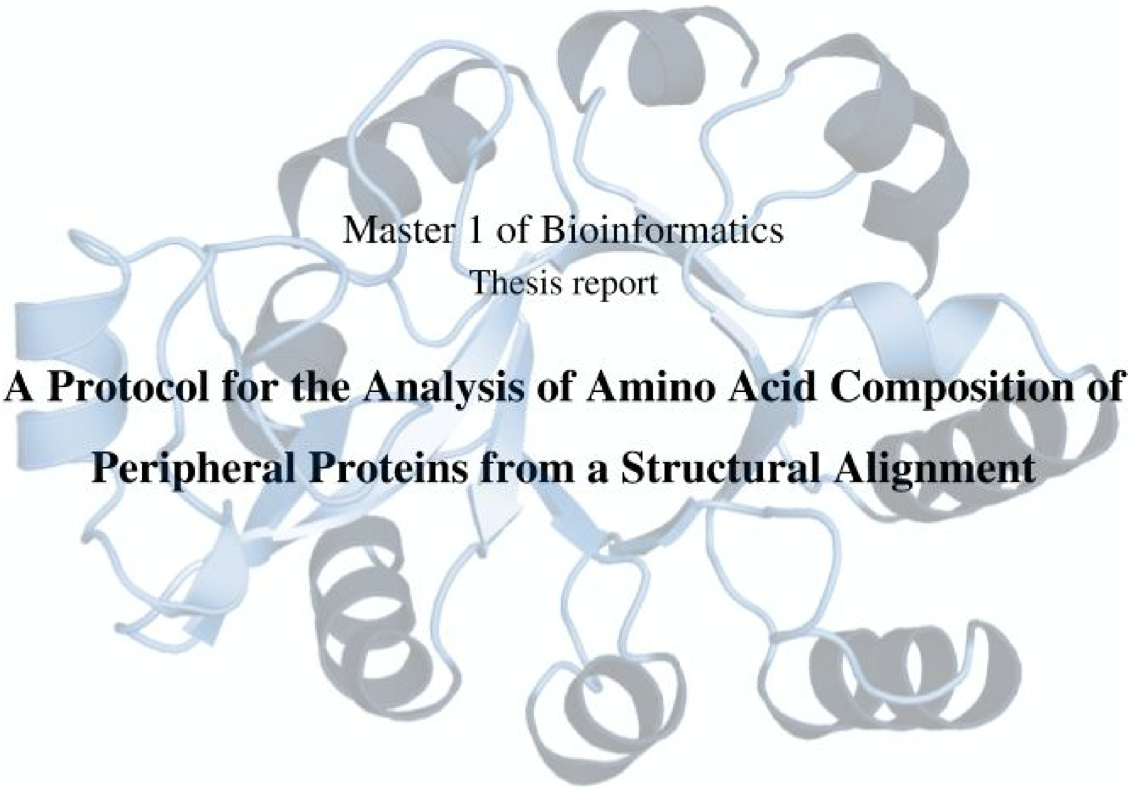
Academic year :
2017/2018



Master 1 of Bioinformatics
Thesis report

A Protocol for the Analysis of Amino Acid Composition of Peripheral Proteins from a Structural Alignment

Author :
Hélène Kabbech
helene.kabbech@gmail.com
+33 6 37 61 80 97

Supervisor :
Pr. Nathalie Reuter
nathalie.reuter@cbu.uib.no
+47 55 58 40 40

Host laboratory :
Computational Biology Unit (CBU)
Høyteknologisenteret
Thormøhlensgt 55
N-5008 Bergen, Norway

Acknowledgments

Firstly, I would like to express my sincere gratitude to Prof. Nathalie Reuter for offering me the opportunity to do my internship in her group. Besides being an extremely kind and professional person, her guidance helped me through the research and writing part of this project.

I need to thank the heads of the CBU for welcoming me to their offices and providing me a nice working environment.

A special thanks to : Adam Giess, Gunnar Schulze, Pierre Bédoucha, Lauriane Paquet and all the other CBU colleagues to have integrated me during my stay. Noelly Madeleine for being my officemate during these three months. Angèle Abboud, Bojan Krtenic and Dr. Hanif Muhamad Khan for the discussions and for giving me some advices about my project.

Finally, I am thankful to all the members of the Reuter group : Nathalie Reuter, Angèle Abboud, Bojan Krtenic, Pierre Bédoucha, Emmanuel Edouard Moutoussamy, Noelly Madeleine, Dr. Hanif Muhamad Khan, Dr. Qaiser Waheed, Lauriane Paquet, Olga Klonova and Fahimeh Khorsand for their continuous help and cheerfulness.

Table of contents

Introduction

In the 17th century, magnification technology advanced enough to discover cells; discovery largely attributed to Robert Hooke. The latter misled the cell membrane theory that all cells contained a hard cell wall delimiting the intracellular from the extracellular environment. It is only in the middle of the 20th century, with the emergence of electron microscopy, that the structure of the cell membrane has been characterized. Nowadays, we can easily define a biological membrane as a structure which acts as a selective barrier between different compartments of the cell and the cell itself. It allows or stops some substances to pass through; such substances may be molecules, ions or other small particles. Its structure consists mostly of lipids organized in two layers polar headgroups. Several other types of biomolecules anchor permanently the biological membrane such as integral membrane proteins and cholesterol, or bind its surface transiently as the peripheral membrane proteins. These latter bind the surface of biological membranes in order to accomplish their functions which can be a key step in signaling cascades, lipid transport or lipid processing.

The current textbook model of peripheral protein membrane association consists of (1) an electrostatically-driven approach most often followed by (2) the intercalation of hydrophobic side chains into the lipid bilayer. The first step is characterized by long-range nonspecific electrostatic forces between the negatively charged membrane and clusters of positively charged amino acids on the protein surface. These forces bring the protein into a binding-competent orientation to the lipid bilayer. The second step consists in the insertion of hydrophobic groups into the hydrophobic core of the membrane. [1] The prototypical interfacial binding site (IBS) is described as containing several positively charged amino acids and a few hydrophobic amino acids such as those with aromatic or aliphatic groups, or covalent lipid anchors.

In a previous study carried out by the group [2], the different membrane-binding residues of the peripheral protein *Bc*PI-PLC have been listed and their respective contributions studied in detail. It appeared that it did not fit with the textbook model, for example it lacked clusters of positively charged amino acids which was later shown to be the case for other peripheral proteins. These discoveries call for a better mapping of membrane-binding sites on peripheral proteins. Only few proteins have reliable annotations of their membrane binding site and this represents a challenge for large-scale mapping of amino acid compositions. Although sequence conservation of membrane-binding sites is usually low, they are generally conserved in protein structures in terms of the structural elements involved (*i.e.* loops or turns, helices, strands).

Here I propose to take advantage of structure alignments of membrane-binding domains belonging to the same family and for which the IBS is identified. These alignments can then be augmented with sequence data from related domains for which the structure is unknown. The final alignments can be used to perform an inventory of the amino acid composition of the IBS across the family.

Material and Methods

I developed a python[3] pipeline that integrates sequence data onto structural alignments of peripheral protein families and performs an inventory of the amino acids present at the membrane-binding and the non-membrane-binding regions. This is made possible by the use of existing annotation of the membrane-binding site of a prototype domain. This binding site (called *mb*SSE) is identified as a series of secondary structure elements known from experimental data to be involved in membrane binding. All the other secondary structure elements are assumed on the *solvent* side (*sol*SSE).

1. Pipeline : identification of *mb*SSE on structurally aligned protein domains

The user needs to provide a structural alignment in FASTA format and the corresponding PDB [4] files. A prototype domain must be chosen among the domains of the alignment and its *mb*SSE have to be indicated. This prototype, in the context of a structural sequence alignment, allows to identify *mb*SSE of all the domains. The pipeline allows the visualization of the structural alignment in a terminal colored by *mb*SSE as well as the superposed PDB files in PyMOL [5]. For a chosen SSE, it gives out a CSV file containing the amino acid inventory for each domain. (see **Figure 1**)

a) Identification of secondary structures from PDB file using DSSP (mkdssp)

The Dictionary of Secondary Structure Proteins (DSSP) program [6] assigns to each residue one of eight secondary structure states based on atomic coordinates stored in the PDB file. The eight different states and their associated symbols are : α-helix (H), isolated β-bridge (B), extended strand/participates in β ladder (E), 3-helix/310 helix (G), 5 helix/π-helix (I), hydrogen bonded turn (T), bend (S) and random coil (*blank*). In this study, we restrict the number of states to four : H for helices (corresponding to states H, G and I from DSSP), E for strands (E from DSSP), T for turns (T from DSSP) and C for coils (DSSP states B, S and *blank*).

b) Construction of the *mb*SSE and *sol*SSE dictionary

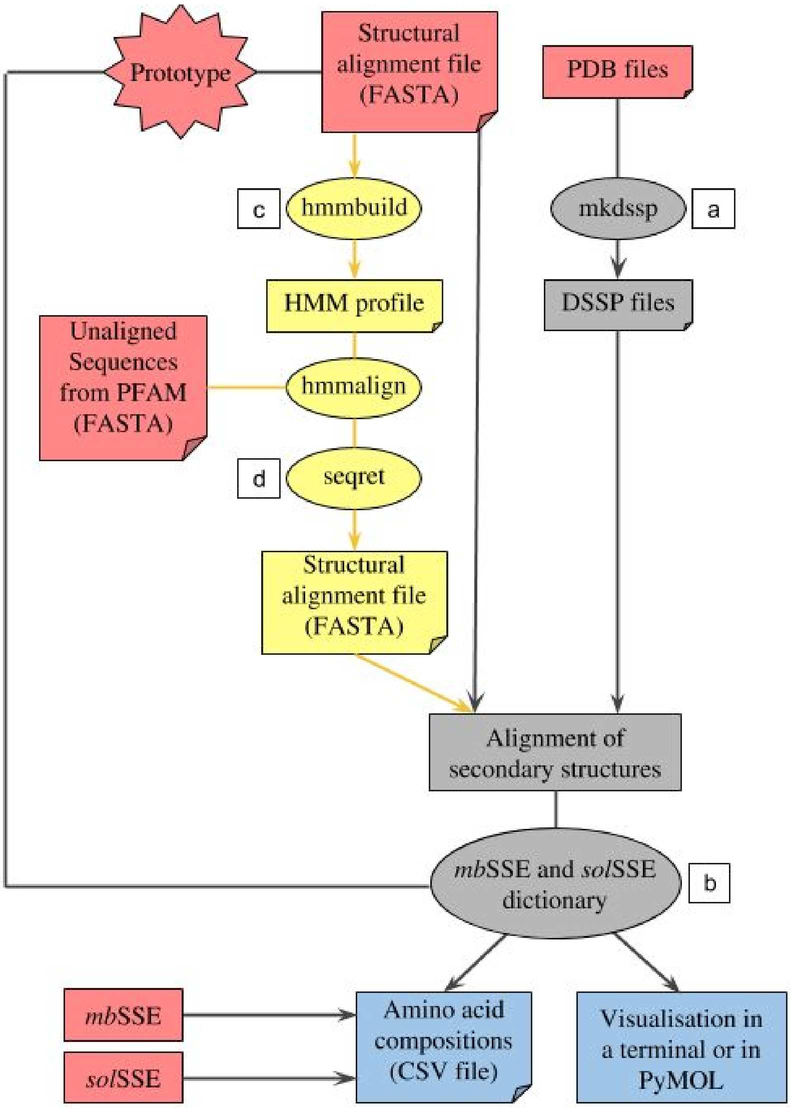
Keys of the dictionary are protein domains. The item corresponding to a key (= a *domain*) is a list of all the *mb*SSE and *sol*SSE mapped along the sequence of that domain. Informations on the secondary structure types (*i.e.* loop, turn, helix and strand) and their name (*e.g.* α_N or β_N) are stored.

c) Alignment generated from unaligned sequences using Hmmer (hmmbuild & hmmalign)

With the aim of enriching the structural alignment we use Hmmer [7], a tool for biosequence analysis using profile hidden Markov models (HMM). I built a HMM profile from the multiple sequence alignment derived from structural alignments retrieved from CATH [8]. This is done with the hmmbuild program. Further sequences of related families are retrieved from the Pfam database [9] which is a large collection of protein families, and aligned to the generated profile with hmmalign. The generated alignment is in the SELEX format. Using this sequence alignment we annotate the position of *mb*SSE in each sequence for which no structure is known. We rely on the principle that these elements are generally conserved through evolution.

d) Conversion format (seqreq)

In order to convert the obtained structural alignment in SELEX format into a FASTA format, seqret program (EMBOSS [10] suit) is used.



Prototype
Structural alignment file (FASTA)
PDB files
c
hmmbuild
mkdssp
a
HMM profile
DSSP files
Unaligned Sequences from PFAM (FASTA)
hmmalign
d
seqret
Structural alignment file (FASTA)
Alignment of secondary structures
*mb*SSE and *sol*SSE dictionary
b
*mb*SSE
*sol*SSE
Amino acid compositions (CSV file)
Visualisation in a terminal or in PyMOL

Figure 1. Diagram of the pipeline. Inputs are in red and outputs in blue. The yellow path describes how a new alignment is generated from unaligned sequences. Main programming steps are in circles (see the corresponding letter for details).

2. Structural alignments and superposed PDB files provided by CATH

a) CATH database

The CATH database [8] provides structural classification of protein domains based on their fold patterns. Protein structures are downloaded from the Protein Data Bank and then classified within the CATH structural hierarchy with 4 different levels called class (C), architecture (A), topology (T) and homologous superfamily (H). At the C-level, domains are grouped according to their secondary structure content (mainly alpha, mainly beta, alpha/beta or few secondary structures). The A-level assigns domains according to the general orientations of their secondary structures. At the T-level, the fold of the domains corresponding to how the secondary structure elements are connected to each other and arranged is taken into account. Then, at the H-level the classification is based on a combination of both sequence similarity and a measure of structural similarity. Superfamily related domains are clustered at different degrees of sequence similarity : 35, 60, 95 and 100%.

b) Choice of the superfamilies studied

We decided to apply our pipeline first on the 2.30.29.30 superfamily (CATH version 4.2.0) which contains 818 domains of Pleckstrin-homology domains (PH domains) and Phosphotyrosine-binding domains (PTB). PH domains are well-known membrane-targeting domains [11]. We then will apply our pipeline on the 3.20.20.190 superfamily (version 4.1.0), also called Phosphatidylinositol (PI) phosphodiesterase, containing 128 protein domains. This superfamily includes *Bc*PI-PLC (*Bacillus cereus*), *Lm*PI-PLC (*Listeria monocytogenes*) and *Sa*PI-PLC (*Staphylococcus aureus*) studied by the group.

c) CATH topological classification of the superfamilies studied

According to the CATH database the topology of a PH-domain (2.30.29), consists of two perpendicular antiparallel β-sheets, followed by a C-terminal α-helix (see **Figure 2A**). The 3.20.20.190 superfamily topology is a TIM barrel (3.20.20). It is one of the most common protein folds [12,13], consisting of eight α-helices and eight parallel β-strands that alternate along the peptide backbone (see **Figure 2B**).

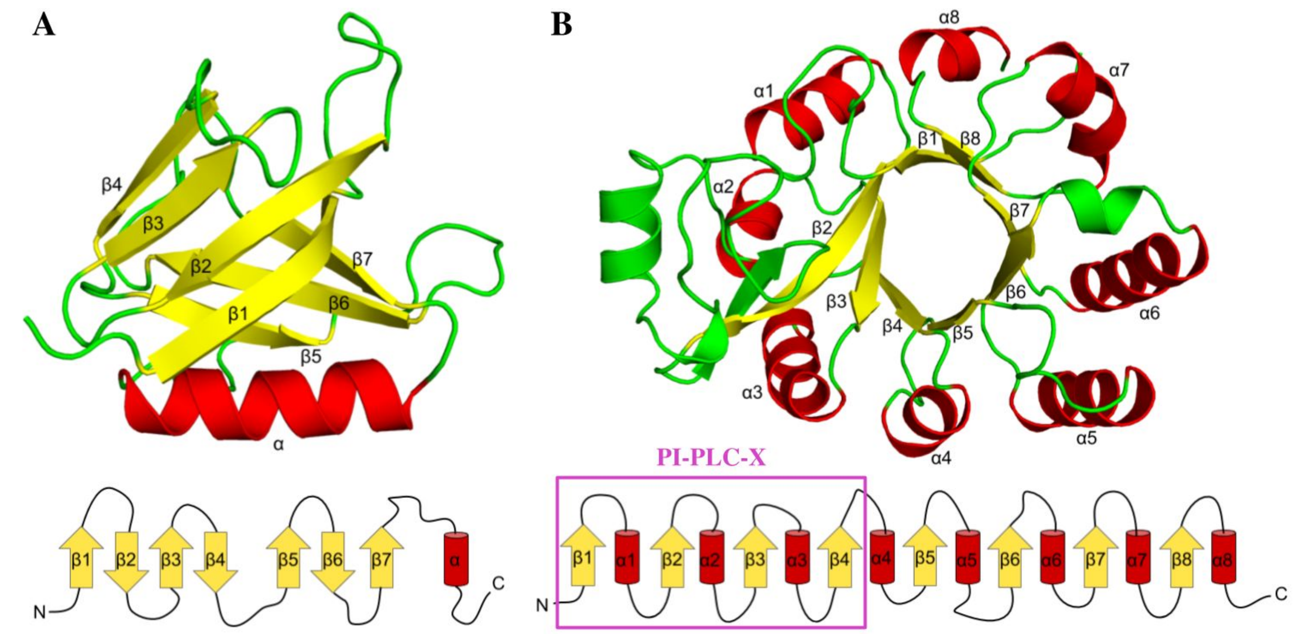


Figure 2. (**A**) Example of the PH-domain topology (PDB 1dynA00) colored by the main structural elements (helices in red and strands in yellow) and segments (in green). (**B**) Example of the TIM barrel topology (PDB 1o1dA00). Below, the topological diagrams related. The PI-PLC-X domain is in pink color.

d) Structural alignments

A structural alignment of protein domain sequences is provided by the CATH developers for the superfamilies studied. Among a superfamily, one domain is selected from each cluster having 35% of sequence identity. As a result, the diversity into a superfamily in term of organisms, protein functions and domain sequences is well represented. The superposed PDB files of the protein domains are also provided by CATH. The alignment for the 2.30.29.30 superfamily is constituted by 149 protein domains, whereas the alignment for the 3.20.20.190 superfamily has 18 domains (see **Table 2**).

3. Generation of peripheral protein datasets for application to the pipeline

Our amino acid composition analyzes are exclusively focused on peripheral proteins. Thereby we kept the domains whose membrane binding function is proven. We studied two families of proteins belonging to the superfamilies studied. The first one is a family of well-known membrane targeting domains; Pleckstrin Homology domain (PH domain) and the second one a family of lipid-processing enzymes : Phosphatidylinositol-specific phospholipase C (PI-PLC). This was particularly necessary for the PI-PLC as the structural alignment from CATH contained non-membrane-binding enzymes. We also added domain sequences to enrich the datasets.

a) PH domain dataset

The set of PH domains was taken from the work of the Gavin group [11], and restricted to those experimentally demonstrated to interact with liposomes. We retrieved the domain sequences of these PH domains from the Pfam database [9]. Indeed, the PF00169 Pfam family matches to the 2.30.29.30 CATH superfamily. As a result, by aligning these sequences to the hmm profile and removing domains with a lack of sequence conservation in the regions containing the identified *mb*SSE, we obtain an alignment of 27 PH domains, including 8 domain structures from the previous CATH structural alignment (see **Table 1**).

PH domains from the CATH alignment
AVO1_YEAST (**3ulbA01**), AKT1_HUMAN (**4ejnA01**), DYN1_HUMAN (**1dynA00**), OSH3_YEAST (**4iapA01**), PDPK1_HUMAN (**1w1hD00**), PLEK_HUMAN (**1plsA00**), RT106_YEAST (**3fssA02**), SOS1_HUMAN (**1dbhA02**).
PH domains added
ATG26_YEAST, G0SB85_CHATD, BEM3_YEAST, G0SHS0_CHATD, BOI2_YEAST, BUD4_YEAST, G0S5F3_CHATD, CA120_YEAST, G0S0E0_CHATD, CLA4_YEAST, SHE3_YEAST, OSH2_YEAST, SIP3_YEAST, SKM1_YEAST, SLM2_YEAST, OSH1_YEAST, G0S1U9_CHATD, LAM1_YEAST, G0S6I2_CHATD.

Table 1. Set of 27 PH domains aligned to the hmm profile and kept for the analysis. Names refer to the entry names used by UniProt [14]. PDB structures are in bold.

b) PI-PLC-X dataset

For our second case, only 4 PDB id are recorded into the OPM database [15] of transmembrane and peripheral protein structures among the 18 domain structures of the CATH alignment (see **Table 2**). An enzymatic protein usually has an EC (Enzyme Commission) number assigned, reflecting its enzymatic activity, for instance EC numbers starting with 3 are enzymes that use a water molecule to hydrolyse bonds of the substrate. By analyzing the EC numbers corresponding to the proteins recorded by OPM (*i.e.* 3.1.4.11, 3.1.4.4 and 4.6.1.13; see **Supplementary datas**), we observed that for each reaction a lipid is implied as a substrate. As a result, we assume that these proteins bind the membrane in order to realize their function. By analyzing the EC numbers of the 14 other domains (*i.e.* 3.1.4.1, 3.1.4.2 and 3.1.4.46 EC; see **Supplementary datas**), it appears that the reactions do not involve membrane lipids as substrates or products. These are therefore discarded from the dataset. The PI-PLC and PI-PD*β2* peripheral proteins detected contain the conserved PI-PLC-X domain (PF00388 Pfam family) which includes the first 4 *β*-strands and the first 3 *α*-helices (see **Figure 2B**). We thus focus the study of amino acid composition on the PI-PLC-X domain and we use only the part of the CATH alignment of the PI-PLC-X domain. The seed of the PF00388 Pfam family was then aligned to the hmm profile. The seed of a Pfam family is constituted by non-redundant reference sequences. By removing domains with a lack of sequence conservation in the identified *mb*SSE, we obtain an alignment of 58 PI-PLC-X domains, including 3 domains from the previous CATH alignment and 3 new domain structures (see **Table 3**).

PDB ID_domain	*Organism*	Protein name	EC number	OPM recording
2plc_A00	*Listeria monocytogenes*	PI-PLC	4.6.1.13	Yes (1aod)
4i90_A00	*Staphylococcus aureus*	PI-PLC	4.6.1.13	Yes (4f2t)
2zkm_X03	*Homo sapiens*	PI-PD$\beta 2$	3.1.4.11	Yes
3rlg_A00	*Loxosceles intermedia*	SM-PDD	3.1.4.4	Yes (3rlh)
3qvq_A00	*Oleispira antarctica*	PD	3.1.4.2	No
1o1z_A00	*Thermotoga maritima*	GDPD	3.1.4.46	No
1vd6_A00	*Thermus thermophilus*	GDPD	3.1.4.46	No
1ydy_A00	*Escherichia coli*	GDPD	3.1.4.46	No
2oog_D00	*Staphylococcus aureus*	GDPD	3.1.4.46	No
3ch0_A00	*Cytophaga hutchinsonii*	GDPD	3.1.4.46	No
2otd_A01	*Shigella flexneri 2a*	GDPD	3.1.4.1	No
3mz2_A00	*Parabacteroides distasonis*	GDPD	-	No
3no3_A00	*Parabacteroides distasonis*	GDPD	-	No
1zcc_A00	*Agrobacterium tumefaciens*	GDPD	-	No
2o55_A00	*Galdieria sulphuraria*	putative GDPD	-	No
3i10_A00	*Bacteroides thetaiotaomicron*	putative GDPD	-	No
3ks6_A00	*Agrobacterium fabrum*	putative GDPD	-	No
3l12_B00	*Ruegeria pomeroyi*	putative GDPD	-	No

Table 2. The 18 protein domains of the 3.20.20.190 CATH superfamily alignment.

PI-PLC : Phosphatidylinositol-specific phospholipase C
PI-PD$\beta 2$: 1-Phosphatidylinositol 4,5-bisphosphate phosphodiesterase β 2
SM-PDD : Sphingomyelin phosphodiesterase D
PD : Phosphodiesterase olei02445
GDPD : Glycerophosphodiester phosphodiesterase

PI-PLC-X domains from the CATH alignment
PLC_LISMO ([*Lm*PIPLC], **2plcA00**), PLC_STAAE ([*Sa*PIPLC], **4i90A00**), PLCB2_HUMAN (**2zkmX03**).
PI-PLC-X domains added
PLC_BACCE ([*Bc*PIPLC], **1gymA00**), PLCD1_RA (**1djgA02**), GNAQ_MOUSE (**3ohmB03**), A0BLC4_PARTE, A0BLM0_PARTE, A2QEJ2_ASPNC, A3LNB7_PICST, A4HCK9_LEIBR, A4I5X4_LEIIN, A7ASE0_BABBO, A7RKT2_NEMVE, A7RPQ0_NEMVE, A7RXP6_NEMVE, A7S279_NEMVE, A7SDH9_NEMVE, A7SXK4_NEMVE, B4H106_DROPE, B6KHA6_TOXGV, E9AEI1_LEIMA, F1R8S3_DANRE, F6HKS4_VITVI, F7DT82_CALJA, G3UXP4_MOUSE, H3CD03_TETNG, H3D5C7_TETNG, H3DBV7_TETNG, H3DF29_TETNG, I1R783_ORYGL, K4A9N8_SETIT, PLC1_SCHPO, PLCB2_RAT, PLCB_CAEEL, PLCD3_ARATH, PLCD4_RAT, PLCE1_HUMAN, PLCH1_HUMAN, PLCH2_HUMAN, PLCZ1_BOVIN, PLCZ1_CHICK, Q013R6_OSTTA, Q0UE70_PHANO, Q16W08_AEDAE, Q16ZC0_AEDAE, Q385H1_TRYB2, Q4DUP6_TRYCC, Q4QBH9_LEIMA, Q59LC4_CANAL, Q5A0L4_CANAL, Q5QIB9_STRPU, Q6CSI9_KLULA, Q7PW25_ANOGA, Q7QHE3_ANOGA, Q8IJR0_PLAF7, Q9XWB7_CAEEL, W4XJU2_STRPU.

Table 3. Set of 58 PI-PLC-X domains aligned to the hmm profile and kept for the analysis. The names refer to the entry names used by UniProt [14]. PDB structures are in bold. Proteins studied by the Reuter group are square brackets.

4. Error bar measurements

Standard error (SE) is used in histogram plots for measuring how variable the mean is. [16]

Results and Discussion

1. The amino acid composition of PH domains conforms with the current textbook model of peripheral protein association

The known Interfacial binding sites (IBS) of PH domains are located in the β_7 structural element and the β_1-β_2 segment. The amino acid compositions of β_7 and the β_1-β_2 segment are compared to that of the *solvent side*. We decided to enlarge the analysis of the β_1-β_2 segment from β_1 to β_2. It appears that small, tiny, polar and negatively charged amino acids are more frequent on the *solvent side* than on the IBS, while hydrophobic, aliphatic and aromatic amino acids are more frequent on the IBS than on the *solvent side*. Moreover, positively charged amino acids are more present on the β_1-β_2 segment than on the *solvent side*. (see **Figure 3**) These results are consistent with the IBS prototype of peripheral proteins that suggests a more important number of positively charged, hydrophobic, aliphatic and aromatic amino acids on the membrane-binding regions.

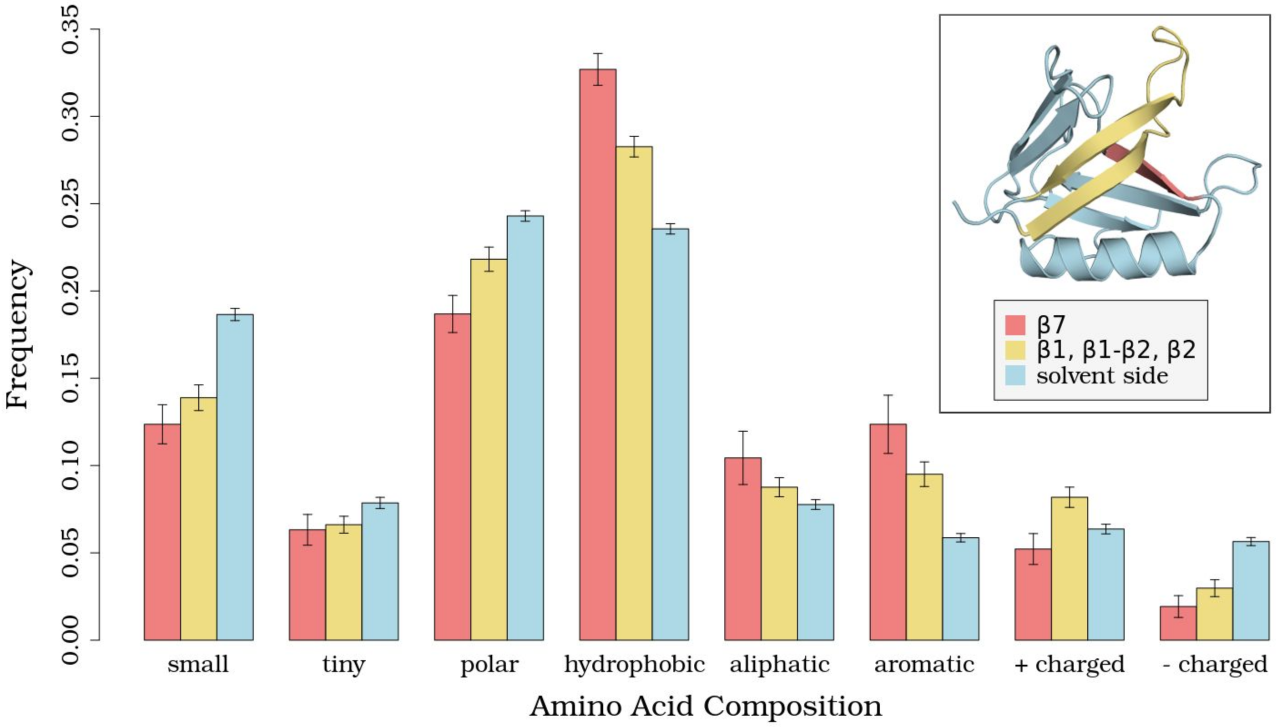


Figure 3. Amino acid composition of the IBS (β_7 and β_1-β_2 segment) and the *solvent side* of 27 PH domains. Prototype : 1dynA00. Amino acids grouped properties : Small : D, N, S, T, C, A, V, P & G; Tiny : S, C, A & G; Polar : R, K, D, E, Q, N, H, S, T, Y, C, W; Hydrophobic : K, H, T, Y, C, W, A, I, L, M,, F, V & G: Aliphatic : I, L & V; Aromatic : Y, W, F & H; Positively charged : R, K & H; Negatively charged : D & E.

2. A more challenging analysis of the amino acid composition of PI-PLC-X

The Reuter group has shown that the IBS of *Bc*PI-PLC, *Lm*PI-PLC and *Sa*PI-PLC are mainly localized near the first two segments β_1-α_1 and β_2-α_2, the α_7 structural element and the β_7-α_7 segment. This analysis is focused on the β_1-α_1 and β_2-α_2 IBS segments because only these are part of the PI-PLC-X domain. The barrel particular topology of these proteins (see **Figure 2B**) and the experimentally data on the location of IBS residues on barrel loops led us to consider the β_i-α_i segments as the *membrane side* and the α_i-β_{i+1} as the *solvent side*. Moreover, we observe that the β_1-α_1 and β_2-α_2 IBS segments are longer than β_3-α_3, α_1-β_2 and α_2-β_3 segments (see **Figure 4**). Consequently, our strategy is to compare the amino acid composition of β_1-α_1 to β_2-α_2 and the *solvent side*. Concerning the β_2-α_2 segment, hydrophobic, aromatic and positively charged amino acids are more frequent than on the *solvent side*. These results are consistent with the IBS prototype, even if there is a lack of aliphatic amino acids. Concerning the β_1-α_1 segment, only the aromatic amino acids are slightly more frequent than on the *solvent side*. Otherwise, the amino acid composition of the β_1-α_1 segment does not fit with the IBS prototype.

It is also interesting to add the comparison of the β_3-α_3 segment which was not identified as an IBS by the experimental studies, although it is close to the IBS segments and could possibly interact with membrane as well. We expect the frequencies for β_3-α_3 to be almost similar of those for the *solvent* *side*. However by analyzing the results (see **Figure 5**), this is not the case for most amino acid groups. We can not exclude that β_3-α_3 could be an IBS for the other proteins of the dataset, but the amino acid frequencies are not exactly what we would expect for a membrane-binding segment. We observe that aromatic and positively charged amino acids are more frequent than on the *solvent side*, but there are too large differences between β_3-α_3 and the *solvent side* for the polar, hydrophobic and aliphatic amino acid frequencies. These results could be due to a low number of amino acids (see **Figure 4**, β_3-α_3), thus it is difficult to conclude.

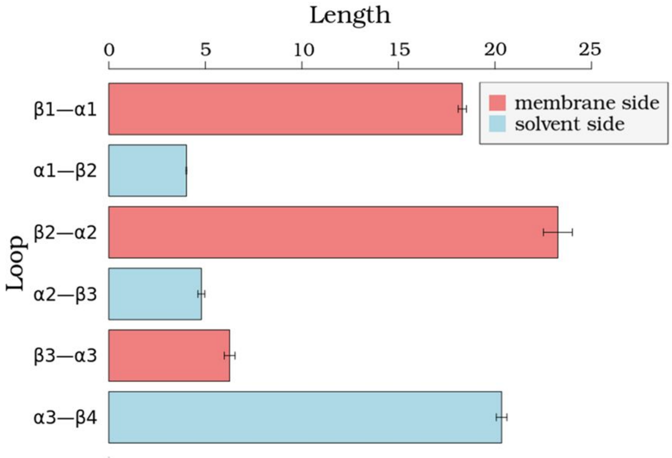


Figure 4. Length (number of amino acids) of 6 segments of 58 PI-PLC-X domain. β_1-α_1 and β_2-α_2 are the IBS.

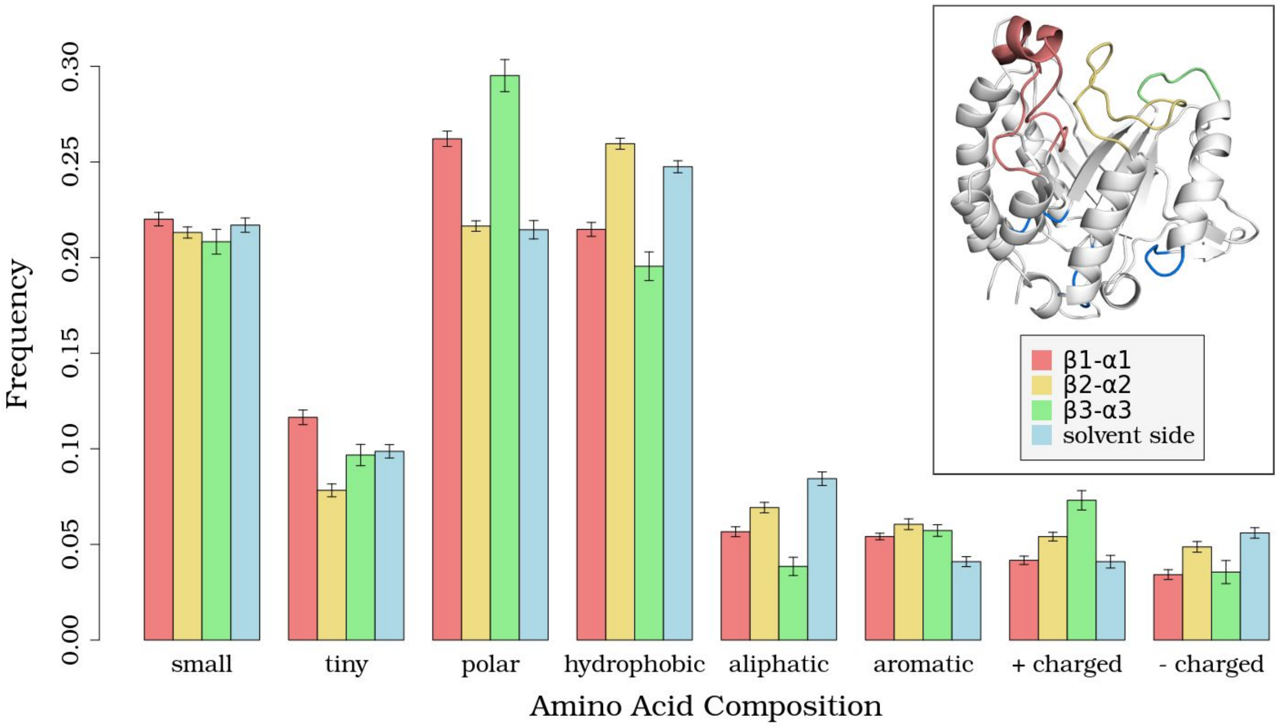


Figure 5. Amino acid composition of 6 different segments of 58 PI-PLC-X domains. The β_1-α_1, β_2-α_2, β_3-α_3 segments are in the membrane side. Prototype : 2plcA00. See details of the amino acid groups in **Figure 3** legend.

Conclusion

We have shown that the amino acid composition of PH domains is close to the textbook model of peripheral proteins which describes interfacial binding sites as containing hydrophobic, aliphatic, aromatic and positively charged amino acids. These regions are experimentally known as being on the β_7 structural element and the β_1-β_2 segment for a set of PH domains. It was more difficult to conclude that the regions identified as membrane-binding for some PI-PLC proteins by the Reuter group previous studies follow the theoretical model. Through two IBS segments studied, only β_2-α_2 may be close to the theoretical model. However, we could improve these results by reduced the analysis to an α-helix presents in the β_1-α_1 segment and experimentally proven as being an important element for the membrane-binding of *Bc*PI-PLC, therefore the amino acid composition should be closer to the model of peripheral proteins. Another idea is to compare the amino acid composition of peripheral proteins to non-membrane proteins, *i.e.* choosing a different reference dataset. Further, the PI-PLCs are enzymes and their function might require a different membrane-binding site as they have to bind and unbind more frequently to find new substrates.

References

1. Lucke M. *Membrane structural biology: with biochemical and biophysical foundations*. Cambridge University Press, 2014.

2. Grauffel C, Yang B, He T, Roberts M. F, Gershenson A and Reuter N. "Cation– π interactions as lipid-specific anchors for phosphatidylinositol-specific phospholipase C." *Journal of the American Chemical Society* 135.15 (2013): 5740-5750.

3. Python Software Foundation. Python Language Reference, version 3.0. http://python.org

4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE. "The Protein Data Bank" *Nucleic Acids Research* 28 (2000): 235-242. http://rcsb.org

5. The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC. https://pymol.org

6. Kabsch W and Sander C. "Dictionary of protein secondary structure: pattern recognition of hydrogen‐bonded and geometrical features." *Biopolymers* 22.12 (1983): 2577-2637.

7. Eddy SR. "Profile hidden Markov models." *Bioinformatics* 14.9 (1998): 755-763. http://hmmer.org/

8. Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, Orengo CA and Sillitoe I. "CATH: an expanded resource to predict protein function through structure and sequence." *Nucleic acids research* 45.D1 (2017): D289-D295. http://cathdb.info/

9. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. "The Pfam protein families database: towards a more sustainable future." *Nucleic acids research* 44.D1 (2015): D279-D285. https://pfam.xfam.org/

10. Rice P, Longden I and Bleasby A. "EMBOSS: The European Molecular Biology Open Software Suite" *Trends in Genetics* 16.6 (2000):276-277

11. Vonkova I, Saliba AE, Deghou S, Ellenberg J, Bork P, Gavin AC and al. "Lipid cooperativity as a general membrane-recruitment principle for PH domains." *Cell reports* 12.9 (2015): 1519-1530.

12. Nozomi N, Orengo CA, and Thornton JM. "One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions." *Journal of molecular biology* 321.5 (2002): 741-765.

13. Tiwari SP and Reuter N. "Similarity in shape dictates signature intrinsic dynamics despite no functional conservation in TIM barrel enzymes." *PLoS computational biology* 12.3 (2016): e1004834.

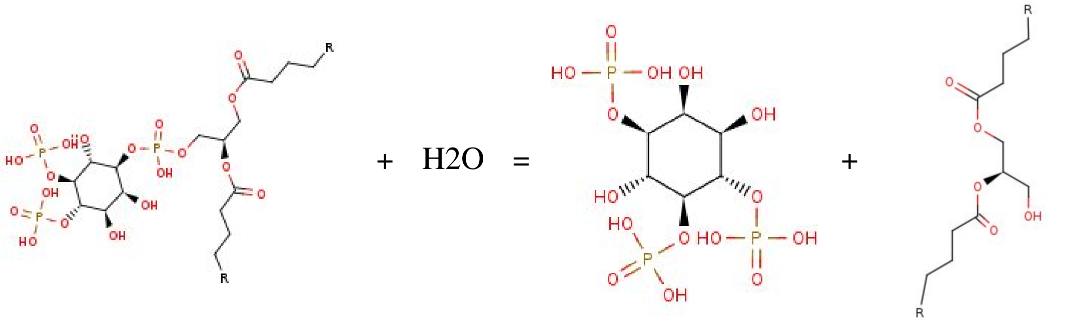
14. Chen C, Huang H, and Wu CH. "Protein bioinformatics databases and resources." *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics* (2017): 3-39. https://uniprot.org

15. Lomize MA, Pogozheva ID, Joo H., Mosberg HI and Lomize AL. "OPM database and PPM web server: resources for positioning of proteins in membranes." *Nucleic acids research* 40.D1 (2012): D370-D376. http://opm.phar.umich.edu/

16. Cumming G, Fidler F and Vaux DL. "Error bars in experimental biology." *The Journal of cell biology* 177.1 (2007): 7-11.

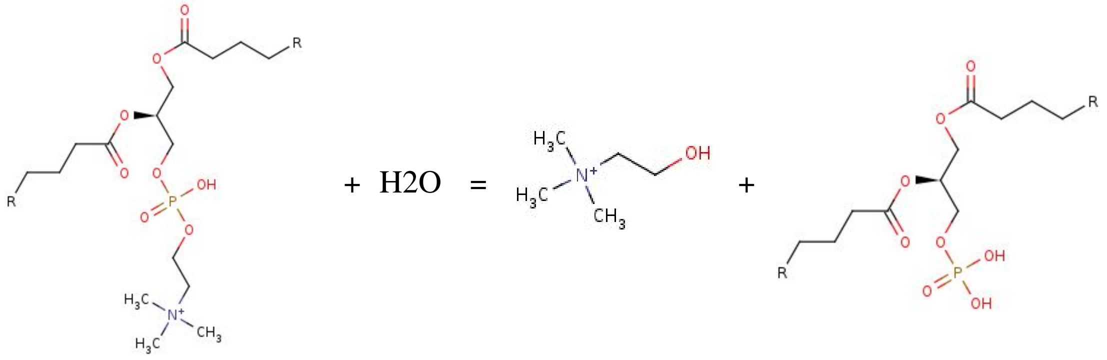
Supplementary datas

E.C.3.1.4.11 - Phosphoinositide phospholipase C



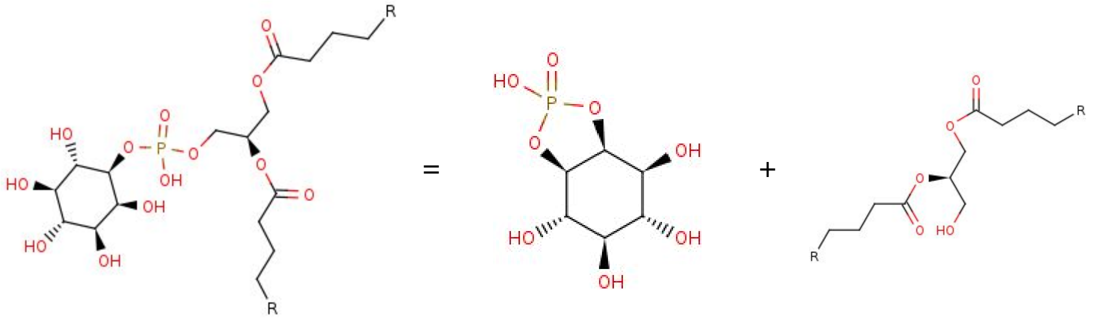
1-phosphatidyl-1D-myo-inositol + H2O = 1D-myo-inositol 1,4,5-triphosphate + diacylglycerol

E.C.3.1.4.4 - Phospholipase D



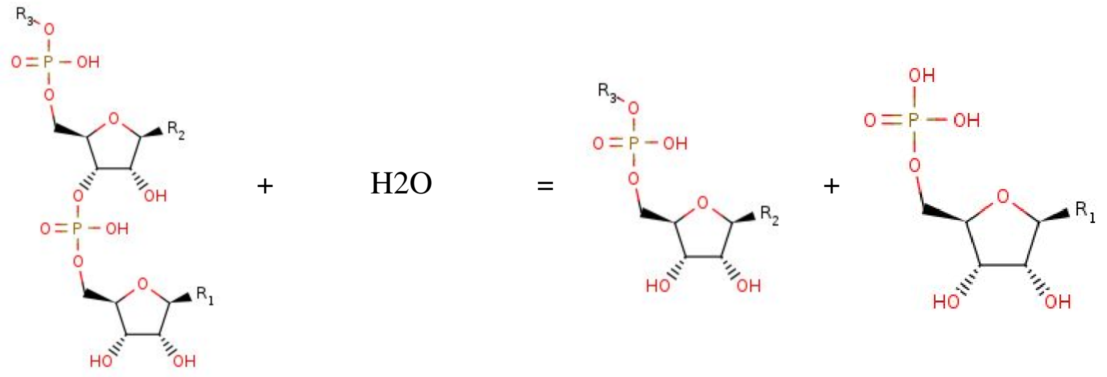
a phosphatidylcholine + H2O = choline + a phosphatidate

E.C.4.6.1.13 - Phosphatidylinositol diacylglycerol-lyase



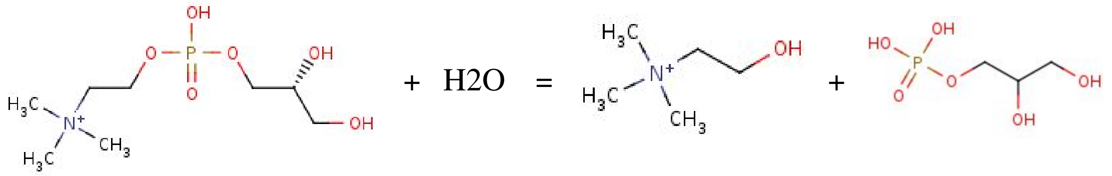
1-phosphatidyl-1D-myo-inositol = 1D-myo-inositol 1,2-cyclic phosphate + 1,2-diacyl-sn-glycerol

E.C.3.1.4.1 - Phosphodiesterase I



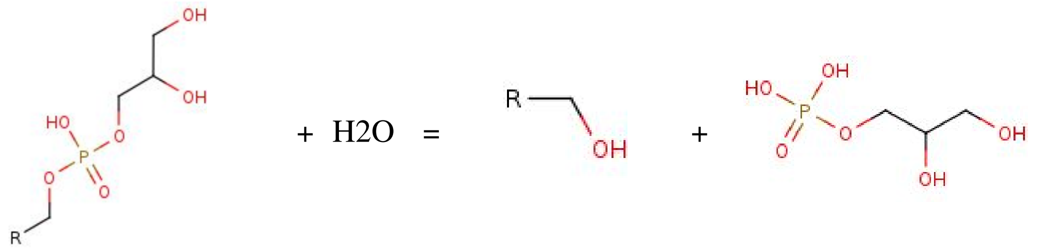
[nucleotide]n + H2O = [nucleotide]m + nucleotide

E.C.3.1.4.2 - Glycerophosphocholine phosphodiesterase



sn-glycero-3-phosphocholine + H2O = choline + sn-glycerol 3-phosphate

E.C.3.1.4.46 - Glycerophosphodiester phosphodiesterase



a glycerophosphodiester + H2O = a primary alcohol + sn-glycerol 3-phosphate

Abstract

The current textbook model of peripheral proteins suggests a higher number of hydrophobic, aliphatic, aromatic and positively charged amino acids on the interface binding sites (IBS) of peripheral proteins than on non-membrane proteins in general. It has recently been shown that some peripheral proteins, which for the different membrane-binding residues contributions have been studied in detail, do not fit this model. These discoveries call for a better mapping of IBS on peripheral proteins. However only few proteins have reliable annotations, consequently this represents a challenge for large-scale mapping of amino acid composition. Here we present the amino acid composition of two protein families : the PH domains and the PI-PLC-X domains. Although we find a composition of amino acids close to the textbook model for the PH domains, the composition of the PI-PLC-X domains were more challenging to analyze and the results are not conclusive. This study was possible thanks to the development of a pipeline which performs an inventory of the amino acids present on the IBS and non-membrane-binding regions of protein domains using a structural alignment augmented with sequence data.

L’actuel modèle théorique des protéines périphériques suggère un nombre plus important d’acides aminés hydrophobes, aliphatiques, aromatiques et chargés positivement sur le ou les sites d’interaction à la membrane de protéines périphériques plutôt que sur les protéines non membranaires en général. Il a été montré récemment que certaines protéines périphériques, dont les contributions des différents résidus se liant à la membrane ont été étudié en détail, ne correspondent pas au modèle décrit. Ces découvertes appellent à une meilleure compréhension des interfaces de liaison membranaire des protéines périphériques. Cependant, seulement quelques protéines ont des annotations fiables de leur site d’interaction membranaire, par conséquent cela représente un challenge pour recenser la composition en acides aminés à grande échelle. Nous présentons ici la composition en acide aminés de deux familles de protéines : Les domaines PH et les domaines PI-PLC-X. Bien que la composition en acides aminés des domaines PH soit proche du modèle théorique, l’analyse des domaines PI-PLC-X représente un défi, car les résultats ne sont pas concluants. Cette étude a été permise par le développement d’un pipeline permettant l’inventaire des acides aminés présents sur l’interface de liaison membranaire et sur les régions non liantes de domaines protéiques, à partir d’un alignement de structure augmenté par des séquences de protéines dont la structure est inconnue.