  | 0.963  | 0.888   | 0.989   | 0.849  | 0.888 | 0.663     |
|           | (0.004)| (0.002)| (0.002)| (0.005) | (0.001)| (0.007)| (0.005)|          |
| 1:4.5     | 0.929  | 0.973  | 0.965  | 0.884   | 0.989   | 0.857  | 0.884 | 0.662     |
|           | (0.004)| (0.001)| (0.002)| (0.005) | (0.000)| (0.007)| (0.005)|          |
| 1:5       | 0.921  | 0.974  | 0.965  | 0.878   | 0.989   | 0.858  | 0.877 | 0.653     |
|           | (0.004)| (0.001)| (0.001)| (0.003) | (0.001)| (0.004)| (0.003)|          |
| 1:5.5     | 0.915  | 0.975  | 0.966  | 0.873   | 0.989   | 0.862  | 0.873 | 0.649     |
|           | (0.006)| (0.001)| (0.001)| (0.004) | (0.000)| (0.006)| (0.004)|          |
| 1:6       | 0.908  | 0.977  | 0.967  | 0.868   | 0.989   | 0.862  | 0.867 | 0.642     |
|           | (0.005)| (0.001)| (0.001)| (0.004) | (0.001)| (0.007)| (0.004)|          |
| 1:6.5     | 0.902  | 0.978  | 0.968  | 0.863   | 0.989   | 0.864  | 0.863 | 0.637     |
|           | (0.006)| (0.001)| (0.001)| (0.006) | (0.001)| (0.005)| (0.006)|          |
| 1:7       | 0.894  | 0.979  | 0.968  | 0.858   | 0.989   | 0.864  | 0.858 | 0.629     |
|           | (0.004)| (0.001)| (0.001)| (0.004) | (0.001)| (0.005)| (0.004)|          |
| 1:7.5     | 0.889  | 0.980  | 0.969  | 0.854   | 0.989   | 0.864  | 0.854 | 0.623     |
|           | (0.007)| (0.001)| (0.001)| (0.007) | (0.001)| (0.007)| (0.006)|          |
| 1:8       | 0.882  | 0.981  | 0.970  | 0.850   | 0.989   | 0.863  | 0.850 | 0.616     |
|           | (0.009)| (0.001)| (0.001)| (0.008) | (0.001)| (0.004)| (0.008)|          |
| 1:8.5     | 0.875  | 0.982  | 0.971  | 0.846   | 0.989   | 0.859  | 0.846 | 0.608     |
|           | (0.009)| (0.001)| (0.001)| (0.007) | (0.000)| (0.005)| (0.007)|          |
| 1:9       | 0.869  | 0.982  | 0.971  | 0.841   | 0.989   | 0.858  | 0.841 | 0.601     |
|           | (0.007)| (0.000)| (0.001)| (0.006) | (0.000)| (0.006)| (0.006)|          |
| 1:9.5     | 0.861  | 0.983  | 0.972  | 0.837   | 0.989   | 0.857  | 0.836 | 0.593     |
|           | (0.005)| (0.001)| (0.001)| (0.006) | (0.000)| (0.002)| (0.006)|          |
| 1:10      | 0.859  | 0.984  | 0.972  | 0.835   | 0.989   | 0.857  | 0.835 | 0.590     |
|           | (0.006)| (0.001)| (0.001)| (0.005) | (0.001)| (0.006)| (0.005)|          |
**Supplementary Table S4.** Comparison of RF and SVM classifiers using D_f features and AMP/non-AMP data ratio of 1:3 in 10-fold cross-validation. Values shown are averages and standard deviations (in brackets) over all corresponding subsets.

| Method | Sn       | Sp       | Acc      | MCC      | AUC-ROC | AUC-PR  | Kappa   | CMeasure |
|--------|----------|----------|----------|----------|---------|---------|---------|----------|
| RF     | 0.950(0.003) | 0.965(0.002) | 0.962(0.002) | 0.900(0.004) | 0.989(0.000) | 0.830(0.009) | 0.889(0.004) | 0.665(0.006) |
| SVM    | 0.532(0.042) | 0.949(0.006) | 0.844(0.012) | 0.552(0.038) | 0.813(0.030) | 0.681(0.034) | 1.0(0.000) | 0.305(0.047) |
**Supplementary Table S5.** A comparison of RF classifiers using different descriptors by 10-fold cross-validation with the AMP data ratio of 1:1. Values shown are averages and standard deviations (in brackets) over 10 times of 10-fold cross validation. The best two results in each performance measure are highlighted.

| Feature set {#}                                      | Sn       | Sp       | Acc      | MCC      | AUC-ROC   | AUC-PR   | Kappa    | C-measure |
|------------------------------------------------------|----------|----------|----------|----------|-----------|----------|----------|-----------|
| AmPEP [105]                                          | 0.978    | 0.945    | 0.962    | 0.924    | 0.988     | 0.698    | 0.923    | 0.588     |
| AAC [20]                                             | 0.948    | 0.946    | 0.947    | 0.894    | 0.985     | 0.77     | 0.894    | 0.606     |
| PAAC [24]                                            | 0.948    | 0.945    | 0.947    | 0.893    | 0.984     | 0.822    | 0.893    | 0.645     |
| K-mer [400]                                          | 0.939    | 0.944    | 0.941    | 0.883    | 0.983     | 0.876    | 0.883    | 0.671     |
| Auto Covariance (AC) {6}                             | 0.761    | 0.844    | 0.802    | 0.606    | 0.870     | 0.814    | 0.604    | 0.259     |
| Cross Covariance (CC) {12}                           | 0.802    | 0.85     | 0.826    | 0.653    | 0.897     | 0.851    | 0.652    | 0.325     |
| Auto-Cross Covariance (ACC) {18}                     | 0.83     | 0.863    | 0.846    | 0.693    | 0.914     | 0.863    | 0.693    | 0.379     |
| Parallel Correlation Pseudo Amino Acid Composition    | 0.948    | 0.945    | 0.947    | 0.893    | 0.984     | 0.806    | 0.893    | 0.633     |
| (PC-PseAAC) {22}                                     |          |          |          |          |           |          |          |           |
| General Parallel Correlation Pseudo Amino Acid Composition | 0.946    | 0.942    | 0.944    | 0.888    | 0.984     | 0.823    | 0.888    | 0.639     |
| (PC-PseAAC-General) {22}                             |          |          |          |          |           |          |          |           |
| Parallel Series Correlation Pseudo Amino Acid Composition | 0.946    | 0.943    | 0.944    | 0.889    | 0.983     | 0.822    | 0.889    | 0.639     |
| (SC-PseAAC-General) {26}                             |          |          |          |          |           |          |          |           |

AAC: Amino Acid Composition, PAAC: Pseudo Amino Acid Composition
AAC and PseAAC were generated using propy 1.0 package (default parameter of propy is used).
Other descriptors, K-mer, AC, CC, ACC, PC-PseAAC, SC-PseAAC, PC-PseAAC-General, SC-PseAAC-General were generated by Pse-in-One-1.0.4 using default parameters.
Supplementary Figure S1. Comparison of the AMP and non-AMP statistical distributions of 105 descriptors.
Supplementary Figure S2. Comparison of the average descriptor value of “first residue” and “100% residues” computed from the AMP sequences of three datasets (M_{model\_train}, C_{test} and C_{train}). Standard deviations are shown as error bars.