Predicting local rmsd between structural fragments using sequence information

Huzefa Rangwala and George Karypis

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

Structure alignment methods score a pair of residues by considering how well fixed-length fragments (i.e., short contiguous backbone segments) centered around each residue align with each other. This score is usually computed as the root mean squared deviation (RMSD) of the optimal superimposition of the two fragments. In this work, we focus on the problem of estimating the RMSD value of a pair of protein fragments by considering only sequence-derived information.

We use a supervised learning framework based on support vector regression and classification for solving the fragment-level RMSD prediction problem. The key contributions of this work include the problem formulation, sequence information encoding, development of novel second-order pairwise exponential kernel functions designed to capture the conserved signals of a pair of local subsequences centered at each of the residues, and use of a fusion-kernel-based approach to use both profile and predicted secondary structure based information.

We perform an extensive experimental evaluation of the algorithms and their parameter space using a dataset of residue-pairs derived from optimal sequence-based local alignments of known protein structures. Our experimental results show that there is a high correlation (0.681 – 0.768) between the estimated and actual fragment-level RMSD scores. Moreover, the performance of our algorithms is considerably better than that obtained by state-of-the-art profile-to-profile scoring schemes when used to solve the fragment-level RMSD prediction problems by about 25%. We also show an improvement in predicting the reliability of residue-pairs dependent on the fragment-level RMSD scores. In this case, we use support vector classification, and evaluate the performance based on ROC scores.

Figures 1 and 2 plot the actual fragment-level RMSD scores against the estimated scores using the support vector regression method and the actual fragment-level RMSD scores against the optimized profile-profile scoring schemes respectively. The color coding used in the figures represents the number of points plotted in a fixed area, and helps us show the superior estimation performance of our algorithm. Notice, in Figure 2 how the plotted points are more scattered and less dense compared to Figure 1.

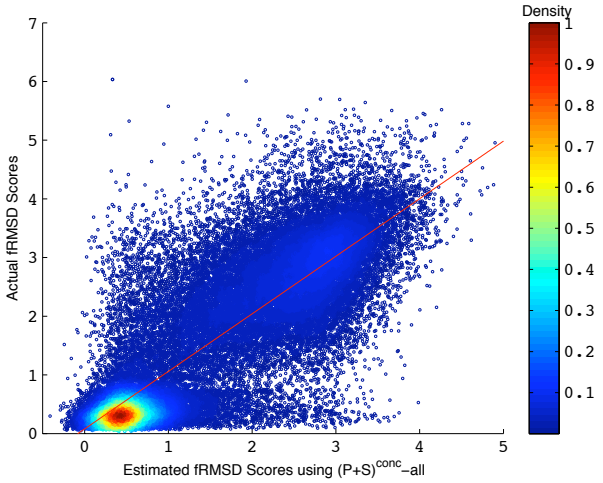

Figure 1: Scatter plot for test protein-pairs at all levels between estimated and actual fragment-level RMSD scores.

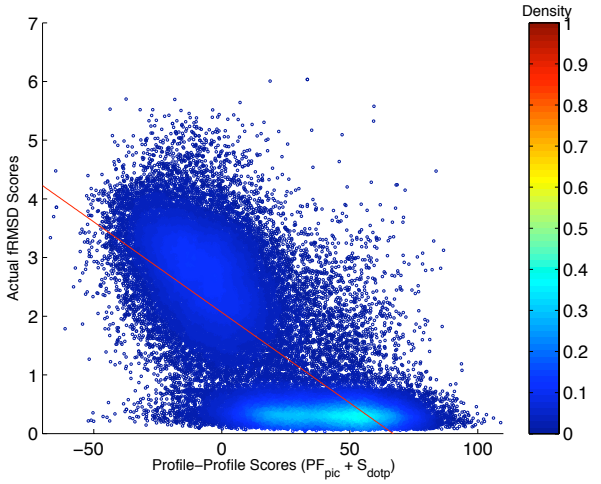

Figure 2: Scatter plot for test protein-pairs at all levels between profile-to-profile and actual fragment-level RMSD scores.

Accurate prediction of the fragment-level RMSD scores has four major applications to protein structure prediction: i) sequence-structure alignment quality assessment, and selection of candidate alignments for comparative modeling, ii) analyze different sequence-structure alignments in order to identify high-quality moderate length fragments to be used in fragment assembly based protein structure prediction methods like ROSETTA, iii) construction of position-to-position specific scoring matrix between all pairs of residues in a pair of proteins to be used for generating sequence alignment, and iv) additional information for other prediction tasks such as remote homology detection and fold recognition.