Comparing and combining predictors of mostly disordered proteins:

Supplementary Information

Christopher J. Oldfield‡,§, Yugong Cheng‡, Mark S. Cortese‡,#, Celeste J. Brown¶, Vladimir N. Uversky‡,#,║,* and A. Keith Dunker‡,#,*

‡ Molecular Kinetics, Inc. 6201 La Pas Trail, Suite 160, Indianapolis, IN 46268, USA; # Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, School of Medicine, Indiana University – Purdue University at Indianapolis, Indianapolis, IN 46202; ¶ Department of Biological Sciences, University of Idaho, Moscow, Idaho 83844, USA; ║ Institute for Biological Instrumentation, Russian Academy of Sciences, 142292 Pushchino, Moscow Region, Russia; § Present Address: Biophysics Department, Institute for Molecular Virology 1525 Linden Drive, Madison, WI 53706
**Receiver Operator Characteristics (ROC) Curves.** ROC curves are plots of the true positive rate (e.g., fraction of correct disorder predictions) as a function of the false positive rate (e.g., fraction of incorrect disordered predictions) generated by varying the prediction threshold. Empirical ROC curves have been generated for the prediction methods with discrete outputs, the consensus score and CDF predictions, and kernel density estimation-based ROC curves\(^{(S1)}\) have been generated for charge-hydropathy prediction. Two sets of curves were generated, one using wholly ordered proteins as the negative set (Fig. S1A) and the other using partially ordered proteins as the negative set (Fig. S2A). Both sets of curves use the wholly disordered proteins as the positive set.

A note should be given on the correspondence of this data to the Tables in the text. Tables I and II in the text are not equivalent to Fig S1. The ROC curve values are generated using a sliding threshold, but the values reported in the tables are generated using prediction margins, where only proteins with prediction values falling a certain distance from the boundary are predicted and proteins with prediction values closer to the boundary are not predicted. The data in Table III corresponds to a single point on each of the curves in Fig S1.

**Wholly ordered proteins** (PDB identification and chain identification, if applicable):

1A0K, 1A3H, 1A8Q, 1AA2, 1AD6, 1AIE, 1AKO, 1AKZ, 1AM5, 1AVO, 1AR2, 1B9K A, 1BD8, 1BEE, 1BFS, 1BG5, 1BK2, 1BLE, 1BM8, 1BU2 A, 1BYW A, 1BZ4 A, 1C25, 1C3G A, 1C44 A, 1C93 A, 1CDZ A, 1CM3 A, 1CQY A, 1CWY A, 1D2K A, 1D2P A, 1DAB A, 1DC9A, 1DT4 A, 1DVO A, 1DYW A, 1EG3 A, 1EH1 A, 1EM7 A, 1EOV A, 1EQ6 A, 1EW4 A, 1EYH A, 1F00 I, 1FNF, 1FSF A, 1FSO A, 1FTS, 1G5Z A, 1G62 A, 1G90 A, 1GC7 A, 1GXQ A, 1GYV A, 1H6T A, 1HH0 A, 1HH2 P, 1HKA, 1HXU A, 1I2T A, 1IS1 A, 1ICX A, 1IFP, 1ILW A, 1I02 A, 1IS5 A, 1JEJ A, 1JLN A, 1JW4 A, 1K6K A, 1KAS, 1L2P A, 1LA1 A, 1LCL, 1LP3 A, 1LPL A, 1MFO A, 1OPS, 1PBA, 1PZC, 1QAU A, 1QCQ A, 1QQH A, 1Q59 A, 1IQTOA, 1SKF, 1SSL, 1TVY, 1UIA, 1UOK, 1VSM A, 1WAL A, 1XGO, 2ALD A, 2ARA, 2BBV A, 2ERA, 2GPR, 2PLC, 2PHT, 2TMY, 6RHN, 8OHH, 9ILB A

**Partially ordered proteins** (PDB identification and chain identification, if applicable, followed by the range of residues with missing electron density):

Wholly disordered proteins (with references):

- P06454(S2), P02315(S3), P07746(S3), P10636(S4), O76070(S5), Q13541(S6), P17639(S7), AAC48720(S8), O06022(S9), AAC48700(S8), Q9XZH1(S10), P02387(S11), P02375(S11), P02427(S11), P02367(S11), P02374(S11), P02379(S11), P02432(S11), P02435(S11), P02436(S11), P02437(S11), P02662(S12), P04324(S13), P19103(S14), P12272(S15), P02687(S16), P38936(S17), P22239(S18), P10622(S19), 20, 21, P03045(S22), 23, P00006(S24), P04554(S25), P81455(S26), AAF13399(S27), P43532(S28), P04541(S29), P55326(S30), P01050(S31), P15340(S32), P02845(S33), P32499(S34), P21815(S35), P10451(S35), P19429(S36), P13795(S37), 38, Q9Z2E4(S39), P16949(S40), P03671(S41), P17096(S42), 43, 44, 2SOB(S45), P62803(S46), Q62100(S47), Q13541(S48), P42692(S49).

FIGURE LEGENDS

**Figure S1.** ROC curves for the consensus score (red line), CDF predictions (blue line), and charge-hydropathy prediction (black line). Marked points on the lines of the consensus score and CDF predictions correspond to the discrete prediction thresholds and only these points are meaningful. Charge-hydropathy predictions are continuous, so the entire line is meaningful. ROC curves where generated using the wholly disordered proteins as a positive set and either (A) the wholly ordered proteins or (B) the partially ordered proteins as a negative set.

REFERENCES


Figure S1.