Supporting information - Protruding hydrophobes on peripheral membrane proteins

Fuglebakk E, Reuter, N

December 21, 2016

S1 Data sets of experimentally identified binding sites

Inspecting structures from proteins with known membrane binding sites, we compiled a list of proteins we judged to have striking protrusions on the surface (Table S1). We used this list to decide on the our more formal definition of protrusions, including the decision of the parameter c and n, described in section 1Definitionssection*.1. The comparison with experimentally identified binding sites reported in Figure 9Comparison of predicted binding sites with experiment. Experimental and predicted orientations are compared via the angle of the vectors connecting the protein centre of mass with the binding site (Eq.**??**). The predicted binding site is taken to be the *Likely Inserted Hydrophobe*. Red marks are proteins where the *Likely Inserted Hydrophobe* has been experimentally identified to be interacting with the membrane. Proteins with no protruding hydrophobes are marked with a cross at 180. Experimetnal data is summarized in Table **??**. The boxplot in the top panel shows the angles of the experimental orientation compared with all protruding amino acids on the protein, not regarding hydropathic properties. label subfigures: a and bfigure.caption.12, was calculated for the data reported in Table S1, and the list of experimentally identified binding sites listed in Table S2.

PDBID	interfacial binding site	family
1RLW	F35, M38, L39, N95, Y96, V97, M98 [27]	C2-domain
1BYN	M173, G174, R233, F234, K235 [33, 34]	C2-domain
1UOV	V304, G305, I367, K369 [35]	C2-domain
1GMI	I89, Y91 [36]	C2-domain
1H6H	F35, Y94, V95 [28]	PX domain
1CZS	W26, W27 [37]	Discodin domain
1D7P	M2199, F2200, L2251, L2251 [12]	Discodin domain
1T6M	W51, Y92, Y208, W246, Y250, Y252 [39, 38, 26]	Bacterial PLC
1H0A	L6, M10 [31]	ENTH domain
1LOX	L195 [40]	Lipoxygenases

Table S1: Structures for which we manually judged insertion to be facilitated by protruding residues. The same structures were used as basis for (manual) parameterisation of the definition of protrusions. Family classifications are taken from OPM[19].

PDBID	interfacial binding site	classification
1DSY	M186, N189, N206, K211, R216, K230, S234, R249, R252 [41]	C2-domain
1O7K	R43 I65 W80 [28]	PX domain
1HYJ	V21, T22 [45]	FYVE PIP_3 domain
1VFY	L185 L186 R193 [29]	FYVE PIP_3 domain
1PTR	L250 W252 L254 [30]	C1 domain
1A8A	T72 S144 W185 S228 S303 [43]	Annexins
1DM5	E142 S144 G145 [44]	Annexins
1IAZ	W112 W116 [46]	Pore-forming Equinatoxin
1NB1	C1 G2 E4 T5 V6 G7 S18 W19 P20 V21 C22 G26 L27 P28 V29 [47]	Cyclotide
1POC	2I 14K 78I [48]	Insect sec. PLA_2
1N28	V3 K10 L19 F23 F63 K115 [49]	Vertebrate sec. PLA_2
1POA	W61 F64 Y110 [32]	Vertebrate sec. PLA_2
1VAP	W20 W109 [50]	Vertebrate sec. PLA_2
4P2P	W3 [55]	Vertebrate sec. PLA_2
1COY	M81 [51]	GMC oxidoreductases
1PFO	W464 W466 [52]	Chol.-dep. Cytolysin
1D1H	W30 [53]	Spider toxins
1PXQ	W30 [54]	Subtilosin A
2FNQ	W413 W449[56]	Lipoxygenases
1G13	T90 L126 N136 [57]	ML domain
1EIN	P42 D96 T123 I252 [25]	Fungal lipases
3PAK	Y164 R216 Y221 R222 [42]	Lectin domain
1F6S	K98 V99 [58]	C-type lysozyme
2DA0	K18 K19 I23 K25 N30 N48 N77 [59]	Pleckstrin-homology domain

Table S2: Membrane interacting residues used for systematic comparison with experiment. Family classifications are taken from OPM[19], except for 3PAK and 1F6S, which is not present in OPM and was taken from SCOPe [24]. Quaternary structures are also taken from OPM, except for 3PAK and 1F6S which was chosen based on the studies cited for their membrane interacting residues. Residue numbering corresponds to the listed PDBID. All structures are either monomers or homo-oligomers were all chains are likely to interact with the membrane. Chain identifiers are therefore not given.

S2 Analysis of alternative reference set and alternative quaternary structure determination

S2.1 Definition of alternative data sets

The reference set used for the reanalysis of protruding hydrophobes on peripheral membrane binders was obtained in the following way.

- We selected all PDBIDs classified in SCOPe [24] in the classes *All alpha proteins* (sunid: 46456), *All beta proteins* (sunid: 48724), *Alpha and beta proteins (a+b)* (sunid: 51349), *Alpha and beta proteins (a/b)* (sunid: 53931) or *Multi domain proteins* (sunid: 56572).
- We removed from this set any PDBIDs that has domains that are classified in the same SCOPe-family as any domain in the OPM-database[19]. This excludes not only the peripheral membrane binders, but also any transmembrane protein found in the reference set used for our primary analysis.
- To reduce redundancy, we iteratively removed proteins with domains that share SCOPe-family classification with any other domain in the set, until there were no such shared classifications left. This process ensures that there is at most one representative for each family in the set.
- We removed any structure not determined by X-ray crystallography, to be able to treat quaternary structure prediction consistently.
- We generated quaternary structure models using PISA [22].

For a fair comparison we similarly created a modified version of the set of peripheral proteins.

- We obtained the list of PDBIDs used for the set of peripheral binders in our primary analysis.
- We iteratively removed proteins with domains that share SCOPe-family classification with any other domain in the set, until there were no such shared classifications left. This process ensures that there is at most one representative for each family in the set.
- We removed any structure not determined by X-ray crystallography, to be able to treat quaternary structure prediction consistently.
- We generated quaternary structure models using PISA.

After further filtering out structures due to technical reasons like parsing problems, the final set of peripheral membrane binders contained 170 proteins, and the final reference set contained 2250 proteins.

SCOPe version 2.06 was used. PISA predictions were obtained through the "Protein interfaces, surfaces and assemblies" service PISA at the European Bioinformatics Institute. (`http://www.ebi.ac.uk/pdbe/prot_int/pistart.html`). Where PISA predicted that the asymmetric unit represents the most stable quaternary structure in solution, we obtained structures from the protein data bank (`http://www.rcsb.org/`)[20].

S2.2 Results and discussion

For our primary analysis we contrasted peripheral membrane binders to a reference set consisting of those fractions of protein surfaces of transmembrane complexes that are adequately far away from identified transmembrane regions. See section 1Data setssection*.1. This was done in order to have high confidence that most of the reference structures did not interact with membrane. It does however come with some concerns for relevance as we are comparing protein fragments with models of real proteins. In order to assess the robustness of the analysis of protruding hydrophobes to the composition of the reference set, we constructed an alternative reference, seeking to ensure complete protein models, and rather compromising on the confidence in lack of membrane interaction. We reanalysed the peripheral membrane proteins contrasting with this alternative reference. The data sets used for this analysis is described in section S2.1. In contrast to the exposed regions of transmembrane protein used in our primary analysis, we have no particular indication that the proteins in this reference set are not membrane binding. They are however chosen from the SCOPe classification and all pdb codes classified in OPM are excluded, so we find it reasonable to assume that it is less enriched in peripheral membrane binders than the set of peripheral proteins obtained from OPM. For a fair comparison, the set of peripheral proteins has also been obtained a bit differently from our primary analysis. In particular, family representatives are chosen differently, only X-ray determined structures are retained, and the quaternary structures are determined differently.

We performed calculations on the alternative data sets exactly as in our primary analysis, with two exceptions. Firstly, the alternative sets defined in section S2.1 were used. Secondly, wherever a mean over OPM-families were used in the primary analysis, we in stead used the values obtained for each protein. This was necessary as OPM-families are not defined for this reference set. To compensate, the alternative sets are in stead filtered for redundancy using the SCOPe classification of protein domains, as explained in section S2.1. Figures S2, S3, S4, S5 and S6 presents the reanalsysis of Figures 3The frequency of hydrophobes on different surface residues in peripheral proteins and the reference set. Panel A and C show the distributions for hydrophobes on protrusions, panel B and D for all solvent exposed residues. The histograms show the mean fraction (Eq.**??**) of hydrophobic amino acids in a family. The vertical axes denote the fraction of families in each data set.figure.caption.6, 5Comparing the weighted frequencies (Eq.**??**) of *hydrophobes* on *vertex* residues of varying local protein density (d). For each bin, the frequencies are compared between the peripheral membrane proteins and the reference set, using the odds ratio (Eq.**??**), so that positive values reflect higher frequencies in the peripheral proteins. More precisely, the vertical axis show $\text{peripheral}, \text{reference}, \hat{F}_{\text{hydrophobe}|\text{vertex}\cap l<d\leq u}$ where l and u denote the lower and upper limits of the ranges given on the vertical axis, and d is the local protein density defined in section **??**. Error bars are 95% confidence intervalsfigure.caption.8, 4Occurance of *co-insertable protruding hydrophobes* on protein surfaces. Panel A and B shows the weighted fraction of proteins (Eq. **??**) in the peripheral proteins (cyan) and the reference set (red). The analysis is done separately for two groups of proteins according to the total number of protrusions on the protein surface ($[0, 25\rangle$ in panel A, $[25, 50\rangle$ in panel B). We have differentiated here between protrusions that have at least one co-insertable pro-

truding hydrophobe (labeled "Co-ins."), and those that have not (labeled "isolated"). Panel C shows the frequency distribution of total number of protruding residues ("# protrusions") for all proteins. The selections analysed in panel A and B are found between the dashed lines in panel C. Error bars in panel A and B are 95% confidence intervals.figure.caption.7, 7The tendency for *protruding hydrophobes* to be *co-insertable*. The weighted frequency of co-insertion (Eq. **??**) is compared between each data set and a null model, using the odds ratio, so that positive values reflect higher frequencies of co-insertion than in the null model. More precisely, we show the comparisons $set, null, \hat{F}^{\text{pair}}_{\text{one,both}}$, where *set* represents the set of peripheral proteins (cyan) and the reference set (red), and *null* represent their respective null models where hydrophobes have been relocated randomly among protrusions as described in the methods section. Error bars are 95% confidence intervals.figure.caption.10, 2Comparison of frequencies of different amino acids at protruding positions. Panel A shows the weighted fractions (Eq. **??**) in peripheral proteins (cyan) and the reference set (red). In panel B, the contrast between the sets are quantified by the odds ratio (Eq. **??**), so that positive values reflect higher frequencies in the peripheral proteins. More precisely the vertical axis denote $\ln \text{peripheral}, \text{reference}, \hat{F}_{aa,\text{protruding}}$, with aa representing each of the standard amino acids. Error bars are 95% confidence intervals.figure.caption.5 and 6Panel A shows the weighted number (Eq. **??**) of *protruding hydrophobes* with the given secondary structures., grouped by their number of co-insertable protruding hydrophobes. We have differentiated between protrusions that have at least one co-insertable protruding hydrophobe (labeled "Co-ins."), and those that have not (labeled "isolated"). Panel B compares the weighted frequencies (Eq. **??**) of hydrophobes on protruding secondary structures. The frequencies are compared between the peripheral membrane proteins and the reference set, using the odds ratio (Eq. **??**), so that positive values reflect higher frequencies in the peripheral proteins. More precisely, panel A show the values $N_{\text{hydrophobe}|\text{protrusion}\cap sse}$, and panel B the comparisons $\text{peripheral}, \text{reference}, \hat{F}_{\text{hydrophobe}|\text{protrusion}\cap sse}$ where sse specifies the different secondary structures given on the horisontal axis. Error bars in panel B are 95% confidence intervals.figure.caption.9 from our primary analysis. The data representing peripheral membrane binders are colored in cyan in all plots, and and the data representing the alternative reference set is colored in red.

All main conclusions are supported by this reanalysis. We reproduce the relative importance of big aliphatic residues on protruding locations in peripheral proteins (Figure S6), the higher density of hydrophobes on protruding locations on the surface of peripheral membrane binders (Figure S2), the propensity for protruding hydrophobes to be co-insertable (Figure S5), the stronger contrast between the data sets when the analysis is restricted to low protein density (Figure S3), the stronger contrast between the data set when the analysis is restricted to co-insertable pairs of protruding hydrophobes (Figure S4), and the results from decomposition on secondary structure (Figure S7).

Even so the contrast between the data sets are smaller than in our primary analysis, which is particularly striking when comparing Figure S2 with Figure 3The frequency of hydrophobes on different surface residues in peripheral proteins and the reference set. Panel A and C show the distributions for hydrophobes on protrusions, panel B and D for all solvent exposed residues. The histograms show the mean fraction (Eq.**??**) of hydrophobic amino acids

Figure S1: Histogram of differences between the number of chains in the protein models used in OPM and the models predicted by PISA. The histogram is calculated for all PDBIDs that occur in both the set of peripheral binders used in our primary analysis and in the alternative analysis presented in this supplementary material. When more chains are present in the PISA models, the difference is negative.

in a family. The vertical axes denote the fraction of families in each data set.figure.caption.6 and Figure S4 with Figure 4Occurance of *co-insertable protruding hydrophobes* on protein surfaces. Panel A and B shows the weighted fraction of proteins (Eq. **??**) in the peripheral proteins (cyan) and the reference set (red). The analysis is done separately for two groups of proteins according to the total number of protrusions on the protein surface ($[0, 25\rangle$ in panel A, $[25, 50\rangle$ in panel B). We have differentiated here between protrusions that have at least one co-insertable protruding hydrophobe (labeled "Co-ins."), and those that have not (labeled "isolated"). Panel C shows the frequency distribution of total number of protruding residues ("# protrusions") for all proteins. The selections analysed in panel A and B are found between the dashed lines in panel C. Error bars in panel A and B are 95% confidence intervals.figure.caption.7. Partly this can be explained by the lack of negative assertions for the reference set, it might well contain some peripheral membrane binders, but it is also evident that the prevalence of protruding hydrophobes is smaller in the set of peripheral proteins considered in the re-analysis. We suspect this is due to the difference in how quaternary structure or oligomeric states are determined. Consistent with this, we find that PISA do tend to predict higher order complexation than the quaternary structure models chosen by the OPM curators. See Figure S1. We do expect the curated models used in OPM to be more accurate in this respect, but choose to analyse both sets with PISA models for consistency.

Comment on S5 ?

Figure S2: Reanalysis of Figure 3The frequency of hydrophobes on different surface residues in peripheral proteins and the reference set. Panel A and C show the distributions for hydrophobes on protrusions, panel B and D for all solvent exposed residues. The histograms show the mean fraction (Eq.**??**) of hydrophobic amino acids in a family. The vertical axes denote the fraction of families in each data set.figure.caption.6.

Figure S3: Reanalysis of Figure 5Comparing the weighted frequencies (Eq.**??**) of *hydrophobes* on *vertex* residues of varying local protein density (d). For each bin, the frequencies are compared between the peripheral membrane proteins and the reference set, using the odds ratio (Eq.**??**), so that positive values reflect higher frequencies in the peripheral proteins. More precisely, the vertical axis show $\text{peripheral, reference, } \tilde{F}_{\text{hydrophobe}|\text{vertex}\cap l<d\leq u}$ where l and u denote the lower and upper limits of the ranges given on the vertical axis, and d is the local protein density defined in section **??**. Error bars are 95% confidence intervalsfigure.caption.8.

Figure S4: Reanalysis of Figure **??**.

Figure S5: Reanalysis of Figure 7The tendency for *protruding hydrophobes* to be *co-insertable*. The weighted frequency of co-insertion (Eq. **??**) is compared between each data set and a null model, using the odds ratio, so that positive values reflect higher frequencies of co-insertion than in the null model. More precisely, we show the comparisons $set, null, \hat{F}^{\text{pair}}_{\text{one,both}}$, where *set* represents the set of peripheral proteins (cyan) and the reference set (red), and *null* represent their respective null models where hydrophobes have been relocated randomly among protrusions as described in the methods section. Error bars are 95% confidence intervals.figure.caption.10.

Figure S6: Reanalysis of Figure 2Comparison of frequencies of different amino acids at protruding positions. Panel A shows the weighted fractions (Eq. **??**) in peripheral proteins (cyan) and the reference set (red). In panel B, the contrast between the sets are quantified by the odds ratio (Eq. **??**), so that positive values reflect higher frequencies in the peripheral proteins. More precisely the vertical axis denote $\ln \text{peripheral, reference}, \hat{F}_{aa,\text{protruding}}$, with aa representing each of the standard amino acids. Error bars are 95% confidence intervals.figure.caption.5.

Figure S7: Reanalysis of Figure 6Panel A shows the weighted number (Eq. **??**) of *protruding hydrophobes* with the given secondary structures., grouped by their number of co-insertable protruding hydrophobes. We have differentiated between protrusions that have at least one co-insertable protruding hydrophobe (labeled "Co-ins."), and those that have not (labeled "isolated"). Panel B compares the weighted frequencies (Eq. **??**) of hydrophobes on protruding secondary structures. The frequencies are compared between the peripheral membrane proteins and the reference set, using the odds ratio (Eq. **??**), so that positive values reflect higher frequencies in the peripheral proteins. More precisely, panel A show the values $N_{\text{hydrophobe}|\text{protrusion}\cap sse}$, and panel B the comparisons $\text{peripheral, reference}, \hat{F}_{\text{hydrophobe}|\text{protrusion}\cap sse}$ where sse specifies the different secondary structures given on the horisontal axis. Error bars in panel B are 95% confidence intervals.figure.caption.9.

References

See main article