Pablo Mier* and Miguel A. Andrade-Navarro

Avoided motifs: short amino acid strings missing from protein datasets

<https://doi.org/10.1515/hsz-2020-0383>

Received December 9, 2020; accepted February 19, 2021;

published online March 2, 2021

Abstract: According to the amino acid composition of natural proteins, it could be expected that all possible sequences of three or four amino acids will occur at least once in large protein datasets purely by chance. However, in some species or cellular context, specific short amino acid motifs are missing due to unknown reasons. We describe these as Avoided Motifs, short amino acid combinations missing from biological sequences. Here we identify 209 human and 154 bacterial Avoided Motifs of length four amino acids, and discuss their possible functionality according to their presence in other species. Furthermore, we determine two Avoided Motifs of length three amino acids in human proteins specifically located in the cytoplasm, and two more in secreted proteins. Our results support the hypothesis that the characterization of Avoided Motifs in particular contexts can provide us with information about functional motifs, pointing to a new approach in the use of molecular sequences for the discovery of protein function.

Keywords: amino acid motif; avoided motif; cellular location; protein sequence analysis.

Introduction

The identification of over-represented residue patterns in biological sequences is a crucial step in the discovery of protein functionality (Edwards et al. 2007; Rigoutsos and Floratos 1998; Ye et al. 2007). Conversely, if a functional motif is deleterious or dangerous in some way for a species, it may as well be under-represented or depleted in its proteome.

Avoidance of motifs was applied to the study of negative results in nucleotide sequences from bacterial genomes revealing, for example, the palindromic motifs characteristic of restriction enzymes that are used by bacteria against viral DNA (Fuglsang 2003; Gelfand and Koonin 1997). More recently, this approach has also been applied in several other genomes (Poirot et al. 2019; Sadovsky et al. 2017).

A particular chain of amino acids may not be found in a proteome, or protein dataset, for several reasons. First of all, in a small dataset there might be simply not enough amino acid combinations to find some motifs even by chance. For example, there are $20^5 = 3,200,000$ possible sequence combinations of length five amino acids. If we take the proteome of *Escherichia coli* (4391 proteins, 1,336,798 overlapping motifs of five amino acids), not finding a five amino acid pattern does not necessarily mean that it was negatively selected in evolution, or avoided, but that the proteome of *E. coli* does not have a sufficient amount of sequences to find it by chance.

A second possibility is that a particular amino acid combination is rarely used, or not used at all, because it interferes with the protein folding. Amino acid frequencies are not completely position-independent, as described by the neighbor effect (Xia and Xie 2002), because amino acids influence their neighbors to fold in a secondary structure (Borguesan et al. 2017). Particular motifs could be negatively selected in proteins, and thus avoided, not because of their functionality but for their structural properties.

Last, if the functionality of a motif can be deleterious in a given context, then its absence in a large number of sequences can bring it to the light. For example, let us assume that there exists a functional motif present in one nuclear protein that triggers its cleavage by another protein. The presence of that motif in other nuclear proteins would result in them being cleaved, which is an undesired result. Therefore, the motif would be generally avoided by nuclear proteins. But not by proteins situated in other subcellular locations. Thus, detecting that this motif is avoided in nuclear proteins could point to its functionality.

Following this idea, here we identify the shortest motifs missing from the human proteome and from bacterial proteins, paying special attention to those expected to be found many times, and discuss their possible function by analysis of the sequences in other species where they are present. Our work presents a theoretical approach to detect

*Corresponding author: Pablo Mier, Faculty of Biology, Institute of Organismic and Molecular Evolution, Johannes Gutenberg University Mainz, Hanns-Dieter-Hüsch-Weg 15, D-55128 Mainz, Germany, E-mail: munoza@uni-mainz.de. <https://orcid.org/0000-0003-3663-2352>

Miguel A. Andrade-Navarro, Faculty of Biology, Institute of Organismic and Molecular Evolution, Johannes Gutenberg University Mainz, Hanns-Dieter-Hüsch-Weg 15, D-55128 Mainz, Germany. <https://orcid.org/0000-0001-6650-1711>

and characterize avoided motifs that could have a potential function; the relevance of our approach requires experimental follow-up verification, which we facilitate by providing categorized sets of avoided motifs.

Materials and methods

Protein sequence retrieval

The complete SwissProt release 2020_01 (561,911 proteins) and the reference proteome of *Homo sapiens* (74,811 proteins) were retrieved from UniProtKB release 2020_01 (UniProt Consortium 2019).

Subcellular location of human proteins was taken into account to generate four datasets of proteins located in a specific subcellular location: nuclear (3459 proteins), cytoplasm (1534 proteins), membrane (1817 proteins) and secreted (1437 proteins). Proteins located in more than one location were discarded.

Identification of avoided motifs

Given an input protein dataset, we counted the occurrences of all overlapping sequence fragments of length 2, 3 and 4 amino acids. Avoided Motifs (AMs) are defined as amino acid combinations that are not found.

Estimation of expected number of hits

Amino acid background frequency was calculated per protein dataset. Also, the total number of subsequences found for each length. For length = 4, this number is 25,448,103 for the human proteome and 200,487,979 for SwissProt. For a given motif, the expected number of hits in a dataset was estimated as the product of the individual amino acid background frequencies, times the number of total motifs (e_1). We use as an example the human Avoided Motif (hAM) 'CPMF'. The expected frequency of the motif computed in this way is $1.098\text{e-}06$ (frequencies are 0.022, 0.063, 0.022 and 0.036, for C, P, M and F, respectively) times the number of four amino acid subsequences in the human proteome, equals 28.0, which would be the expected number of hits. This calculation assumes that amino acids occur independently.

A more realistic estimation of the expected number of hits was performed using the frequencies of the pairs of consecutive amino acids forming the motif (e_2), to account for the neighbor effect (Xia and Xie 2002). We calculated these frequencies for all our datasets. Following this alternative procedure, the expected number of hits (e_2) for the hAM 'CPMF' in the human proteome was 26.4.

Motif pattern retrieval

A total of 1311 motif patterns were downloaded from PROSITE release 2020_02 (Sigrist et al. 2013). To select patterns with certain level of specificity, we discarded those with more than one consecutive uncertainty "x", and those having not even one amino acid specific position. We also discarded patterns with less than four amino acids (the

length of the studied AMs), and very long patterns (longer than 14 amino acids). These filtering steps left us with 120 unique patterns.

Results and discussion

Identification of human avoided motifs

Research regarding protein sequences is usually focused on the actual sequences, rather than on what is not on them. The identification of amino acid motifs missing from large protein datasets may indicate functional motifs that for some reason are avoided in them. For this purpose, we looked in the human proteome for all possible amino acid combinations, starting from motifs of two amino acids. We found all possible combinations of lengths two and three, discarding the presence of Avoided Motifs. However, there were 209 human Avoided Motifs (hAMs) of length four (Supplementary File S1). That means there are 209 combinations of just four amino acids that are not found in the human proteome.

To compute the expected frequencies of hAMs, we used two measurements. One is the product of the frequencies of the individual amino acids composing the motif (e_1 , see Methods). But this follows the wrong assumption that amino acids appear in a sequence independently from their neighboring residues. As a more realistic measurement, we also used the product of the observed frequencies of consecutive amino acid pairs (e_2). This second measurement may not be accurate when dealing with small backgrounds but this is difficult to assess. As a pragmatic solution we used both measurements and compared the values. Differences between the values will be discussed, otherwise, amino acid pair expectation (e_2) will be used.

The values of these two measurements for the 209 hAMs expected to be found in the human proteome at least once indicate that most of those with high e_1 have lower e_2 , suggesting that there could be some structural component in the selection of hAMs (Figure 1a). However, this does not happen generally since there are some cases with high e_1 values and $e_1 < e_2$. In fact, for the most expected hAM, 'CPMF', the e_1 and e_2 values were very similar (26.4 and 28.0 times, respectively).

hAMs are highly enriched in tryptophan, methionine and cysteine residues (Figure 1b), the three less abundant amino acids in the human proteome (with frequencies of 1.24, 2.19, and 2.21%, respectively). These amino acids are metabolically expensive to produce, but they also have atypical side chains that make them difficult to place into structures. Other amino acids that are also not frequent, such as tyrosine and histidine (2.61% each), are not as

Figure 1: Identification of human avoided motifs.

(a) Expected number of hits in the human proteome per hAM (human avoided motif) of four amino acids, calculated using the product of the frequencies of individual amino acids ($e1$) and of amino acid pairs ($e2$); hAMs showed when they are expected to be more than 15 times using either measure. Red line indicates $x = y$. (b) Amino acid frequency per hAM position, and background frequency per amino acid in the human proteome.

highly enriched in the hAMs. Enriched residues in hAMs are mostly position-independent, except for cysteine. These residues seem to be more enriched at the N-terminal of the hAMs than at the C-terminal. On the other hand, there are no leucine residues in any hAM, and they are highly depleted in alanine, glycine, arginine and serine.

Subcellular location specific avoided motifs

To find out if AMs are associated to biological function, and since function is often specific to subcellular location, we repeated the previous analysis (starting from a length of two amino acids) considering proteins that are annotated as present in just one subcellular location. For example, there could be a motif for protein cleavage that should be present only in secreted proteins and it would be avoided in cytoplasmic proteins to avoid their undesired cleavage. We focused on four distinct subcellular locations: nucleus, cytoplasm, membrane and secreted.

This time we found no AMs of length two amino acids, but two AMs of length three in the cytoplasm ('MCW' and 'WHW') and secreted ('CMW' and 'TMW') datasets, but with relatively low numbers of expected hits: 5.45 for 'MCW', 3.63 for 'WHW', 3.96 for 'CMW', and 9.83 for 'TMW'. In

general, lower $e1/e2$ values suggest that our observations could be just due to chance.

To try a different strategy to assess the significance of these subcellular specific three amino acid motifs, we studied in detail the most expected AM, the secreted Avoided Motif (secAM) 'TMW', and looked for it in the complete human proteome. We found it 275 times, in 273 different proteins (102 from SwissProt, 171 from TrEMBL), when 380.3 hits were expected. This motif is enriched in human transmembrane proteins: 43% of these proteins (118/273) have at least one transmembrane (TM) region, compared to the 16% of human proteins with them (11,954/74,811) (a 2.7 fold enrichment). Regarding the orientation of the protein, from the 118 TM proteins containing the motif, it was 46 times in an outside loop, 53 inside, with the remainder 19 cases overlapping the TM. We believe there may be some functionality behind the motif 'TMW' that explains this enrichment, probably related to protein transport to membranes, so it would be negatively selected in secreted proteins, but this function would not be related to the orientation of the protein in the membrane (e.g. used as a mechanism to detect wrongly inserted proteins). To the best of our knowledge, 'TMW' has not been described yet as a functional motif. Examination of protein structures in the PDB database show that

We found many subcellular-specific AMs of length four: 16,696 in nucleus, 24,494 in cytoplasm, 21,339 in membrane and 33,575 in secreted proteins. Here, we report those expected to occur 10 or more times ($e2 \geq 10$): 104 in nucleus, six in cytoplasm, 11 in membrane, and 17 secreted (Supplementary Figure S1, Supplementary File S2).

Our next strategy to characterize AMs is to study their species specificity. AMs due to structural reasons (e.g. complicating protein folding) should be avoided in every species, whereas functional AMs, for example, reserved by one species to recognize proteins from pathogens and thus avoided in own proteins, should be more species specific. To find out if some hAMs fall in this category, we evaluated the presence of hAMs in the comprehensive dataset of annotated proteins in SwissProt.

We observed that hAMs have a large variation in the ratio $e2$ in SwissProt versus the number of actual hits (Figure 2a). Many hAMs are more than 2-fold under-represented, suggesting a structural reason. Interestingly, some hAMs such as ‘WMHD’, ‘WWMV’ and ‘HHMW’ are more than 2-fold over-represented in SwissProt compared

Figure 2: Characterization of human avoided motifs.

(a) Number of found and expected (e2) hits in SwissProt per hAM. Blue, red and green lines indicate a $x = 2y$, $x = y$ and $2x = y$ value, respectively. (b) Taxonomic distribution of proteins with hAMs, for all hAMs found in SwissProt at least 50 times.

Figure 3: Characterization of bacterial avoided motifs.

(a) Expected number of hits in SwissProt_bacteria per bAM (bacterial avoided motif) of four amino acids, calculated using the combination of the frequencies of amino acids ($e1$) and of amino acid pairs ($e2$); bAMs showed when they are expected to be more than 20 times using either measure. Red line indicates $x = y$. (b) Amino acid frequency in hAMs and bAMs, compared to the background frequency in the complete human proteome and SwissProt_bacteria datasets.

to what was expected given their frequencies of amino acid pairs. These would not seem to have negative structural effects, suggesting that their absence in human proteins might have a functional explanation.

Most hAMs are found in bacteria, and many are absent or very infrequent in eukaryotic species (Figure 2b), which is hinting at a mechanism of peptide recognition specific to eukaryotes. An intriguing result is obtained for hAM 'MWMD', found overall 112 times in SwissProt, almost 75% of them in viral proteins. They are all in protein PA-X from different strains of the Influenza A virus, involved in the suppression of the host antiviral and immune responses (Hayashi et al. 2015). There are two versions of the PA-X protein, one full-length and one truncated, given by the length of the C-terminal X-ORF (61 or 41 amino acids, respectively). The truncation of the protein is due to a TGG(Trp)-to-TAG(stop) mutation in position 232 of the protein (Shi et al. 2012), leading to the loss of the last 20 amino acids of the PA-X protein, which results in increased viral replication and pathogenicity (Gao et al. 2015). The mutated Trp residue is in the second position of the hAM 'MWMD'. The elimination of the AM in the shorter version could explain its increased pathogenicity.

Identification of bacterial avoided motifs

Most of the hAMs can be found in bacterial proteins from SwissProt. We wondered if we could find bacterial Avoided Motifs (bAMs) in all SwissProt proteins from bacteria (334,482 proteins), and how similar these would be from hAMs.

We identified 154 bAMs of length four amino acids (Supplementary File S1), and none of length two or three. Most of them (149/154) were expected to be found more than once in the dataset (Figure 3a), with the extreme case of 'CIMY' ($e1 = 43.6$ and $e2 = 28.6$). They are enriched in Trp, Met and Cys residues, as are hAMs (Figure 3b), but they are all position-independent (data not shown). Cysteines, in fact, are almost three times more prevalent in bAMs than in hAMs.

There are 18 shared Avoided Motifs (hbAM) (Supplementary File S1). All of them were found in SwissProt less than 10 times, except for 'NMWC' and 'CHWY' (found 22 and 11 times, respectively). Intriguingly, the hbAM 'NMWC' is present in the C-terminal region of protein KC1G3 (Casein kinase I isoform gamma-3) from *Mus musculus* (UniProt:Q8C4X2, positions 405–408) and *Pongo abelii* (UniProt:Q5R4V3,

positions 437–440). There is a human SwissProt entry for KC1G3 (UniProt:Q9Y6M4), highly similar to both orthologous proteins (Supplementary File S3), but lacking the motif ‘NMWC’. The motif is neither functionally annotated nor structurally resolved for any of the proteins containing it.

AM functionality

As discussed above, a motif may be avoided, or not selected, in a proteome but functional in another. The functionality of an AM could be the reason of its avoidance in the proteome in which it is not present. To assess this possibility in a high-throughput manner, we performed a comprehensive functional analysis of the found hAMs and bAMs by means of a comparison to known motif patterns from the PROSITE database.

To find AMs specifically matching a pattern, we considered PROSITE motifs in which at least three residues from the AM were met in the motif, allowing a maximum of one uncertainty (‘x’), but not a single mismatch.

We found 17 AMs matching six PROSITE patterns (Supplementary File S4). Only one AM is matched unequivocally, the human avoided motif ‘WEWW’ to the pattern PS00481: [RK]-E-C-T-G-L-x-W-E-W-W-[RK]. hAM

‘NWEW’ also matches it. This PROSITE pattern is the signature of thiol-activated cytolysins (or cholesterol-binding cytolysins, CBC), toxins secreted by taxonomically diverse species of gram-positive bacteria that attack humans in different ways (Billington et al. 2000; Rossjohn et al. 1997), involved in membrane binding and intercalation (Jacobs et al. 1999). We studied how frequent in the human proteome are all possible motifs of length four amino acids that match the pattern PS00481. Both the number of expected hits (e_2) and found hits decrease drastically from the N-terminal to the C-terminal part of the pattern (Figure 4).

Our interpretation of this result is that there may be a molecular process against thiol-activated cytolysins, and that the avoidance of hAM ‘WEWW’ in human proteins has a protective effect against this mechanism. In fact, mutational studies show that the hemolytic activity of thiol-activated cytolysins carrying this hAM vastly rely on the last two tryptophan residues (95 and 99.9% decrease in hemolytic activity, respectively) (Michel et al. 1990), which explains why a mechanism aimed at destroying this bacterial protein would target such a motif. Experimental work must be carried out to support this finding, for example, by performing a cell viability assay after introducing the ‘WEWW’ motif in a human essential protein. In the complete SwissProt database there are only eight eukaryotic proteins with this motif (three from Viridiplantae, two from Fungi, two from Metazoa, and one other eukaryota), indicating that this protective mechanism may be shared by other eukaryotes.

Conclusions

The absence of a sequence in a well-characterized and completely sequenced proteome is used in comparative analyses to infer protein functionality, or the lack of it, in a species. Here we replicated this process in a different scale, focusing on identifying missing short protein motifs, or Avoided Motifs (AMs), from protein datasets.

We identified 209 human AMs and 154 bacterial AMs (Supplementary File S1). They are enriched in Trp, Met and Cys residues, some of the less abundant amino acids in these datasets. However, almost all of the AMs were expected to be found by chance in the database at least once, some of them even 43 times. Here we followed a theoretical approach defining an Avoided Motif as strictly avoided, that is, it must not be found in the whole dataset. Easing this cutoff and using a less stringent one to define AMs (i.e. allowing one or a few hits, depending on the dataset size) would increase the number of found AMs but complicate

Figure 4: Motif occurrence of bacterial pattern PS00481 in the human proteome.

Number of expected hits (e_2) and found hits (in log10 scale) in the human proteome per motif of length four amino acids from PROSITE pattern PS00481. The two hAMs found within this pattern, ‘NWEW’ and ‘WEWW’, are located in the C-terminal part of the pattern.

the analysis. At this point, to excite interest in our approach, we decided to limit the analysis for simplicity. Targeted searches for AMs (for example, restricting them to smaller taxonomic divisions, or introducing restrictions in the sequence space to scan) may be used in the future to engage in further searches of AMs (for example of length five amino acids) without expanding the number of AMs for analysis to unmanageable numbers.

We have shown here a few examples of putative functionality associated to some of these AMs, meaning that at least for them their absence may indeed be an active avoidance. The concept of Avoided Motifs itself opens the possibility to characterize under-represented amino acid combinations that could turn out to be functional. Our intention here is to provide guidelines for experimental work about hundreds of these motifs with a significant anomalous distribution: occurring in thousands of proteins but avoided in precise sets of species and/or cellular locations.

Figure S1: Number of expected hits per subcellular location-specific avoided motif.

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: This work was supported with funds from the Deutsche Forschungsgemeinschaft [AN735/4-1 to M.A.A.N.].

Conflict of interest statement: The authors declare that they have no conflict of interests.

References

- Billington, S.J., Jost, B.H., and Songer, J.G. (2000). Thiol-activated cytolytins: structure, function and role in pathogenesis. *FEMS Microbiol. Lett.* 182: 197–205.
- Borguesan, B., Inostroza-Ponta, M., and Dorn, M. (2017). NIAS-Server: neighbors influence of amino acids and secondary structures in proteins. *J. Comput. Biol.* 24: 255–265.
- Edwards, R.J., Davey, N.E., and Shields, D.C. (2007). SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS One* 2: e967.
- Fuglsang, A. (2003). Distribution of potential type II restriction sites (palindromes) in prokaryotes. *Biochem. Biophys. Res. Commun.* 310: 280–285.
- Gao, H., Sun, H., Hu, J., Qi, L., Wang, J., Xiong, X., Wang, Y., He, Q., Lin, Y., Kong, W., et al. (2015). Twenty amino acids at the C-terminus of PA-X are associated with increased influenza A virus replication and pathogenicity. *J. Gen. Virol.* 96: 2036–2049.
- Gelfand, M.S. and Koonin, E.V. (1997). Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. *Nucleic Acids Res.* 25: 2430–2439.
- Hayashi, T., MacDonald, L.A., and Takimoto, T. (2015). Influenza A virus protein PA-X contributes to viral growth and suppression of the host antiviral and immune responses. *J. Virol.* 89: 6442–6452.
- Jacobs, T., Cima-Cabal, M.D., Darji, A., Mendez, F.J., Vazquez, F., Jacobs, A.A., Shimada, Y., Ohno-Iwashita, Y., Weiss, S., and de los Toyos, J.R. (1999). The conserved undecapeptide shared by thiol-activated cytolytins is involved in membrane binding. *FEBS Lett.* 459: 463–466.
- Michel, E., Reich, K.A., Favier, R., Berche, P., and Cossart, P. (1990). Attenuated mutants of the intracellular bacterium *Listeria monocytogenes* obtained by single amino acid substitutions in listeriolysin O. *Mol. Microbiol.* 4: 2167–2178.
- Poirot, O., Jeudy, S., Abergel, C., and Claverie, J.M. (2019). A puzzling anomaly in the 4-mer composition of the giant pandoravirus genomes reveals a stringent new evolutionary selection process. *J. Virol.* 93: e01206–e01219.
- Rigoutsos, I. and Floratos, A. (1998). Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics* 14: 55–67.
- Rossjohn, J., Feil, S.C., McKinstry, W.J., Tweten, R.K., and Parker, M.W. (1997). Structure of a cholesterol-binding, thiol-activated cytolytin and a model of its membrane form. *Cell* 38: 685–692.
- Sadovsky, M., Fontaine, J.F., Andrade-Navarro, M.A., Yakubailik, Y., and Rudenko, N. (2017). Lost strings in genomes: what sense do they make? In: Rojas I. and Ortuño F. (Eds.), *Bioinformatics and biomedical engineering. IWBBIO 2017. Lecture notes in computer science*, Vol. 10209. Springer, Cham, pp. 20–29.
- Shi, M., Jagger, B.W., Wise, H.M., Digard, P., Holmes, E.C., and Taubenberger, J.F. (2012). Evolutionary conservation of the PA-X open reading frame in segment 3 of influenza A virus. *J. Virol.* 86: 12411–12413.
- Sigrist, C.J.A., de Castro, E., Cerutti, L., Cuče, B.A., Hulo, N., Bridge, A., Bougueleret, L., and Xenarios, I. (2013). New and continuing developments at PROSITE. *Nucleic Acids Res.* 41: D344–D347.
- UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47: D506–D515.
- Xia, X. and Xie, Z. (2002). Protein structure, neighbor effect, and a new index of amino acid dissimilarities. *Mol. Biol. Evol.* 19: 58–67.
- Ye, K., Kusters, W.A., and Ijzerman, A.P. (2007). An efficient, versatile and scalable pattern growth approach to mine frequent patterns in unaligned protein sequences. *Bioinformatics* 23: 687–693.

Supplementary Material: The online version of this article offers supplementary material (<https://doi.org/10.1515/hsz-2020-0383>).