**Supplementary Information**

**Protein–RNA binding residues prediction method based on deep feature fusion**

Zhen Feng , Hui Yu, Hongwei Wu, Lichuan Gu , Ke Li and Xiaobo Zhou

School of Information and Artificial Intelligence, Anhui Agricultural University, Hefei,Anhui 230601,China.

All correspondence should be addressed to Ke Li (E-mail: [like@ahau.edu.cn)](mailto:chunhuali@bjut.edu.cn))

**Table S1.** The 2624 features used in our study

|  |  |  |  |
| --- | --- | --- | --- |
| Feature type | Feature number | Feature name | Description |
| Physical and chemical properties of amino acids | 1 | Da | The unit of molecular mass,Dalton |
| 2 | Hydrophobicity | Measures the tendency of an amino acid to avoid water, indicating its non-polar nature |
| 3 | Hydrophilicity | Measures the affinity of an amino acid for water, indicating its polar nature |
| 4 | Electrostatic Charge | Represents the net charge of an amino acid at a specific pH, categorized as positive, negative, or neutral |
| 5 | Potential Hydrogen Bond Count | The number of potential hydrogen bonds an amino acid can form |
| 6 | Polarity | Indicates the degree of an amino acid's polarity, affecting its interaction with other molecules |
| 7 | Polarizability | Measures an amino acid's ability to be polarized by an external electric field |
| 8 | Propensity | The tendency of an amino acid to occur in certain structural motifs, like α-helices or β-sheets |
| 9 | Average Accessible Surface Area | The average surface area of an amino acid exposed to the solvent |
| 10 | Actual Hydrogen Bond Count | The actual number of hydrogen bonds an amino acid participates in |
| 11 | Bull & Breese hydrophobicity | Bull & Breese hydrophobicity |
| 12 | Grantham polarity | Grantham polarity |
| 13 | Zimmerman polarity | Zimmerman polarity |
| 14 | Kate & Doolittle hydrophobicity | Kate & Doolittle hydrophobicity |
| 15 | isoelectric point | isoelectric point |
| 16 | bulkiness | bulkiness |
| 17 | Eisenberg hydrophobicity | Eisenberg hydrophobicity |
| 18 | Hope & Woods hydrophobicity propensities | Hope & Woods hydrophobicity propensities |
| RASA、DPX & CX | 19 | RASA | relative solvent accessible surface area |
| 20 | average DPX | The average depth index value |
| 21 | average CX | The average Protrusion index value |
| Position-Specific Scoring Matrix | 22 | PSSM1 | PSSM1 |
| …… | …… | …… |
| 41 | PSSM20 | PSSM20 |
| Position-Specific Scoring Matrix based on spatial neighborhood | 42 | SNB-PSSM1 | SNB-PSSM1 |
| …… | …… | …… |
| 541 | SNB-PSSM500 | SNB-PSSM500 |
| SPIDER3-based features | 542 | ASA | solvent accessible surface area |
| 543 | Phi | torsion angle descriptors：Phi |
| 544 | Psi | torsion angle descriptors：Psi |
| 545 | Theta(i-1=>i+1) | backbone angle descriptors between Cα atoms ：Theta |
| 546 | Tau(i-2=>i+2) | backbone angle descriptors between Cα atoms ：Tau |
| 547 | HSE\_alpha\_up | hemisphere exposure within upper hemispheres |
| 548 | HSE\_alpha\_down | hemisphere exposure within lower  hemispheres |
| 549 | P(C) | position information： |
| 550 | P(H) | position information：α helix |
| 551 | P(E) | position information：β fold |
| Topology characteristics | 552 | tdegree | tdegree |
| 553 | cluster\_coefficient | cluster\_coefficient |
| 554 | degrees\_centrality | degrees\_centrality |
| 555 | closenesses\_centrality | closenesses\_centrality |
| 556 | betweennesses | betweennesses |
| Interfacial propensity | 557 | IP | Interfacial propensity |
| Dynamic parameters | 558 | F1 | relative residual fluctuations of the m1 motional modes |
| 559 | F2 | relative residual fluctuations of the m2 motional modes |
| 560 | F3 | relative residual fluctuations of the m3 motional modes |
| 567 | F4 | relative residual fluctuations of the m4 motional modes |
| 568 | F5 | relative residual fluctuations of the m5 motional modes |
| 569 | F6 | relative residual fluctuations of the m6 motional modes |
| Supplement 0 | 570-576 | 0 | 0 |
| Sequence Embedding features | 577-1600 | Total 1024 dimensions | Embedding features extracted based on Protbert protein language model |
| graph Embedding features | 1601-2624 | Total 1024 dimensions | Embedding features extracted based on graph attention network |

**Table S2.** The numerical values of 18 physicochemical properties of the 20 amino acids

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Residue（1-10）** | **Da** | **Hydrophobicity** | **Hydrophilicity** | **Electrostatic Charge** | **Potential Hydrogen Bond Count** | **Polarity** | **Polarizability** | **Propensity** | **Average Accessible Surface Area** | **Actual Hydrogen Bond Count** |
| A | 1 | 2.01 | 0.328 | 0.937 | 7 | 0.22 | 0.35 | -0.01 | 1.6 | 0.0373 |
| C | 1 | 1.98 | 0 | 1.004 | 5.5 | 0.2 | 0.5 | 0.12 | 1.2 | 0.0829 |
| D | 3.2 | -2.05 | 3.379 | 1.64 | 13 | 0.73 | 2.16 | 0.15 | 2.6 | 0.1263 |
| E | 1.7 | 0.93 | 0 | 0.679 | 12.5 | 0.08 | 0.65 | 0.07 | 2 | 0.0058 |
| F | 1 | 2.68 | 1.336 | 0.803 | 5 | 0.08 | 0.89 | 0.03 | 0.9 | 0.0946 |
| G | 1 | 0.12 | 0.5 | 0.901 | 7.9 | 0.58 | 2.4 | 0 | 0.9 | 0.005 |
| H | 1 | -0.14 | 1.204 | 1.085 | 8.4 | 0.14 | 1.19 | 0.08 | 0.7 | 0.0242 |
| I | 0.6 | 3.7 | 2.078 | 0.178 | 4.9 | 0.22 | 0.12 | -0.01 | 0.7 | 0 |
| K | 0.7 | 2.55 | 0.835 | 1.254 | 10.1 | 0.27 | 0.83 | 0 | 1 | 0.0371 |
| L | 1 | 2.73 | 0.414 | 0.808 | 4.9 | 0.19 | 0.58 | -0.01 | 0.3 | 0 |
| M | 1 | 1.75 | 0.982 | 0.886 | 5.3 | 0.38 | 0.22 | 0.04 | 1 | 0.0823 |
| N | 1.7 | 0.03 | 1.498 | 1.08 | 10 | 0.42 | 2.12 | 0.06 | 0.7 | 0.0036 |
| P | 1 | 0.41 | 0.415 | 0.748 | 6.6 | 0.46 | 0.43 | 0 | 0.5 | 0.0198 |
| Q | 1 | 1.02 | 0 | 1.078 | 8.6 | 0.26 | 0.73 | 0.05 | 0.8 | 0.0761 |
| R | 0.7 | 0.84 | 2.088 | 1.725 | 9.1 | 0.28 | 0.75 | 0.04 | 0.9 | 0.0959 |
| S | 1.7 | 1.47 | 1.089 | 1.145 | 7.5 | 0.55 | 1.24 | 0.11 | 0.8 | 0.0829 |
| T | 1.7 | 2.39 | 1.732 | 1.487 | 6.6 | 0.49 | 0.85 | 0.04 | 0.7 | 0.0941 |
| V | 0.6 | 3.5 | 0.946 | 0.625 | 5.6 | 0.08 | 0.43 | 0.01 | 0.6 | 0.0057 |
| W | 1 | 2.49 | 1.781 | 0.803 | 5.3 | 0.43 | 0.62 | 0 | 1.7 | 0.0548 |
| Y | 1 | 2.23 | 0 | 1.227 | 5.7 | 0.46 | 1.44 | 0.03 | 0.4 | 0.0516 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Residue（11-18）** | **Bull&Breese hydrophobicity** | **Grantham polarity** | **Zimmerman polarity** | **Kate & Doolittle hydrophobicity** | **isoelectric point** | **bulkiness** | **Eisenberg hydrophobicity** | **Hope & Woods hydrophobicity propensities** |
| A | 2.01 | 0.328 | 0.937 | 7 | 0.22 | 0.35 | 1.6 | 0.0373 |
| C | 1.98 | 0 | 1.004 | 5.5 | 0.2 | 0.5 | 1.2 | 0.0829 |
| D | -2.05 | 3.379 | 1.64 | 13 | 0.73 | 2.16 | 2.6 | 0.1263 |
| E | 0.93 | 0 | 0.679 | 12.5 | 0.08 | 0.65 | 2 | 0.0058 |
| F | 2.68 | 1.336 | 0.803 | 5 | 0.08 | 0.89 | 0.9 | 0.0946 |
| G | 0.12 | 0.5 | 0.901 | 7.9 | 0.58 | 2.4 | 0.9 | 0.005 |
| H | -0.14 | 1.204 | 1.085 | 8.4 | 0.14 | 1.19 | 0.7 | 0.0242 |
| I | 3.7 | 2.078 | 0.178 | 4.9 | 0.22 | 0.12 | 0.7 | 0 |
| K | 2.55 | 0.835 | 1.254 | 10.1 | 0.27 | 0.83 | 1 | 0.0371 |
| L | 2.73 | 0.414 | 0.808 | 4.9 | 0.19 | 0.58 | 0.3 | 0 |
| M | 1.75 | 0.982 | 0.886 | 5.3 | 0.38 | 0.22 | 1 | 0.0823 |
| N | 0.03 | 1.498 | 1.08 | 10 | 0.42 | 2.12 | 0.7 | 0.0036 |
| P | 0.41 | 0.415 | 0.748 | 6.6 | 0.46 | 0.43 | 0.5 | 0.0198 |
| Q | 1.02 | 0 | 1.078 | 8.6 | 0.26 | 0.73 | 0.8 | 0.0761 |
| R | 0.84 | 2.088 | 1.725 | 9.1 | 0.28 | 0.75 | 0.9 | 0.0959 |
| S | 1.47 | 1.089 | 1.145 | 7.5 | 0.55 | 1.24 | 0.8 | 0.0829 |
| T | 2.39 | 1.732 | 1.487 | 6.6 | 0.49 | 0.85 | 0.7 | 0.0941 |
| V | 3.5 | 0.946 | 0.625 | 5.6 | 0.08 | 0.43 | 0.6 | 0.0057 |
| W | 2.49 | 1.781 | 0.803 | 5.3 | 0.43 | 0.62 | 1.7 | 0.0548 |
| Y | 2.23 | 0 | 1.227 | 5.7 | 0.46 | 1.44 | 0.4 | 0.0516 |

**Text S1. Evaluation Methods**

ACC represents the ratio of correctly predicted samples to the total number of samples. Specifically, ACC can be defined as follows:

ACC =

where TP (True Positive) and TN (True Negative) represent the numbers of correctly predicted positive and negative instances, respectively, while FP (False Positive) and FN (False Negative) represent the numbers of incorrectly predicted positive and negative instances. A higher accuracy indicates a more precise prediction by the classifier.

The F1 score is an evaluation metric that harmonizes the balance between a model's precision (PRE) and sensitivity (recall, SEN). It is calculated by taking the harmonic mean of precision and sensitivity, effectively capturing the trade-off between them. The fundamental objective of the F1 score is to achieve an equilibrium where both precision and sensitivity are equally prioritized, thereby ensuring a more robust measure of a model's accuracy in classifying data correctly.

F1 =

The Matthews Correlation Coefficient (MCC) is a comprehensive evaluation metric that incorporates all quadrants of the confusion matrix: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). MCC provides a value between -1 and 1, where 1 indicates perfect prediction, 0 indicates no better than random prediction, and -1 indicates total disagreement between prediction and observation. This metric is particularly useful because it gives a high score only if the prediction succeeds in both classes of the binary classification (positive and negative), effectively capturing both the balance and the strength of the predictive performance across different classes.

MCC =

ROC reflects the probability of the model correctly distinguishing between positive and negative samples. It is calculated as the area under the ROC curve, where the horizontal axis represents the False Positive Rate (FPR) and the vertical axis represents the True Positive Rate (TPR). A larger ROC value indicates better performance of the model. FPR and TPR can be calculated as follows:

FPR =

TPR =

**Text S2. Data Dimensions: Discussion on the data dimension of two embedding features**

Different amino acid positions and relationships in protein sequences are complex, making it essential to use an appropriate feature dimension to capture and represent protein sequence and structural information. In the ProtBert model, which extracts 1024-dimensional features, we sought to validate this dimension for protein-RNA interface residue prediction. To explore the effectiveness of different dimensions, we conducted experiments applying Principal Component Analysis (PCA) to reduce the dimensionality from 1024 to 512. This step aimed to remove redundant and highly correlated features while retaining the most significant ones. Subsequently, we used an autoencoder to compress the data into a 2048-dimensional space, intended to reveal more complex patterns that might be missed in the original feature set. The three sets of data were then inputted into the same prediction model (Supplementary Table S3). The results indicate that reducing dimensions to 512 significantly decreases model accuracy, suggesting the loss of important features and patterns. Conversely, increasing dimensions to 2048 does not notably improve prediction accuracy and may introduce redundant information and noise, complicating model training and reducing efficiency.

In the graph embedding features, we transformed the data to 512 and 2048 dimensions and used a single channel network for prediction (Supplementary Table S3). The results indicate that 1024-dimensional features outperform both 512 and 2048 dimensions.

indent The experimental results show that 1024-dimensional features strike the best balance between capturing protein sequence and structural information. In contrast, 512-dimensional features oversimplify the data and lose critical details, while 2048-dimensional features add excessive noise and redundancy, impairing model performance. Thus, 1024-dimensional features offer the optimal trade-off between expressiveness and computational complexity.

**Table S3.** Comparison of different feature dimensions of Protein Sequence Embedding and Graph Embedding in single channel TCFFPre

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature dimension** | **Protein sequence embedding** | | | | | | | **Protein graph embedding** | | | | | | |
| **ACC** | **SEN** | **SPE** | **PRE** | **F1** | **MCC** | **AUC** | **ACC** | **SEN** | **SPE** | **PRE** | **F1** | **MCC** | **AUC** |
| 512 | 0.728 | 0.716 | 0.733 | 0.61 | 0.725 | 0.507 | 0.711 | 0.717 | 0.693 | 0.755 | 0.632 | 0.731 | 0.53 | 0.725 |
| 1024 | **0.805** | **0.752** | 0.81 | **0.692** | **0.801** | **0.596** | **0.802** | **0.768** | **0.747** | **0.788** | **0.674** | **0.786** | **0.574** | **0.767** |
| 2048 | 0.793 | 0.746 | **0.818** | 0.688 | 0.791 | 0.583 | 0.789 | 0.733 | 0.724 | 0.749 | 0.625 | 0.744 | 0.534 | 0.752 |