

# *f*RMSDPred: Predicting local rmsd between structural fragments using sequence information (extra tables)

Huzefa Rangwala and George Karypis

Computer Science & Engineering, University of Minnesota, Minneapolis, MN 55455  
rangwala@cs.umn.edu, karypis@cs.umn.edu

Table ?? optimizes the weight parameter on the individual kernels integrated for the fusion kernel. It also shows why using equal weighting on the individual kernels was a valid choice for the paper.

Table 1: Classification performance of the fusion kernels for the hard datasets optimizing the weighting on pssm part

Scheme	weight	<i>all</i>		<i>fam</i>		<i>suf</i>		<i>fold</i>	
		<i>ROC</i> <sub>5</sub>	<i>ROC</i>	<i>ROC</i> <sub>5</sub>	<i>ROC</i>	<i>ROC</i> <sub>5</sub>	<i>ROC</i>	<i>ROC</i> <sub>5</sub>	<i>ROC</i>
$(\mathcal{P}+\mathcal{S})^{conc} -all$	0.0	0.756	0.894	0.732	0.864	0.753	0.915	0.797	0.914
	0.1	0.775	0.902	0.754	0.876	0.772	0.921	0.809	0.919
	0.2	0.781	0.904	0.760	0.878	0.779	0.923	0.814	0.920
	0.3	0.785	0.906	0.764	0.880	0.784	0.925	0.816	0.921
	0.4	0.788	0.907	0.766	0.882	0.789	0.927	0.819	0.922
	0.5	0.790	0.908	0.768	0.883	0.791	0.927	0.820	0.923
	0.6	0.791	0.908	0.770	0.885	0.793	0.926	0.821	0.924
	0.7	0.791	0.908	0.770	0.885	0.791	0.925	0.821	0.924
	0.8	0.790	0.908	0.769	0.884	0.791	0.924	0.818	0.924
	0.9	0.787	0.907	0.768	0.884	0.789	0.922	0.813	0.923
1.0	0.705	0.872	0.729	0.866	0.690	0.883	0.688	0.867	
$(\mathcal{P}+\mathcal{S})^{pair} -all$	0.0	0.678	0.852	0.627	0.806	0.682	0.883	0.745	0.886
	0.1	0.756	0.889	0.741	0.866	0.751	0.902	0.783	0.906
	0.2	0.761	0.891	0.746	0.869	0.759	0.905	0.785	0.908
	0.3	0.764	0.893	0.748	0.870	0.760	0.907	0.791	0.911
	0.4	0.765	0.893	0.749	0.870	0.760	0.906	0.792	0.912
	0.5	0.766	0.893	0.750	0.871	0.761	0.907	0.793	0.912
	0.6	0.767	0.894	0.750	0.871	0.761	0.906	0.797	0.913
	0.7	0.766	0.894	0.750	0.871	0.760	0.907	0.797	0.914
	0.8	0.764	0.893	0.747	0.869	0.757	0.906	0.797	0.914
	0.9	0.763	0.892	0.745	0.868	0.756	0.906	0.794	0.912
1.0	0.665	0.851	0.700	0.850	0.627	0.857	0.657	0.847	

The test and training set consisted of proteins from the *all*, *fam*, *suf*, and *fold* sets. The numbers in parentheses for the profile-to-profile scoring schemes indicate the value of  $w$  for the  $wmers$  that were used. The numbers in bold show the best performing schemes for the kernel-based and profile-to-profile scoring based schemes. The underlined results show the cases where the pairwise coding scheme performs better than the concatenation coding scheme.

Table ?? shows the regression performance on a dataset using residue-pairs from pairs of proteins randomly rather than aligned positions. We also show good root mean squared error performance along with correlation coefficient performance.

Table 2: Regression Performance of the fusion kernels on the dataset for a random set of positions

Scheme	<i>all</i>			<i>fam</i>			<i>suf</i>			<i>fold</i>		
	<i>CC</i>	<i>rmse</i>	<i>dev</i>	<i>CC</i>	<i>rmse</i>	<i>dev</i>	<i>CC</i>	<i>rmse</i>	<i>dev</i>	<i>CC</i>	<i>rmse</i>	<i>dev</i>
$\mathcal{PF}_{pic} + \mathcal{S}_{dotp}(3)$	-0.450	21.461	26.746	-0.442	21.945	25.505	-0.445	21.304	24.280	-0.464	21.088	30.980
$(\mathcal{P}+\mathcal{S})^{conc} -fam$	<b>0.609</b>	<b>0.727</b>	<b>0.859</b>	<b>0.587</b>	<b>0.754</b>	<b>0.876</b>	<b>0.625</b>	<b>0.709</b>	<b>0.799</b>	<b>0.617</b>	<b>0.713</b>	<b>0.904</b>
$(\mathcal{P}+\mathcal{S})^{conc} -suf$	<b>0.613</b>	<b>0.727</b>	<b>0.861</b>	<b>0.602</b>	<b>0.750</b>	<b>0.868</b>	<b>0.615</b>	<b>0.716</b>	<b>0.808</b>	<b>0.623</b>	<b>0.713</b>	<b>0.911</b>
$(\mathcal{P}+\mathcal{S})^{conc} -fold$	<b>0.611</b>	<b>0.728</b>	<b>0.857</b>	<b>0.594</b>	<b>0.753</b>	<b>0.876</b>	<b>0.623</b>	<b>0.714</b>	<b>0.801</b>	<b>0.619</b>	<b>0.714</b>	<b>0.896</b>
$(\mathcal{P}+\mathcal{S})^{conc} -all$	<b>0.630</b>	<b>0.716</b>	<b>0.840</b>	<b>0.617</b>	<b>0.741</b>	<b>0.859</b>	<b>0.643</b>	<b>0.699</b>	<b>0.788</b>	<b>0.631</b>	<b>0.702</b>	<b>0.874</b>
$(\mathcal{P}+\mathcal{S})^{pair} -fam$	<u>0.567</u>	<u>0.759</u>	<u>0.911</u>	<u>0.538</u>	<u>0.785</u>	<u>0.911</u>	<u>0.568</u>	<u>0.750</u>	<u>0.842</u>	<u>0.597</u>	<u>0.739</u>	<u>0.987</u>
$(\mathcal{P}+\mathcal{S})^{pair} -suf$	<u>0.569</u>	<u>0.761</u>	<u>0.895</u>	<u>0.545</u>	<u>0.784</u>	<u>0.906</u>	<u>0.573</u>	<u>0.751</u>	<u>0.843</u>	<u>0.593</u>	<u>0.743</u>	<u>0.939</u>
$(\mathcal{P}+\mathcal{S})^{pair} -fold$	<u>0.577</u>	<u>0.761</u>	<u>0.911</u>	<u>0.547</u>	<u>0.783</u>	<u>0.903</u>	<u>0.585</u>	<u>0.753</u>	<u>0.846</u>	<u>0.600</u>	<u>0.743</u>	<u>0.992</u>
$(\mathcal{P}+\mathcal{S})^{pair} -all$	<u>0.588</u>	<u>0.748</u>	<u>0.898</u>	<u>0.563</u>	<u>0.772</u>	<u>0.894</u>	<u>0.591</u>	<u>0.741</u>	<u>0.831</u>	<u>0.614</u>	<u>0.727</u>	<u>0.977</u>

The test and training set consisted of proteins from the *all*, *fam*, *suf*, and *fold* sets. The number in parentheses for the profile-to-profile scoring scheme indicates the value of  $w$  for the  $w_{MER}$  that was used. Good correlation coefficient values will be negative for the profile-to-profile scoring scheme and positive for the kernel-based schemes. The numbers in bold show the best performing schemes. The underlined results show the cases where the pairwise coding scheme performs better than the concatenation coding scheme.