Making intrinsic disorder prediction practical with quality assessment

Gang Hu[1] , Zhonghua Wu[1], Christopher Oldfield[2], Chen Wang[2], **Lukasz Kurgan**[2,*]

[1]School of Mathematical Sciences and LPMC, Nankai University, Tianjin, 300071, PR China

[2]Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA.

Abstract:
Putative intrinsic disorder is widely used in computational and experimental studies. Perhaps surprisingly, these predictions do not include quality assessment (QA) scores that quantify correctness/confidence at a residue level. In essence, QA scores reveal which predictions are more likely to be correct. To date there are no QA predictors for the intrinsic disorder, which is in a stark contrast to the QA for putative tertiary structures that has been intensely researched. The complicating factor is that QA scores must be optimized for specific predictors of disorder since these methods rely on different types of disorder annotations and have widely different predictive architectures.

We introduce first-of-its-kind toolbox of methods that provide accurate QA scores for ten popular disorder predictors. The QUARTER (QUality Assessment for pRotein inTrinsic disordEr pRedictions) tool is optimized for each of the ten disorder predictors, guided by an empirical feature selection and disorder predictor-specific putative propensities for disorder. Empirical tests on a large test dataset reveal that QUARTER generates high quality scores which significantly outperform the disorder propensities output by the original predictors. QUARTER is available as a convenient webserver at http://biomine.cs.vcu.edu/servers/QUARTER/.

The main application of the QA scores produced by QUARTER is annotation/selection of a high-quality subset of residue-level disorder predictions. We show that when combining results from the ten disorder predictors, the QA scores can be used to find 40% of residues for which disorder is predicted with 95% precision. We also present results of analysis of high-quality disorder predictions in the human proteome that were generated by combining the ten predictions and their corresponding QA scores.