PPI-hotspot^{ID} summary

PPI-Hotspot^{ID} identifies PPI-hot spots using an ensemble of classifiers and only 4 residue features (conservation, aa type, solvent-accessible surface area (SASA), and gas-phase energy ΔG^{gas}).

Per-Residue Conservation Score. To calculate the conservation score, k_i^C, of residue *i* in a protein, we implemented a method similar to ConSurf[1,2] to run in parallel with the energy evaluation code. First, we searched the UNIREF-90 database[3] using HMMER[4] to find sequences similar to the target sequence. Near-duplicates were removed by clustering matched sequences with ≥95% pairwise sequence identity using CD-hit[5] and keeping only one representative. Since HMMER[4] may only find good matches for a small proportion of the target sequence, we compared the HMMER sequences with the target sequence. We kept only those with > 60% overlap with the target sequence, and discarded sequences that were dissimilar (≤ 35% sequence identity) or nearly identical (≥ 95% sequence identity). Next, we pairwise aligned the remaining sequences, and if two sequences overlapped by > 10% of the sequence, we rejected the shorter sequence. After this filtering process, the resulting HMMER hits were used or if the number of hits exceeded 300, we selected the top 300 hits. These sequences were then aligned to the target sequence using MAFFT-LINSi.[6] We then used the Rate4Site program[7] to compute position-specific evolutionary rates from the generated multiple sequence alignment. These rates were normalized and grouped into ConSurf grades ranging from 1 to 9, where $k^C = 1$ represents the most rapidly evolving residues, and $k^C = 9$ indicates the most conserved residues.

Per-Residue Free Energy Contributions. For a given free protein structure, the Reduce program[8] was used to add hydrogens and assign the protonation states

of ionizable residues. Additional missing heavy and hydrogen atoms were added using the AmberTools version 20[9] and the Amber FF19SB forcefield.[10] To eliminate any steric clashes, we performed a conjugate gradients minimization with constraints on the heavy atoms using the Generalized Born model for 500 steps. The resulting structure was used to compute the per-residue energy/free energy contributions using the MMPBSA (Molecular Mechanics Poisson-Boltzmann Surface Area) module in AmberTools.[9] For each residue i in the protein, we computed the (i) molecular mechanics energy $E_i^{gas} = E_i^{MM,int} + E_i^{MM,vdW} + E_i^{MM,ele}$, where $E_i^{MM,int}$ includes contributions from bonded terms, $E_i^{MM,vdW}$ is the vdW interaction energy, and $E_i^{MM,ele}$ is the electrostatic interaction energy, (ii) the polar solvation free energy ($\Delta G_i^{solv,pol}$), and (iii) the nonpolar solvation free energy ($\Delta G_i^{solv,npl}$) relative to the corresponding values of residue i in an extended reference state, $CH_3–aa_i–NHCH_3$, where the residues do not interact with one another.[11]

The SASA of each residue was computed using FreeSasa.[12]

1 Glaser, F. *et al.* ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. *Bioinformatics* **19**, 163-164, doi:10.1093/bioinformatics/19.1.163 (2003).

2 Landau, M. *et al.* ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* **33**, 299-302 (2005).

3 Wu, C. H. *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **34**, D187-191, doi:10.1093/nar/gkj161 (2006).

4 Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431, doi:10.1186/1471-2105-11-431 (2010).

5 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

6 Nakamura, T., Yamada, K. D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34**, 2490-2492, doi:10.1093/bioinformatics/bty121 (2018).

7 Pupko, T., Bell, R., Mayrose, I., Glaser, F. & Ben-Tal, N. Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of

evolutionary determinants within their homologues. *Bioinformatics (Oxford, England)* **18 Suppl 1**, S71-77, doi:10.1093/bioinformatics/18.suppl_1.S71 (2002).

8 Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **285**, 1735-1747 (1999).

9 AMBER 2020 v. 20 (University of California, San Francisco., 2020).

10 Tian, C. *et al.* ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **16**, 528-552, doi:10.1021/acs.jctc.9b00591 (2020).

11 Chen, Y. C., Wu, C. Y. & Lim, C. Predicting DNA-binding amino acid residues from electrostatic stabilization upon mutation to Asp/Glu and evolutionary conservation. *Proteins-Structure Function and Bioinformatics* **67**, 671-680, doi:10.1002/prot.21366 (2007).

12 Mitternacht, S. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research*, S 189, doi:10.12688/f1000research.7931.1 (2016).