Supplementary Text S2: SVM parameter selection

In this supplementary text, we describe the procedure of selecting support vector machine (SVM) kernel parameters. In-order to run a grid of values for parameter tuning [[1](#_ENREF_1)], we need a set of features. Different features can lead to different parameters and different parameters can lead to different performance. To address this, we use successive feature selection scheme in the forward direction [[2](#_ENREF_2)] with default SVM parameters to first select reasonable number of features. Then, we use these features to run grid of values to find appropriate values of C and gamma. Below we describe the steps for selecting SVM parameters for models constructed using BigramMoRF and StructMoRF methods.

**Parameter tuning for model constructed using BigramMoRF method**

Using steps of BigramMoRF method, we use flank size of 8 amino acids to extract samples (flank size of 8 was selected for the purpose of computational processing). We perform 5-fold cross validation with successive feature selection scheme on training data with balanced sampling, this resulted in 6 structural attributes with 36 features. Using these features with 5-fold cross validation on training data, we run a grid of values as shown in TableS2-1. Cells in each grid hold the value of AUC. We selected the cell approximately central to the high AUC values.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| gamma | C | | | | |
| 100 | 500 | **1000** | 4000 | 8000 |
| 0.0005 | 0.639 | 0.655 | 0.662 | 0.682 | 0.696 |
| 0.0008 | 0.643 | 0.659 | 0.669 | 0.693 | 0.698 |
| 0.0011 | 0.648 | 0.665 | 0.677 | 0.696 | 0.701 |
| 0.0014 | 0.65 | 0.669 | 0.683 | 0.699 | 0.698 |
| 0.0017 | 0.654 | 0.671 | 0.689 | 0.701 | 0.706 |
| 0.002 | 0.657 | 0.677 | 0.689 | 0.704 | 0.707 |
| 0.0023 | 0.658 | 0.681 | 0.69 | 0.703 | 0.705 |
| 0.0026 | 0.662 | 0.682 | 0.696 | 0.7 | 0.709 |
| 0.0029 | 0.66 | 0.688 | 0.697 | 0.705 | 0.704 |
| 0.0032 | 0.663 | 0.691 | 0.697 | 0.702 | 0.709 |
| 0.0035 | 0.667 | 0.692 | 0.699 | 0.707 | 0.712 |
| **0.0038** | 0.668 | 0.69 | **0.713** | 0.708 | 0.712 |
| 0.0041 | 0.668 | 0.696 | 0.702 | 0.705 | 0.713 |
| 0.0044 | 0.671 | 0.694 | 0.699 | 0.711 | 0.709 |
| 0.0047 | 0.671 | 0.695 | 0.702 | 0.707 | 0.714 |
| 0.005 | 0.673 | 0.694 | 0.699 | 0.713 | 0.712 |
| 0.011 | 0.687 | 0.702 | 0.709 | 0.713 | 0.717 |
| 0.017 | 0.692 | 0.704 | 0.709 | 0.714 | 0.712 |
| 0.03 | 0.699 | 0.713 | 0.713 | 0.713 | 0.709 |
| 0.05 | 0.706 | 0.711 | 0.71 | 0.715 | 0.703 |
| 0.07 | 0.706 | 0.711 | 0.709 | 0.704 | 0.704 |
| 0.1 | 0.710 | 0.716 | 0.712 | 0.703 | 0.695 |
| 0.3 | 0.714 | 0.707 | 0.695 | 0.687 | 0.685 |
| 0.5 | 0.711 | 0.697 | 0.684 | 0.686 | 0.68 |
| 0.7 | 0.705 | 0.689 | 0.694 | 0.682 | 0.67 |
| 0.9 | 0.701 | 0.69 | 0.692 | 0.674 | 0.669 |
| 1 | 0.698 | 0.683 | 0.679 | 0.674 | 0.655 |
| 2 | 0.696 | 0.672 | 0.662 | 0.653 | 0.65 |
| 3 | 0.682 | 0.659 | 0.662 | 0.647 | 0.647 |
| 4 | 0.678 | 0.66 | 0.662 | 0.651 | 0.653 |
| 5 | 0.679 | 0.668 | 0.664 | 0.657 | 0.651 |

TableS2-1. Grid using 36 selected features. Each row starts with its gamma value and each column starts with C value. Each cell hold the value of AUC and are divided into two regions, light orange cells for higher AUC values and light blue cells for lower AUC values. The selected parameters are in bold.

**Parameter tuning for model constructed using StructMoRFmethod**

Using steps of StructMoRF method with flank size of 8 amino acids, we use feature selection scheme with balanced sampling and 5-fold cross validation on training data to obtain 51 features. With these features we perform 5-fold cross validation on training data to run a grid of values as shown in Table S2-2 (a).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| a) | gamma | C | | | | | b) | gamma | C | | | | |
|  | 100 | 500 | **1000** | 4000 | 8000 |  | 100 | 500 | **1000** | 4000 | 8000 |
|  | 0.0005 | 0.6372 | 0.6562 | 0.6641 | 0.6773 | 0.6805 |  | 0.0005 | 0.6788 | 0.6814 | 0.678 | 0.6689 | 0.6663 |
|  | 0.0008 | 0.6448 | 0.6664 | 0.6753 | 0.6802 | 0.6818 |  | 0.0008 | 0.6811 | 0.6778 | 0.6732 | 0.6647 | 0.6634 |
|  | 0.0011 | 0.6536 | 0.6724 | 0.6787 | 0.6835 | 0.6847 |  | 0.0011 | 0.6821 | 0.674 | 0.6692 | 0.6634 | 0.6633 |
|  | 0.0014 | 0.6587 | 0.6783 | 0.6812 | 0.6841 | 0.686 |  | 0.0014 | 0.6818 | 0.6705 | 0.6672 | 0.6626 | 0.6627 |
|  | 0.0017 | 0.6647 | 0.6797 | 0.6797 | 0.6842 | 0.6855 |  | 0.0017 | 0.6806 | 0.6684 | 0.6648 | 0.6626 | 0.6633 |
|  | 0.002 | 0.6683 | 0.6786 | 0.6829 | 0.6845 | 0.684 |  | 0.002 | 0.6785 | 0.6672 | 0.664 | 0.663 | 0.6633 |
|  | 0.0023 | 0.6716 | 0.6804 | 0.6825 | 0.6847 | 0.6854 |  | 0.0023 | 0.6769 | 0.6657 | 0.6637 | 0.6634 | 0.6629 |
|  | 0.0026 | 0.6739 | 0.6823 | 0.684 | 0.684 | 0.6849 |  | 0.0026 | 0.6749 | 0.6647 | 0.6634 | 0.6629 | 0.6631 |
|  | 0.0029 | 0.676 | 0.6836 | 0.6852 | 0.6853 | 0.6849 |  | 0.0029 | 0.6737 | 0.6641 | 0.6632 | 0.6629 | 0.6634 |
|  | 0.0032 | 0.6776 | 0.6828 | 0.6854 | 0.6846 | 0.6843 |  | 0.0032 | 0.6718 | 0.664 | 0.6626 | 0.663 | 0.6643 |
|  | 0.0035 | 0.6784 | 0.6845 | 0.6855 | 0.686 | 0.6848 |  | 0.0035 | 0.6707 | 0.6634 | 0.6625 | 0.6631 | 0.6633 |
|  | **0.0038** | 0.678 | 0.6847 | **0.6843** | 0.6872 | 0.6866 |  | **0.0038** | 0.6694 | 0.6633 | **0.6625** | 0.6636 | 0.6634 |
|  | 0.0041 | 0.6794 | 0.6823 | 0.6857 | 0.6856 | 0.6853 |  | 0.0041 | 0.6686 | 0.663 | 0.6628 | 0.6636 | 0.6639 |
|  | 0.0044 | 0.6781 | 0.6848 | 0.6858 | 0.6861 | 0.685 |  | 0.0044 | 0.6684 | 0.6629 | 0.6633 | 0.6634 | 0.6635 |
|  | 0.005 | 0.6811 | 0.6845 | 0.6865 | 0.6866 | 0.6874 |  | 0.005 | 0.6663 | 0.6623 | 0.6628 | 0.6634 | 0.663 |
|  | 0.011 | 0.6855 | 0.6868 | 0.6884 | 0.6928 | 0.6921 |  | 0.011 | 0.6627 | 0.6621 | 0.6621 | 0.6598 | 0.6587 |
|  | 0.017 | 0.687 | 0.6895 | 0.6955 | 0.6968 | 0.6969 |  | 0.017 | 0.662 | 0.6613 | 0.6616 | 0.6562 | 0.653 |
|  | 0.02 | 0.6887 | 0.6911 | 0.6948 | 0.6974 | 0.6989 |  | 0.02 | 0.6622 | 0.6609 | 0.6601 | 0.6535 | 0.6487 |
|  | 0.06 | 0.7 | 0.704 | 0.7074 | 0.7113 | 0.7139 |  | 0.06 | 0.6586 | 0.6496 | 0.6421 | 0.6252 | 0.6206 |
|  | 0.1 | 0.7055 | 0.7122 | 0.7163 | 0.7202 | 0.7145 |  | 0.1 | 0.6518 | 0.6356 | 0.6283 | 0.6149 | 0.6132 |
|  | 0.3 | 0.7319 | 0.7347 | 0.7268 | 0.7143 | 0.7154 |  | 0.3 | 0.6277 | 0.6168 | 0.6143 | 0.604 | 0.6001 |
|  | 0.5 | 0.7427 | 0.734 | 0.7301 | 0.7229 | 0.7112 |  | 0.5 | 0.6217 | 0.609 | 0.6031 | 0.5955 | 0.5963 |
|  | 0.7 | 0.747 | 0.7373 | 0.7328 | 0.7211 | 0.712 |  | 0.7 | 0.6169 | 0.6012 | 0.5979 | 0.5981 | 0.5982 |
|  | 0.9 | 0.7494 | 0.7371 | 0.7355 | 0.7215 | 0.7128 |  | 0.9 | 0.6099 | 0.5989 | 0.5969 | 0.5983 | 0.5972 |
|  | 1 | 0.7528 | 0.7401 | 0.7285 | 0.7253 | 0.7129 |  | 1 | 0.6081 | 0.5982 | 0.5974 | 0.5976 | 0.595 |
|  | 2 | 0.7602 | 0.7509 | 0.7372 | 0.7333 | 0.7248 |  | 2 | 0.595 | 0.5961 | 0.5961 | 0.5844 | 0.5819 |
|  | 3 | 0.7621 | 0.751 | 0.7513 | 0.7389 | 0.7382 |  | 3 | 0.592 | 0.5924 | 0.5888 | 0.5789 | 0.5776 |
|  | 4 | 0.7685 | 0.7559 | 0.7552 | 0.744 | 0.7506 |  | 4 | 0.592 | 0.5884 | 0.5833 | 0.5786 | 0.5787 |
|  | 5 | 0.7725 | 0.7618 | 0.7552 | 0.7503 | 0.7555 |  | 5 | 0.591 | 0.5862 | 0.5833 | 0.5811 | 0.5811 |

Table S2-2 (a) Grid using 5-fold cross validation on training. (b) Grid for independent test using training data to train and test data for scoring. (Grid using 51 selected features. Each row starts with its gamma value and each column starts with C value. Each cell hold the value of AUC and are divided into three regions, light orange cells for higher AUC values, gray cells with mid-range AUC values and light blue cells for lower AUC values.)

**Selected kernel parameters**

To see the performance in scoring test sequence using the higher AUC value cell parameters, we use the same feature sets to train each model using training data and we test the model using test data. For model constructed using BigramMoRF method, the higher cell parameters of Table S2-1 gave approximately good performance in scoring test sequence compared to other cell parameters. However, for model constructed using StructMoRF method, when higher AUC value cell parameters of Table S2-2(a) are used for scoring test sequence, the AUCs obtained are lower. This is observed in Table S2-2(b), which is a strong indication of over-fitting. It is also noted that when lower and mid-range AUC value cell parameters of Table S2-2(a) are used for scoring test sequence, the AUCs are little higher.

To construct model using BigramMoRF method, we selected C and gamma values as 1000 and 0.0038, respectively. The selected parameter cell was approximately central to the high AUC values in Table S2-1. And to construct model using StructMoRF method, we have again selected C and gamma values as 1000 and 0.0038, respectively. We have selected this cell, since it is approximately in the central of higher values of AUCs in scoring test as observed in Table S2-2(b) and also AUC values are in mid-range in Table S2-2(a). Using the selected SVM parameters, the attributes and features are then selected to develop the proposed predictor.

[1] Chang, C.C. and Lin, C.J., LIBSVM : a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, 2011; **2**, 1-27.

[2] Sharma, A., Paliwal, K.K., Dehzangi, A., Lyons, J., Imoto, S. and Miyano, S., A Strategy to Select Suitable Physicochemical Attributes of Amino Acids for Protein Fold Recognition, *BMC Bioinformatics*, 2013; **14**, 1-11.