Supplementary Materials

Table. 4. The details of the features that is used in our work. These features are used to analyse the distribution of the feature importance scores.

Feature extraction methods	Details
AB-PSSM	The features extracted through PSSM-based method AB-PSSM (1×400)
DP-PSSM	The features extracted through PSSM-based method DP-PSSM (1×240)
PSSM-composition	The features extracted through PSSM-based method PSSM-composition (1×400)
RPM-PSSM	The feature extracted through residue probing method (RPM-PSSM), a PSSM-based method (1×400)
S-FPSSM	The features extracted through PSSM-based method S-FPSSM (1×400)
DPC-PSSM	The dipeptide-composition features extracted through DPC-PSSM, a PSSM-based method (1×400)
EEDP-PSSM	The features based on evolutionary difference formula, a PSSM-based feature extraction method (1×400)
KSB-PSSM	The k-separated bigrams features extracted through PSSM-based method KSB-PSSM (1×400)
CKSAAGP	These features are composition of k-spaced amino acid pairs, extracted by sequence-based method CKSAAP ($\mathrm{k} = 5$). And the sequence is simplified through 5-letter alphabet. (1×150)
CKSAAGP-chem	These features are composition of k-spaced amino acid pairs, extracted by sequence-based method CKSAAP ($\mathrm{k} = 0$). And the sequence is simplified through chemical-property alphabet. (1×64)
AAC	Amino acid composition features extracted by sequence-based method AAC. They consists of AAC features of primitive sequence (1×20) and sequence simplified by both two kinds of amino acid alphabets, 5-letter alphabet (1×5) and chemical-property alphabet (1×8). (1×33)
TPC	These features are tripeptide composition extracted by sequence-based method TPC. And the sequence is simplified through 5-letter alphabet. (1×125)
TPC-chem	These features are tripeptide composition extracted by sequence-based method TPC. And the sequence is simplified through chemical-property alphabet. (1×512)
CT	Conjoint triad features extracted throu a sequence-based method CT (1×343)

Table. 6. The details of the subsets, which are used to find the best combination of features.

Subsets	Details

all	Consists of all the features used in the work of feature selection (a 1×4267 vector)
subset_30	consists of 30 top-scoring (higher than 4.03×10^{-4}) features selected through PU extra trees (a 1×30 vector)
subset_50	consists of 50 top-scoring (higher than 3.47×10^{-4}) features selected through PU extra trees (a 1×50 vector)
subset_100	consists of 100 top-scoring (higher than 2.72×10^{-4}) features selected through PU extra trees (a 1×100 vector)
subset_150	consists of 150 top-scoring (higher than 2.37×10^{-4}) features selected through PU extra trees (a 1×150 vector)
subset_200	consists of 200 top-scoring (higher than 2.2×10^{-4}) features selected through PU extra trees (a 1×201 vector)
gaac	33-dimension AAC features (a 1×33 vector)
cksaagp	features of CKSAAP (k=5) simplified through 5-letter alphabet (a 1×150 vector)
cksaagpchem0	features of CKSAAP (k=0) simplified through chemical-property alphabet (a 1×64 vector)
cksaagpchem1	features of CKSAAP (k=1) simplified through chemical-property alphabet (a 1×128 vector)
cksaagpall	consists of CKSAAGP and CKS_chem_0 (a 1×214 vector)
gaac_cksaagp	consists of GAAC and CKSAAGP (a 1 ×183 vector)
pssm_30	consists of 30 top-scoring (higher than 3.39×10^{-4}) PSSM-based features selected through PU extra trees (a 1×30 vector)
finalpos	consists of GAAC_CKS and pssm_30 (a 1×213 vector)