

Appendix I: Parameter selection

Qidong Zhang

Parameter selection for the SVM prediction system

When training SVM, we need to select a proper kernel function. There are four typical SVM kernel functions: linear, polynomial, radial basis function (RBF) and sigmoid. Here radial basis function (RBF) is used, which is given by:

$$K(\bar{x}_i, \bar{x}_j) = \exp(-\gamma|\bar{x}_i - \bar{x}_j|^2)$$

There is one parameter γ in this equation. A series of kernel parameters γ and the regularization parameter C are tested. The results are summarized in Table A1~A2. For the limited computational power, we use 365 protein chains as training set and 61 protein chains as testing set to select the optimal kernel parameters and regularization parameter for the prediction system. Since our dataset contains natural occurring percentage of β -turn residues which comprise about 25% of all protein residues, a cost-factor is used to overweight errors of positive examples. Here the cost factor is set to $j = 2$.

Table A1. Performance comparisons for various values of regularization parameter C . Cost factor is set to $j = 2$.

C	$Q_{\text{total}} (\%)$	$Q_{\text{predicted}} (\%)$	$Q_{\text{observed}} (\%)$	MCC
4	70.0	43.1	76.0	0.382
8	71.1	44.1	74.9	0.389
16*	73.5	46.8	71.5	0.405
32	74.7	48.2	67.6	0.404
64	75.0	48.6	64.7	0.394
128	75.3	49.0	63.4	0.392
256	75.2	48.9	60.6	0.378
512	74.7	47.8	55.8	0.347

Table A2. Performance comparisons when various values of γ in the RBF kernel are adopted. Regularization parameter is set to $C = 16$. Cost factor $j = 2$.

γ	Q_{total} (%)	$Q_{\text{predicted}}$ (%)	Q_{observed} (%)	MCC
0.0039	70.0	43.0	76.1	0.382
0.0078	70.0	43.1	76.0	0.382
0.0156	73.5	46.8	71.5	0.405
0.0186*	74.4	47.8	70.2	0.411
0.0313	75.4	49.1	64.6	0.399
0.0625	75.3	48.8	52.9	0.343
0.125	76.3	51.0	36.2	0.286
0.25	77.3	54.7	33.3	0.297