Figure S1. Discovering evolutionary signal for conformations of flexible and disordered proteins

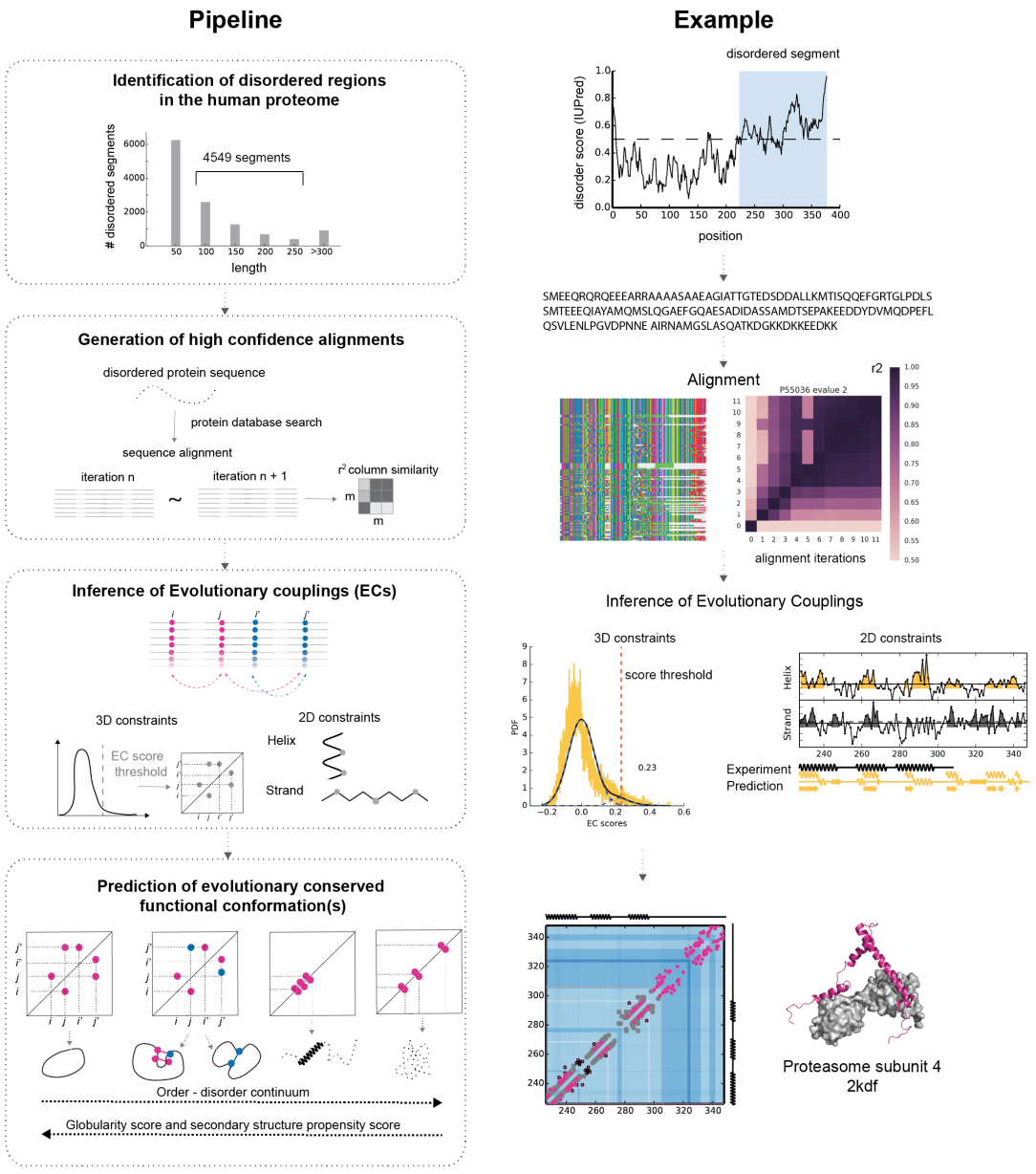

Figure S1. Discovering evolutionary signal for conformations of flexible and disordered proteins.

Pipeline: We applied the evolutionary couplings method on ~4500 100-300 residues long disordered regions of the human proteome. First we created high-quality alignments and judged the number of sequences and the robustness of the alignment. The alignment robustness score represents the agreement of the amino acid composition of the alignment columns after different rounds of re-alignment iterations. Then we applied a maximum entropy model to identify evolutionarily coupled pairs of columns in the alignments as described previously (Marks et al., 2011). We inferred the parameters of our model using penalized Maximum Likelihood with a pseudo-likelihood approximation (Ekeberg et al., 2013; Hopf et al., 2014) and excluded gap states from the calculation of the likelihoods (plmC, code available upon request). Then we assessed the significance of ECs based on a statistical model of scores. We automated the detection of significant EC pairs using a mixture model distribution providing consistency across all proteins. Using local ECs, we calculated the propensity for α-helical and β-strand secondary structure elements (Experimental Procedures). Based on the predicted 2D and 3D constraints, we proposed the structural constraints of a protein and predicted the residue-level secondary structure propensities and long-range residue-residue contacts. We can determine whether there is evolutionary signal for ordered states.

Example: Proteasome subunit 4 (PTM4_HUMAN). We define disordered regions using a sequence-based predictor, IUPred. First, we searched Uniprot for homologous proteins and created alignments. Then we tested the robustness of the alignment after different numbers of re-alignment iterations. If the alignment converged after 9 to 11 iterations, we proceeded with the evolutionary coupling calculations. We fit the distribution of the ECs with a Gaussian-lognormal mixture model, and determined a significance threshold. We applied the novel secondary structure propensity score to predict helices and strands along the sequence. We judged our prediction against known experimental structures if available.