Assessing the Performance of Macromolecular Sequence Classifiers

Cornelia Caragea, Jivko Sinapov, Michael Terribilini, Drena Dobbs, and Vasant Honavar

Machine learning approaches offer some of the most cost-effective approaches to building predictive models (e.g., classifiers) in a broad range of applications in computational biology, e.g., given an amino acid sequence, identifying the amino acid residues that are likely to bind to RNA. Comparing the effectiveness of different algorithms requires reliable procedures for accurately assessing the performance (e.g., accuracy, sensitivity, and specificity) of the resulting predictive classifiers. There are, broadly speaking, two approaches to evaluating the performance of such classifiers: *sequence-based* cross-validation and *window-based* cross-validation. In the former, the training and test data typically correspond to disjoint sets of sequences. In the latter, they typically correspond to disjoint sets of sequence windows.

We compare *sequence-based* and *window-based cross-validation* procedures on four representative sequence-based prediction tasks: identifying glycosylation sites, protein-protein interface residues, protein-RNA interface residues, and secondary structure from amino acid sequence. Our experiments with two representative classifiers (Naive Bayes [1] and Support Vector Machines [2]) show that *sequence-based* and *window-based cross-validation* procedures and *data selection* methods can yield different estimates of commonly used performance measures such as Accuracy (Acc), Matthews Correlation Coefficient (MCC) and Area under the Receiver Operating Characteristic curve (AUC) defined in [3] (see Table 1). Our results suggest that *window-based cross-validation* can significantly over-estimate the performance of the classifiers under certain conditions. On tasks that require labeling residues in a novel macromolecular sequence (as in the case of identifying RNA binding residues from an amino acid sequence), we believe that the estimates obtained using *sequence-based cross-validation* provide more natural estimates of performance than those obtained using *window-based cross-validation*.

Classifier/ PerfMeasure	**SVM-WinCV**	**SVM-SeqCV**	**NB-WinCV**	**NB-SeqCV**
Acc	**0.94**	0.89	**0.90**	0.89
MCC	**0.77**	0.56	**0.60**	0.56
AUC	**0.94**	0.89	**0.91**	0.88

Table 1. Experimental results for window-based cross-validation (**WinCV**), and sequence-based cross-validation (**SeqCV**) for the glycosylation dataset using Support Vector Machine (**SVM**) and Naïve Bayes (**NB**) classifiers.

[1] T.M.Mitchell. *Machine Learning*. McGraw Hill, 1997.

[2] V.Vapnik. *Statistical Learning Theory*. Springer-Verlag, New-York,1998.

[3] P.Baldi, S. Brunak, Y. Chauvin, C. Andersen, and H. Nielsen. *Assessing the Accuracy of Prediction Algorithms for classification: an overview*. Bioinformatics, 16(5):412-424, 2000.