Journal of Structural Biology: X 5 (2021) 100055

Contents lists available at ScienceDirect

Journal of Structural Biology: X

journal homepage: www.sciencedirect.com/journal/journal-of-structural-biology-x

Graphical Structural Biology Review

Systematized analysis of secondary structure dependence of key structural features of residues in soluble and membrane-bound proteins

Mohammed H. AL Mughram [a,1], Noah B. Herrington [a,1], Claudio Catalano [a], Glen E. Kellogg [a,b,*]

[a] Department of Medicinal Chemistry and Institute for Structural Biology, Drug Discovery, and Development, Virginia Commonwealth University, Richmond, Virginia, USA

[b] Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, Virginia, USA

ARTICLE INFO

Keywords:
Protein structure
Solvent accessible surface area
Hydropathic interactions
Soluble proteins
Membrane proteins
Amino acid residue populations

ABSTRACT

Knowledge of three-dimensional protein structure is integral to most modern drug discovery efforts. Recent advancements have highlighted new techniques for 3D protein structure determination and, where structural data cannot be collected experimentally, prediction of protein structure. We have undertaken a major effort to use existing protein structures to collect, characterize, and catalogue the inter-atomic interactions that define and compose 3D structure by mapping hydropathic interaction environments as maps in 3D space. This work has been performed on a residue-by-residue basis, where we have seen evidence for relationships between environment character, residue solvent-accessible surface areas and their secondary structures. In this graphical review, we apply principles from our earlier studies and expand the scope to all common amino acid residue types in both soluble and membrane proteins. Key to this analysis is parsing the Ramachandran plot to an 8-by-8 chessboard to define secondary structure bins. Our analysis yielded a number of quantitative discoveries: 1) increased fraction of hydrophobic residues (alanine, isoleucine, leucine, phenylalanine and valine) in membrane proteins compared to their fractions in soluble proteins; 2) less burial coupled with significant increases in favorable hydrophobic interactions for hydrophobic residues in membrane proteins compared to soluble proteins; and 3) higher burial and more favorable polar interactions for polar residues now preferring the interior of membrane proteins. These observations and the supporting data should provide benchmarks for current studies of protein residues in different environments and may be able to guide future protein structure prediction efforts.

Introduction

Our knowledge and understanding of protein structure and function are continuously evolving, in large part aided by studies that solve three-dimensional structures of interest. These data have become a foundation for insight into the functional dependence of protein structure and for future drug discovery endeavors. To date, the Protein Data Bank (Berman et al., 2000) has close to 200,000 deposited entries and continues to expand as we continue to explore protein structure. Identifying and exploiting commonalities of structure should lead to schemes to predict the structure of new and structurally uncharacterized proteins. These are largely a mix of homology modeling, i.e., to define the overall shape by applying sequence-structure relationships (Webb and Sali, 2016; Waterhouse et al., 2018; Roy et al., 2010), rotamer selection from published libraries, and molecular mechanics-based optimization of atom-atom interactions. The recent AlphaFold program (Jumper et al., 2021) and implementations of Baker's Rosetta (Yang et al., 2020) are particularly noteworthy. Our approach has focused on characterizing and cataloguing those inter-atomic interactions and other structural characteristics as a function of secondary structure in three-dimensional maps (Ahmed et al., 2015; Ahmed et al., 2019; AL Mughram et al, 2021; Catalano et al., 2021; Herrington and Kellogg, 2021).

One interesting area of study concerns the major structural differences between soluble and membrane-bound proteins. Soluble proteins, notably, are known to have exteriors covered with polar residues facing outward into solvent where they interact favorably with water

* Corresponding author at: Department of Medicinal Chemistry and Institute for Structural Biology, Drug Discovery, and Development, Virginia Commonwealth University, Richmond, Virginia, USA.

E-mail addresses: almughrammh@vcu.edu (M.H. AL Mughram), herringtonnb@vcu.edu (N.B. Herrington), ccatalano@vcu.edu (C. Catalano), glen.kellogg@vcu.edu (G.E. Kellogg).

[1] Contributed equally.

https://doi.org/10.1016/j.yjsbx.2021.100055
Received 24 September 2021; Received in revised form 18 November 2021; Accepted 27 November 2021
Available online 30 November 2021

molecules, while their hydrophobic residues are usually packed within their interiors. Most membrane proteins, on the other hand, consist of extra- and intracellular domains connected by a large membrane-spanning domain, whose outer environment is the hydrophobic membrane interior. For this reason, it is much more common to find hydrophobic residues on the surface of this membrane-spanning domain. Many reports highlight this and the fact that many membrane protein interiors are also characterized by a greater abundance of polar residues linked by combinations of hydrogen bonding and salt bridges (von Heijne, 1992; Wimley and White, 1996; Luckey, 2016; Tamm et al., 2004; Zhang et al., 2015). Although α-helical and β-pleat structures add additional stability to both soluble and membrane-bound proteins, the tertiary structures they comprise can be drastically different, and the organization of their constituent residues merits further study.

In this graphical review, we describe a study of the solvent-accessible surface area (SASA) and hydropathic character of all residue types present in structures of both soluble and membrane proteins. We believe that the relative solvent accessibility and the hydropathic natures of residues found in both types of proteins are secondary structure-dependent and shed light on an important relationship between structure and protein folding in polar and hydrophobic environments. This work supplements findings of our previous contributions studying the hydropathic environments of specific amino acid types (Ahmed et al., 2015; Ahmed et al., 2019; AL Mughram et al., 2021; Catalano et al., 2021; Herrington and Kellogg, 2021), by encompassing all residue types in the context of soluble and membrane protein environments. As a result, we highlight notable differences in residue populations, SASAs and interaction characters between both. With this information, we further illuminate a complex relationship between protein secondary structure and hydropathic environments that satiate the "hydropathic valence" (Ahmed et al., 2019) of compositional residues.

Results

Our two datasets are as follows: 2703 soluble proteins described earlier (Ahmed et al., 2015) and 369 membrane protein structures extracted (Catalano et al., 2021) from the MemProtMD database (Newport et al., 2019). This set includes an artificial lipid molecule (dipalmytoylphosphatidylcholine, DPPC) to simulate the (not structurally-characterized) membrane. We use a Ramachandran (Ramachandran et al., 1963; Chen et al., 2010) "chessboard" schema (Ahmed et al., 2015) to bin residues by their backbone angles and hence, secondary structure (see Fig. 1A–C); i.e., each chess square (***a1*** to ***h8***) corresponds to a range of ϕ and ψ angles that correlate with secondary structure, with each residue studied binned into its corresponding chess square (see also Fig. 1D). The residue-type frequencies, for both datasets, are shown in Fig. 1E. Numerical data for the latter are available in Table S1 (Supporting Information). Generally, the frequencies for hydrophobic residues are higher in the membrane proteins, while the frequencies for polar residues (especially those normally charged) are higher in the soluble proteins. As in our earlier reports, we employed Fraczkiewicz and Braun's (1998) GETAREA algorithm to calculate the SASAs of every protein residue in our datasets to identify relationships between secondary structure and solvent accessibility. We also utilized the HINT force field (Kellogg and Abraham, 2000; Sarkar and Kellogg, 2010) to calculate the relative contributions of hydrophobic and polar interactions between those residues and their environments.

To gain insight into the relationship between residue type and secondary structure preferences, the relative residue populations of each chess square were tabulated (see Table S2). These are displayed in Figs. 2 and 3, for soluble and membrane proteins, respectively, as the sizes (represented in a $\log_{10}$ scale) of the solid boxes in each chess square (see legend). The colors of the boxes represent their solvent accessibility, calculated using a metric, $f_{outside}$, described previously (AL Mughram et al., 2021). In brief, $f_{outside}$ represents the fraction of residues in the set reported by GETAREA to have SASAs $\geq$ 50% of reference random-coil values. Note that GETAREA is not normally parameterized to recognize the DPPC molecule as either a solvent or to be part of the protein, so residues interacting with it are considered solvent-accessible (*vide infra*). In our analyses of SASA (Figs. 2 and 3), a few patterns are immediately apparent. First, the more hydrophobic residues tend to have lower solvent accessibility in both soluble and membrane-bound proteins, compared to their more polar counterparts. Residues such as phenylalanine, leucine, isoleucine, and valine have large clusters of $f_{outside}$ <0.4 in their right hand α-helix and β-pleat regions, indicating these residues tend to be highly buried within those secondary structures. Polar residues, such as glutamine and aspartic acid, on the other hand, are more likely to consistently appear in higher solvent-accessible environments across secondary structures, indicated by large clusters of $f_{outside}$ >0.6 squares. Additionally, all residues, apart from glycine and proline,

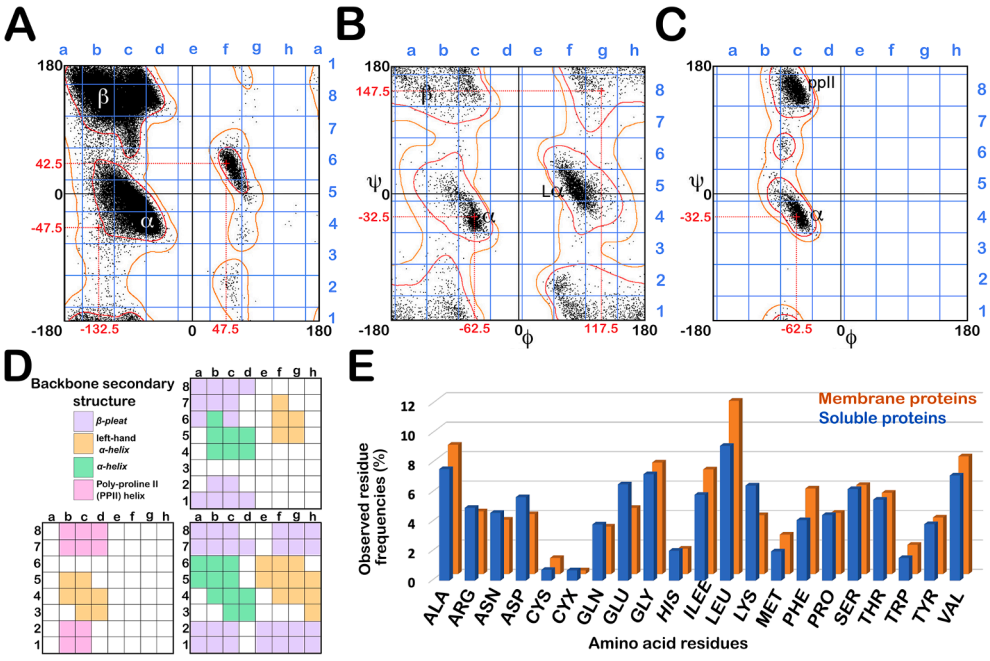

Fig. 1. Ramachandran (φ versus ψ) plots and the observed frequencies for residues in soluble and membrane proteins in our datasets. A) Ramachandran plot based on data of Lovell et al. (2003) with superimposed chessboard schema (blue) for all residues (except glycine and proline); **B)** Plot for glycine; and **C)** Plot for proline. Chess square centroids are illustrated with red dotted lines, e.g., in 1C, ***b4*** (φ = -132.5, ψ = -47.5); **D)** Mapping of secondary structure motif elements onto chessboards (upper right, all except PRO and GLY; lower left, PRO; lower right GLY); **E)** Observed amino acid residue frequencies for soluble (blue) and membrane (orange) proteins in the datasets. Images in A, B and C were adapted from https://en.wikipedia.org/wiki/Ramachandran_plot using the Creative Commons License 3.0 (https://creativecommons.org/licenses/by/3.0/legalcode) by superimposition of guidelines and labels from our chessboard schema. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

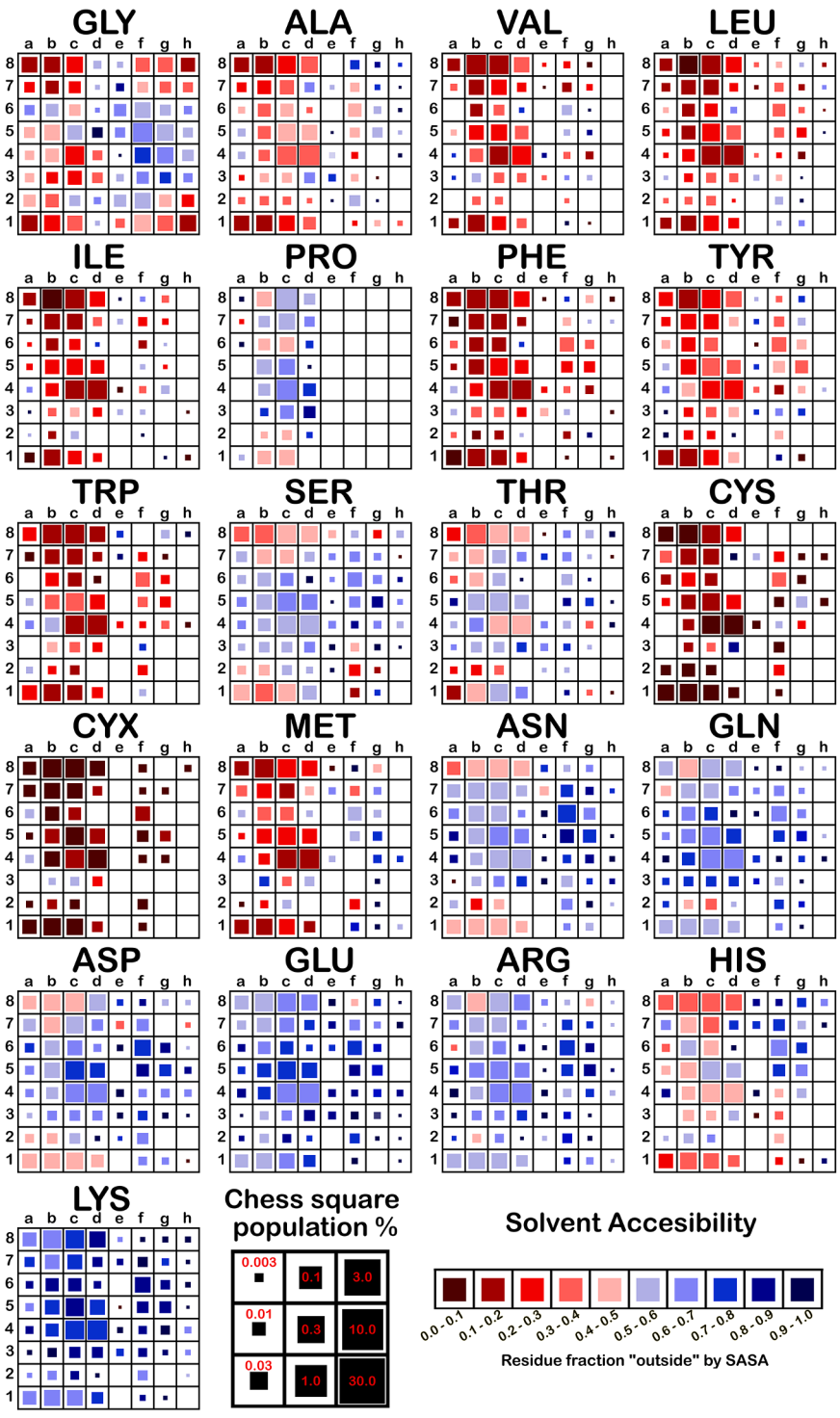

Fig. 2. Ramachandran chessboards illustrating chess square populations and solvent accessibility (fraction outside) in the soluble protein dataset. The size of the colored box within each square represents the population of each residue in $\log_{10}$ scale, as indicated by the key in the legend. Each represents the fractional percent of that residue's total population. Solvent accessibility is represented by fraction outside, which is derived from the GETAREA solvent accessible surface area, colored from dark red = most buried to dark blue = most exposed, as shown in the legend. For reference, chess squares associated with secondary structure motifs are mapped in Fig. 1D. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

exhibit their highest solvent accessibility in the left-hand α-helix region. These observations were anticipated, as we have seen them previously (Ahmed et al., 2015; Ahmed et al., 2019; AL Mughram et al., 2021; Catalano et al., 2021), and they are consistent with literature observations of SASA with respect to secondary structure (Lins et al. 2003).

Also of note is how solvent accessibility changes in hydrophobic and polar residues between soluble and membrane-bound proteins: those that are hydrophobic tend to become more solvent-accessible in membrane proteins, while polar residues tend to bury themselves in the protein interior. The former is particularly evident in the α-helix regions where the ***d4*** chess square is nearly always the most populated – contributing ~50% of the total population of several residue types in membrane proteins (Figs. 2 and 3, also Table S2). This is consistent with previous reports (Luckey, 2016; Tamm et al., 2004; Zhang et al., 2015). The Zhang et al. article reported that soluble helical bundles are spaced further apart with bulky hydrophobic residues, while transmembrane helices are brought closer together via hydrogen bonding and electrostatic networks. Clearly, solvent environment has a significant impact on the folding and other association phenomena of proteins and sequence identity is not the sole determining factor for a protein's three-dimensional structure.

We also analyzed each residue type's interaction character as a

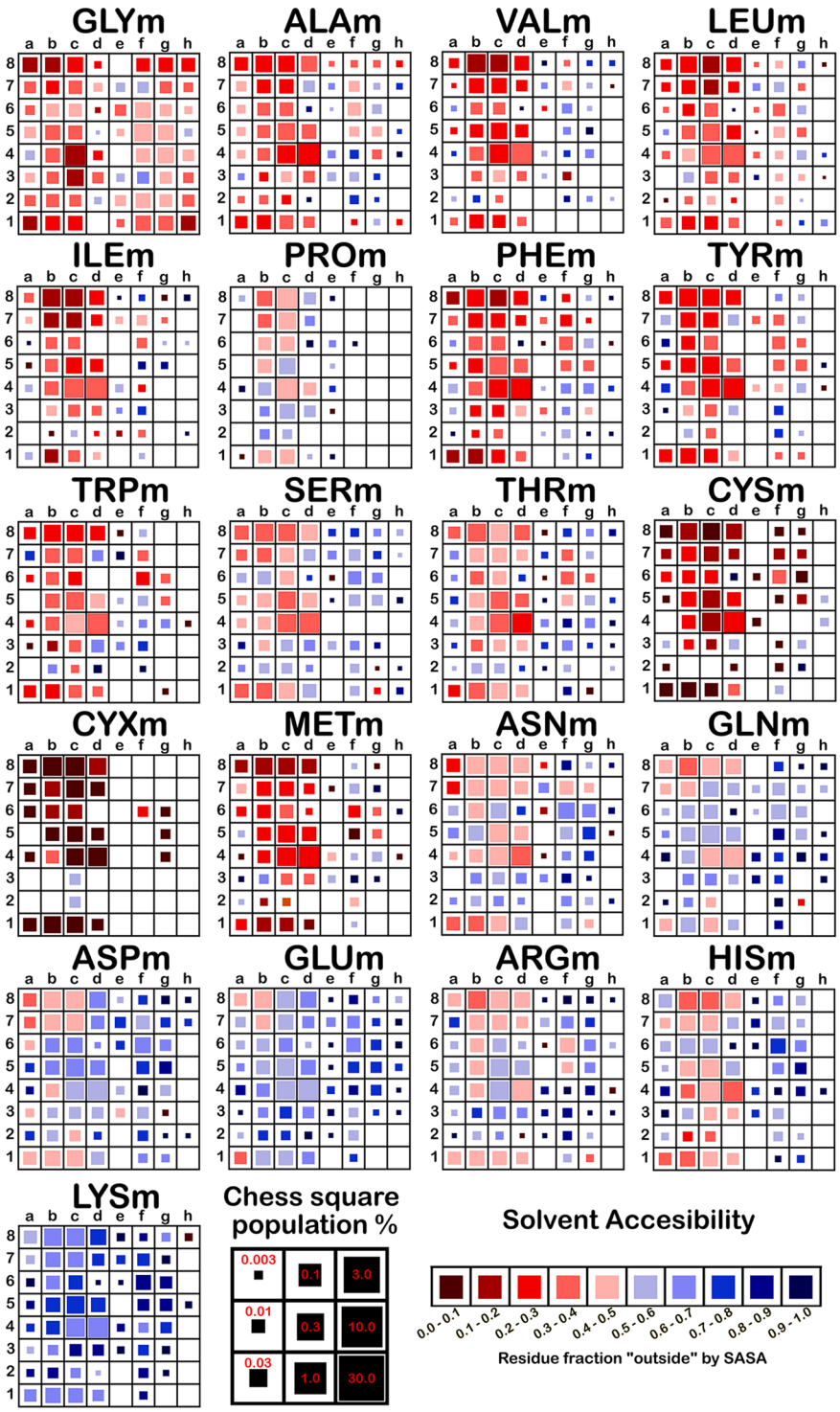

Fig. 3. Ramachandran chessboards illustrating chess square populations and solvent accessibility (fraction outside) in the membrane protein dataset. See caption for Fig. 2.

function of its secondary structure. Using the HINT force field (Kellogg and Abraham, 2000; Sarkar and Kellogg, 2010), we evaluated the hydropathic environments around each residue in our data set as a quartet of interaction types: positive polar (i.e., hydrogen bonding, salt bridges, π-cation), negative polar (i.e., acid-acid, base-base), positive hydrophobic (i.e., hydrophobic packing, π-π stacking), and negative hydrophobic (i.e., hydrophobic-polar – a desolvation cost). We determined the average fractional contribution by each interaction type for each chess square for each residue type (Figs. 4 and 5, also Table S2). One expected trend was that hydrophobic residues, being more "solvent" (actually lipid) exposed in membrane-bound proteins, would exhibit a greater fraction of positive hydrophobic interactions. Indeed, many residues of this type, including valine, isoleucine, tryptophan, and especially phenylalanine show a larger proportion of these interactions across many chess squares, as indicated by more substantial green bars in Fig. 5, which is attributable to their increased presence on protein surfaces, as suggested by Figs. 2 and 3, and hence in the artificial membrane environment. This raised the question of how polar residues would be

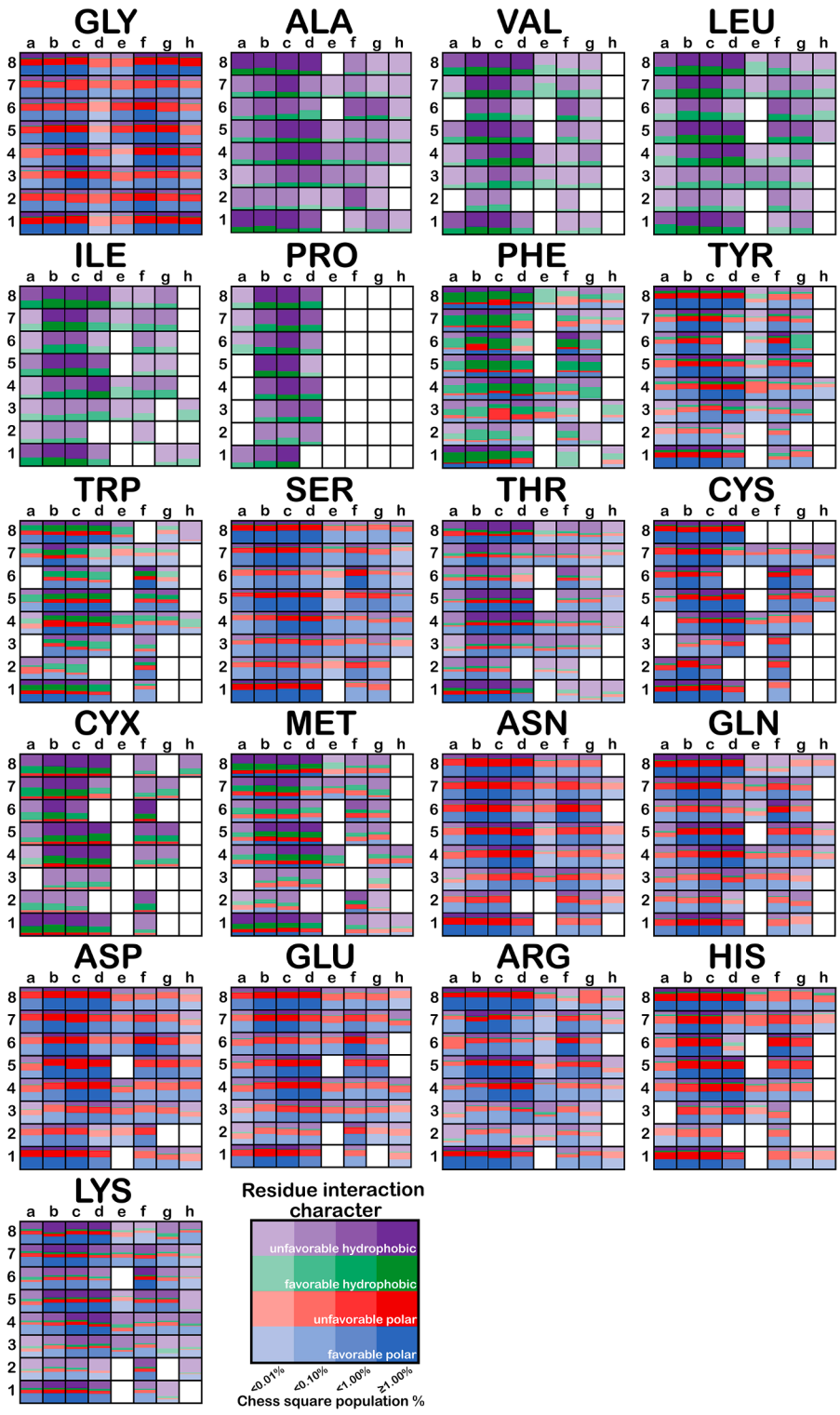

Fig. 4. Ramachandran chessboards illustrating fractional hydropathic interaction characters in the soluble protein dataset. The Interaction character of each residue was extracted from its hydropathic interaction maps as described in the text as four classes, layered, from top, within each chess square as: unfavorable hydrophobic (purple horizontal bars), favorable hydrophobic (green), unfavorable polar (red) and favorable polar (blue). These were averaged over all residues of the specified type within the chess square. The heights of these colored bars represent the character fractions for each of the four interaction types, while the transparencies of the sets of bars for each chess square indicate its relative population in terms of fractional percent (see legend, most solid = most populous). For reference, chess squares associated with secondary structure motifs are mapped in Fig. 1D. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

impacted by being shifted inward toward the protein interior. Our analysis shows a mixture of effects. Residues such as serine show decreased fractions of negative polar interactions (thinner red bars) in membrane protein structures and small increases in negative hydrophobic interactions (thicker purple bars). Serine and similar residues, preferring the more polar protein interior, force its Cβ into an environment with more unfavorable hydrophobic interactions with other polar residues. Lysine and arginine, based on the same principle of buriedness, show increased favorable polar and decreased unfavorable polar interactions in membrane proteins, likely due their being able to make closer, stronger polar interactions in the membrane protein interior. However, it should be mentioned that, even in membrane-bound proteins, there are extra- and intracellular components (loops, etc.) that are not within the membrane interior, and are exposed to water solvent at their surfaces. At present, our computational protocol does not include crystallographic or simulated waters of solvation, i.e., that surround extended sidechains for highly polar residues. Thus, we are likely underestimating the favorable and unfavorable polar interactions of

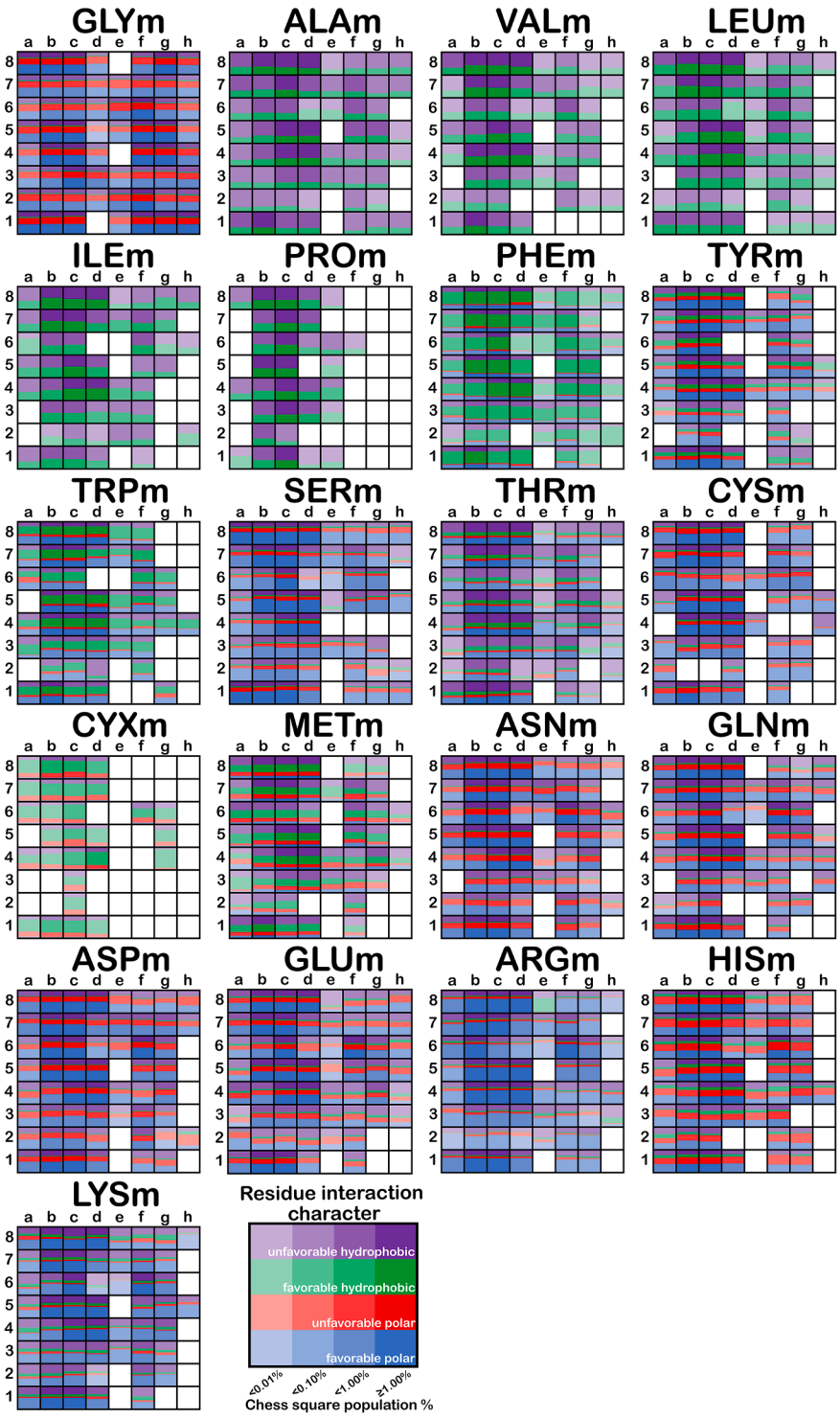

Fig. 5. Ramachandran chessboards illustrating fractional hydropathic interaction characters in the membrane protein dataset. See caption for Fig. 4.

such residues in soluble proteins.

An alternative arrangement of these plots is given in Supporting Information Fig. S1A–U, where the four plots for each residue type are grouped together in each image.

Summary and conclusions

Since the dawn of structural biology, science has made leaps and bounds in developing our understanding of the relationship between structure and function. Numerous studies have already analyzed this phenomenon from a variety of perspectives. Here, we put forth a comprehensive study that unites current understanding of protein secondary structure elements with their relative solvent accessibility and interaction environments with both graphical and numerical data. We feel that these results will be useful to numerous researchers investigating protein structure from both experimental and predictive viewpoints. It is difficult to formulate strong and definitive conclusions from these data as there are multiple moving pieces with complex and multifactorial relationships. Nevertheless, this graphical review begins to tie together secondary structure, solvent accessibility and interaction

character and sheds light on the more delicate details of residue-specific environments. A key has been our "chessboard" schema, which we believe captures subtle differences within secondary structural regions (e.g., the seven squares within the right-hand α-helix regions), while including enough data in each bin for the results to be meaningful.

One caveat of significance is that, while the set of soluble protein structures is fairly mature and likely reasonably diverse, there are fewer solved membrane protein structures available, and the difficulties inherent in their data collection and solution suggest that, over time, the nature of this group of proteins may change. For example, what we have reported appears to be overly weighted towards right-hand α-helical residue backbones, compared to left-hand α-helical or β-pleated backbone conformations. In conclusion, we hope this review offers a point of reference and new understanding for protein structural elements as they compose larger, more complex bodies. Moving forward, we intend to use the information we have gathered through dissecting these large datasets of structures to develop protein structure prediction tools.

Materials and methods

The chessboard schema was constructed from the standard Ramachandran plots (Lovell et al., 2003) by superimposing an 8×8 two-dimensional grid of 45°×45° bins, slightly frame shifted to more efficiently encompass secondary structure elements (Ahmed et al., 2015).

We conducted our analyses for this work primarily using two tools. All solvent-accessible surface area calculations for each residue in our data set were conducted with Fraczkiewicz and Braun's (1988) GETAREA server. The average SASA was calculated for each chess square using all residues in that chess square; other metrics were calculated as described previously (AL Mughram et al., 2021). Hydropathic character fractions were calculated using our in-house HINT force field (Kellogg and Abraham, 2000; Sarkar and Kellogg, 2010). Briefly, HINT utilizes residue-specific dictionaries of atomistic partial calculated partition coefficients ($\log P_{1\text{-octanol/water}}$) encoding free energy terms used to score the favorability of inter-atomic interactions. 3D maps of interactions between each residue's sidechain and its environment were calculated for each of the (quartet of) interaction types, positive polar, negative polar, positive hydrophobic, and negative hydrophobic, using methods reported earlier (Ahmed et al., 2015; Ahmed et al., 2019). Each resulting map was analyzed by summing the grid points by interaction type, and the fractional characters for each residue were calculated as described in a previous communication (Catalano et al., 2021). We then calculated the average fraction of each interaction type over all residues in a specific chess square.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge the structural insight into protein structures provided by Professors Martin K. Safo and Youzhong Guo.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.yjsbx.2021.100055.

References

Ahmed, M.H., Koparde, V.N., Safo, M.K., Neel Scarsdale, J., Kellogg, G.E., 2015. 3D interaction homology: The structurally known rotamers of tyrosine derive from a surprisingly limited set of information-rich hydropathic interaction environments described by maps. Proteins 83 (6), 1118–1136.

Ahmed, M.H., Catalano, C., Portillo, S.C., Safo, M.K., Neel Scarsdale, J., Kellogg, G.E., 2019. 3D interaction homology: The hydropathic interaction environments of even alanine are diverse and provide novel structural insight. J. Struct. Biol. 207 (2), 183–198.

AL Mughram, M.H., Catalano, C., Bowry, J.P., Safo, M.K., Scarsdale, J.N., Kellogg, G.E., 2021. 3D interaction homology: Hydropathic analyses of the "π-cation" and "π-π" interaction motifs in phenylalanine, tyrosine, and tryptophan residues. J. Chem. Inf. Model. 61 (6), 2937–2956.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. Nucleic Acids Research 28, 235–242.

Catalano, C., AL Mughram, M.H., Guo, Y., Kellogg, G.E., 2021. 3D interaction homology: Hydropathic interaction environments of serine and cysteine are strikingly different and their roles adapt in membrane proteins. Curr. Res. Struct. Biol. 3, 239–256.

Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., Richardson, D.C., 2010. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr. 66 (1), 12–21.

Fraczkiewicz, R., Braun, W., 1998. Exact and efficient analytical calculation of the accessible surface area and their gradients for macromolecules. J. Comput. Chem. 19, 319–333.

Herrington, N.B., Kellogg, G.E., 2021. 3D interaction homology: Computational titration of aspartic acid, glutamic acid and histidine can create pH-tunable hydropathic environment maps. Front. Mol. Biosci. 8 https://doi.org/10.3389/fmolb.2021.773385.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596 (7873), 583–589.

Kellogg, G.E., Abraham, D.J., 2000. Hydrophobicity: Is LogP(o/w) more than the sum of its parts? Eur. J. Med. Chem. 35 (7–8), 651–661.

Lins, L., Thomas, A., Brasseur, R., 2003. Analysis of accessible surface of residues in proteins. Protein Sci. 12 (7), 1406–1417.

Lovell, S.C., Davis, I.W., Arendall, W.B., de Bakker, P.I.W., Word, J.M., Prisant, M.G., Richardson, J.S., Richardson, D.C., 2003. Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. Proteins. 50 (3), 437–450.

Luckey, M., 2016. Introduction to the structural biology of membrane proteins. Computational Biophysics of Membrane Proteins 1–18. https://doi.org/10.1039/9781782626695-00001.

Newport, T. D., Sansom, M. S. P., Stansfeld, P. J., 2019. The MemProtMD database: a resource for membrane-embedded protein structures and their lipid interactions. Nucleic Acids Res. 47, D390-D397.

Ramachandran, G.N., Ramakrishnan, C., Sasisekharan, V., 1963. Stereochemistry of polypeptide chain configurations. J. Mol. Biol. 7 (1), 95–99.

Roy, A., Kucukural, A., Zhang, Y., 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. 5 (4), 725–738.

Sarkar, A., Kellogg, G.E., 2010. Hydrophobicity—Shake flasks, protein folding and drug discovery. Curr. Top. Med. Chem. 10, 67–83.

Tamm, L.K., Hong, H., Liang, B., 2004. Folding and assembly of β-barrel membrane proteins. Biochimica et Biophysica Acta 1666 (1-2), 250–263.

von Heijne, G., 1992. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J. Mol. Biol. 225 (2), 487–494.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T., 2018. SWISS-MODEL: Homology modelling of protein structures and complexes. Nucleic Acids Res. 46, W296-W303.

Webb, B., Sali, A., 2016. Comparative protein structure modeling using Modeller. Curr. Protoc. Bioinformatics 54, John Wiley & Sons, Inc., 5.6.1-5.6.37.

Wimley, W.C., White, S.H., 1996. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat. Struct. Biol. 3 (10), 842–848.

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., Baker, D., 2020. Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. U.S. A. 117 (3), 1496–1503.

Zhang, S.-Q., Kulp, D., Schramm, C., Mravic, M., Samish, I., DeGrado, W., 2015. The membrane- and soluble-protein helix-helix interactome: Similar geometry via different interactions. Structure 23 (3), 527–541.