Polar profile of random proteins and their incidences in Uniprot Database

C. Polanco

Department of Mathematics, Faculty of Sciences, Universidad Nacional Autónoma de México, 04510 México.

In order to deepen in the polymerization mechanism of proteins, we considered the possibility of nature having a random pattern to design proteins. With this in mind, we built, evaluated, and compared two groups of proteins with different level of randomness Random proteins, and Pseudo random proteins. The degree of randomness was statistically verified for both groups, their fragments were searched in Uniprot Database and their electromagnetic profile was evaluated with the Polarity Index Method. Our results show that the "random pattern" of the protein fragments, in both groups, has a presence (measured with the number of hits) that is far from being an isolated or accidental case. It was particularly observed that the total fragments of the Pseudo Random proteins found in Uniprot Database was 10 times greater than the total fragments observed in the Random proteins.

Keywords: proteomics; structural bioinformatics; random proteins; pseudo random proteins; electromagnetic profile; synthetic proteins; catastrophic bifurcation points; supervised programs; quantitative structure–activity relationship models.

1. Introduction

Polymerization is an eminent chemical process where different monomers group together to form a protein, whose primary structure is an orderly succession of amino acids that adopts a characteristic three-dimensional conformation. There are two factors that have an important impact in protein formation: polymerization and electrical charge of monomers, although they are also influenced by the amino acids concentration, temperature, and acidity of the medium. All these factors bias the formation of a protein. In this work, two groups of proteins called **Synthetic proteins** were computationally built, these groups had a different degree of randomness that was verified with the same statistical test, and the electronegativity profile was determined for both of them. Additionally, all **Synthetic protein** fragments were searched in Uniprot Database [1], with Blast software [2], finding that these synthetic protein fragments are very high. When their electronegativity graphs (Fig. 1) were checked, it was found that they exhibit a large number of inflection points. The Synthetic protein group is formed by two types of proteins: **Pseudo Random proteins**, whose monomers were chosen from 20 possible amino acids, weighted, and grouped by their electrical charge in four polarity groups; and **Random proteins**, whose monomers had an equal probability of being chosen. Both groups of proteins were evaluated with the supervised program Polarity Index Method (PIM) [3-19], that only takes for its metric the electromagnetic profile of the protein linear sequence (Evaluation of electromagnetic profile section).

2. Material and Methods

This work takes two different groups of **synthetic proteins** computationally generated with different level of randomness (Table 1, paragraphs 1, 2), at a later stage, they were sought in Uniprot Database [1], with all their possible fragments.

2.1 Evaluation of electromagnetic profile

The supervised method used here, (PIM), has been previously used to identify different groups of proteins [3-19]. For a review of its use and restrictions, it is recommended to check *Polarity index in Proteins - A Bioinformatics Tool* [19, Appendix to Computational Tool).

2.2 Catastrophic Bifurcation Points

The identification of regularities in the graph of a function, includes the points where the function exhibits maximum, minimum, changes in concavity, and fractality; the inflection points, where it is observed a change in the geometric regularity of the graph before and after, are known as Catastrophic Bifurcation Points [22-24].

2.3 Data acquisition

The groups of **Pseudo random proteins** (Table 1, paragraph 2) and **Random proteins** (Table 1, paragraph 1), were built with a different degree of randomness (Table 1, paragraphs 1 and 2), and they were statistically verified (Table 1, paragraph 4).

2.4 Test plan

The set of tests (Table 2) applied to both groups **Pseudo random proteins** (Table 1, paragraph 2), and **Random proteins** (Table 1, paragraph 1), was the same.

Table 1 Data Acquisition.

#	Description
1	***Random proteins***. *Methodology*: the proteins with **RND** prefix were built forming a random succession of amino acids, using a script linux (Supplementary Materials section; polanco@unam.mx). from a set of twenty letters {H, K, R, D, E, C, G, N, Q, S, T, Y, A, F, I, L, M, P, V, W}, these ten proteins are: RND010, RND020, RND030, RND040, RND050, RND060, RND070, RND080, RND090, and RND100 (Table 5).
2	***Pseudo Random proteins***. *Methodology*: the proteins with **PRND** prefix were built forming a random succession of amino acids with a *Linear Congruential Generator* [20] from a set of 20 letters {H, K, R, D, E, C, G, N, Q, S, T, Y, A, F, I, L, M, P, V, W}, pondering the generation of amino acids with the corresponding proportion of its electrical charge, according to the numerical equivalence: (1) basic hydrophilic = {H, K, R}; (2) acidic hydrophilic = {D, E}; (3) neutral = {C, G, N, Q, S, T, Y}; and (4) non-polar = {A, F, I, L, M, P, V, W} (polarity bias). These ten proteins are PRND010, PRND020, PRND030, PRND040, PRND050, PRND060, PRND070, PRND080, PRND090, and PRND100 (Table 4).
3	**Catastrophic bifurcation points**. *Methodology*: these points were computationally located with a program designed for this purpose, written in Fortran 77 (Supplementary Materials section; polanco@unam.mx).
4	**Kolmogorov–Smirnov test**. *Methodology*: the tests in both groups were carried out with a computational implementation designed for this purpose (Supplementary Materials section; polanco@unam.mx) based on the Kolmogorov-Smirnov test [21], with a p-value 0.01.
List of tests carried out to the ***Pseudo random protein*** and ***Random protein*** *groups*.	

Table 2 Test Plan.

#	Description
1	The level of randomness in the ***Pseudo random protein***, (Table 4), and ***Random protein*** *groups* (Table 5) were verified with the Kolmogorov–Smirnov test (Table 1).
2	The identification of fragments in both, Pseudo random protein (Table 4) and Random protein groups (Table 5), was carried out using Uniprot-BLAST method [1]. Target database: UniprotKB E-Threshold: 10 Matrix: Auto Filtering: None Gapped: Yes Hits: 250).
3	The relative frequencies of the 16 polarity interactions (electronegativity profile) were plotted as smooth curves for both groups **Pseudo random proteins**, (Table 4) and **Random proteins** (Table 5).
4	The catastrophic bifurcation points (Supplementary Materials section; polanco@unam.mx) of each protein were computationally identified in each group **Pseudo random proteins**, (Table 4) and **Random proteins** (Table 5).
5	Both groups, **Pseudo random proteins** (Table 4) and **Random proteins** (Table 5), were evaluated with the PIM (Evaluation of electromagnetic profile section), calibrating the PIM with one group and applying it to both groups (Table 3).
List of tests carried out to the ***Pseudo random protein*** and ***Random protein*** groups.	

3. Results

Eight out of ten **Pseudo random proteins** (Table 4, Test column), influenced by the bias (Table 1, paragraph 2), were statistically rejected (Table 1, paragraph 4). Ten out of ten **Random proteins** (Table 5, test column), not affected by the electromagnetic bias (Table 1, paragraph 1), were statistically accepted (Table 1, paragraph 4).

Table 3 Hits Synthetic Proteins.

Group	Pseudo random proteins	Random proteins
Pseudo random proteins	**70**	0
Random proteins	0	**80**

Hits (%) of the PIM for ***Pseudo random proteins*** and ***Random proteins*** (Test plan section). The PIM calibrated with each group (rows) is compared with the groups (columns).

The number of **Pseudo random protein** fragments, (Table 4, Number of similar fragments column), found in Uniprot Database [1], is ten times less than the number of **Random protein** fragments (Table 5, Number of similar fragments column) found in the same database.

The number of catastrophic bifurcation points (Supplementary Materials section; polanco@unam.mx) found in each protein of both groups, **Pseudo random proteins** (Table 4, Number of CBP in the sequence column), and **Random proteins** (Table 5, Number of CBP in the sequence column), is similar.

Table 4 Pseudo Random Proteins.

#	ID	Sequence	Test	Number of similar fragments located in Uniprot Database	Number of CBP in the sequence	PIM	
						PRP	RP
1	PRND010	YHVEEEKIDY	✔	2	1	×	×
2	PRND020	YKCEVMQGWHYDDDHHPGME	✔	5	1	×	×
3	PRND030	YDDDHHPGMEQFDYEDFRNFRRREADKGYR	×	21	1	✔	×
4	PRND040	QFDYEDFRNFRRREADKGYRVKYGYRDEPWCDFEYVERVL	×	57	2	✔	×
5	PRND050	RRREADKGYRVKYGYRDEPWCDFEYVERVLYHYICEDYRAKWLDRRIPKL	×	2	3	×	×
6	PRND060	VKYGYRDEPWCDFEYVERVLYHYICEDYRAKWLDRRIPKLWHYVHKEYRFYNYVRGCRQY	×	1	5	✔	×
7	PRND070	CDFEYVERVLYHYICEDYRAKWLDRRIPKLWHYVHKEYRFYNYVRGCRQYEIDRQYDHHGRHIKNHDVDG	×	2	7	✔	×
8	PRND080	YHYICEDYRAKWLDRRIPKLWHYVHKEYRFYNYVRGCRQYEIDRQYDHHGRHIKNHDVDGDAKRYNKEYPRRRHNDENDR	×	17	6	✔	×
9	PRND090	KWLDRRIPKLWHYVHKEYRFYNYVRGCRQYEIDRQYDHHGRHIKNHDVDGDAKRYNKEYPRRRHNDENDREQKFCKYCYELNQWQYEMEM	×	56	8	✔	×
10	PRND100	WHYVHKEYRFYNYVRGCRQYEIDRQYDHHGRHIKNHDVDGDAKRYNKEYPRRRHNDENDREQKFCKYCYELNQWQYEMEMFVEIEHFLKGMEFGHAQYLM	×	59	9	✔	×

Pseudo Random proteins. PIM: The PIM calibrated with each group (PRP/RP) compared with the groups (columns). (✔): Protein accepted by the PIM. (×): Protein not accepted by the PIM. (Test plan section). Number of similar fragments located in Uniprot Database [1]: Number of fragments found in Uniprot Database. Test: (✔): Protein accepted by the Kolmogorov–Smirnov test. (×): Protein not accepted by the Kolmogorov–Smirnov test (p-value = 0.01, Table 1).

The PIM efficiently discriminated both groups (Table 3): **Pseudo random proteins** (Table 1, paragraph 2) and **Random proteins** (Table 1, paragraph 1). Individually, The PIM calibrated with the **Pseudo Random proteins**, identified seven out of ten proteins from this group and any of the **Random protein group**. When the PIM was calibrated with the **Random protein** group, it identified eight out of ten **Random proteins** and any **Pseudo Random proteins** (Table 5 PIM column).

Table 5 Random Proteins.

#	ID	Sequence	Test	Number of similar fragments located in Uniprot Database	Number of CBP in the sequence	PIM	
						PRP	RP
1	RND010	HMEVTIPFNS	✔		1	×	✔
2	RND020	WAPECSCHIGAPKDTQFYYK	✔		1	×	✔
3	RND030	FCCCHRVYPSALGISCMDILEEETLGHYFD	✔	3	1	×	✔
4	RND040	ELAMGDGENGFWPNQKIGGMEMIHLHQL GVTASFYLDMQP	✔		2	×	✔
5	RND050	ELAMGDGENGFWPNQKIGGMEMIHLHQL GVTASFYLDLGPKSRHQSCNDF	✔	2	3	×	✔
6	RND060	DGVCVIVKSYFDHFWNHMRIPEEHCEHNP VLAHNNCLPGNPFDAIGWECVMSKGKFHF K	✔	8	6	×	✔
7	RND070	FAPDLVQLGHLDGFPPGMKPMMGPKDKQ GKHGAHHELGNKKNPFEVQFEMLEKHFK QDGQPTVNYQMRDV	✔	6	8	×	×
8	RND080	SDGHETEMWMKHQYADLMAAMMALLD GNMNMAPPIFYFAAKDCMDEQPMICMIEN ICFAYGGHQKNMIYHENFCMFMMFI	✔		9	×	✔
9	RND090	CICHGMADIDDLHKEILSSHDYNVMKEKW CDSHFQEICKTAIQDLHNIMYDREGFHGKG KHEFTKNPNLFNEMGPDMFCGQEWPSKAP CI	✔	2	8	×	×
10	RND100	GHDHLFDAMFIALVTEDCRNGACFAIIRAH QPHFSRDNAMMMSSECEAFIMSDILEPDG MMCEHDIFLGFHGNMTRVCAAMHKGPEC QDAAYPINGAHDD	✔	4	10	×	✔

Random proteins. PIM: The PIM calibrated with each group (PRP/RP) compared with the groups (columns). (✔): Protein accepted by the PIM. (×): Protein not accepted by the PIM. (Test plan section). Number of similar fragments located in Uniprot Database: Number of fragments found in Uniprot Database. Test: (✔): Protein accepted by the Kolmogorov–Smirnov test. (×): Protein not accepted by the Kolmogorov–Smirnov test (p-value = 0.01, Table 1).

The geometric representation of the electromagnetic profile (Fig. 1) of the **Random proteins** (Table 1, paragraph 1) did not match with the electromagnetic profile of the **Pseudo Random proteins** (Table 1, paragraph 2) in any of the 16 polarity interactions.

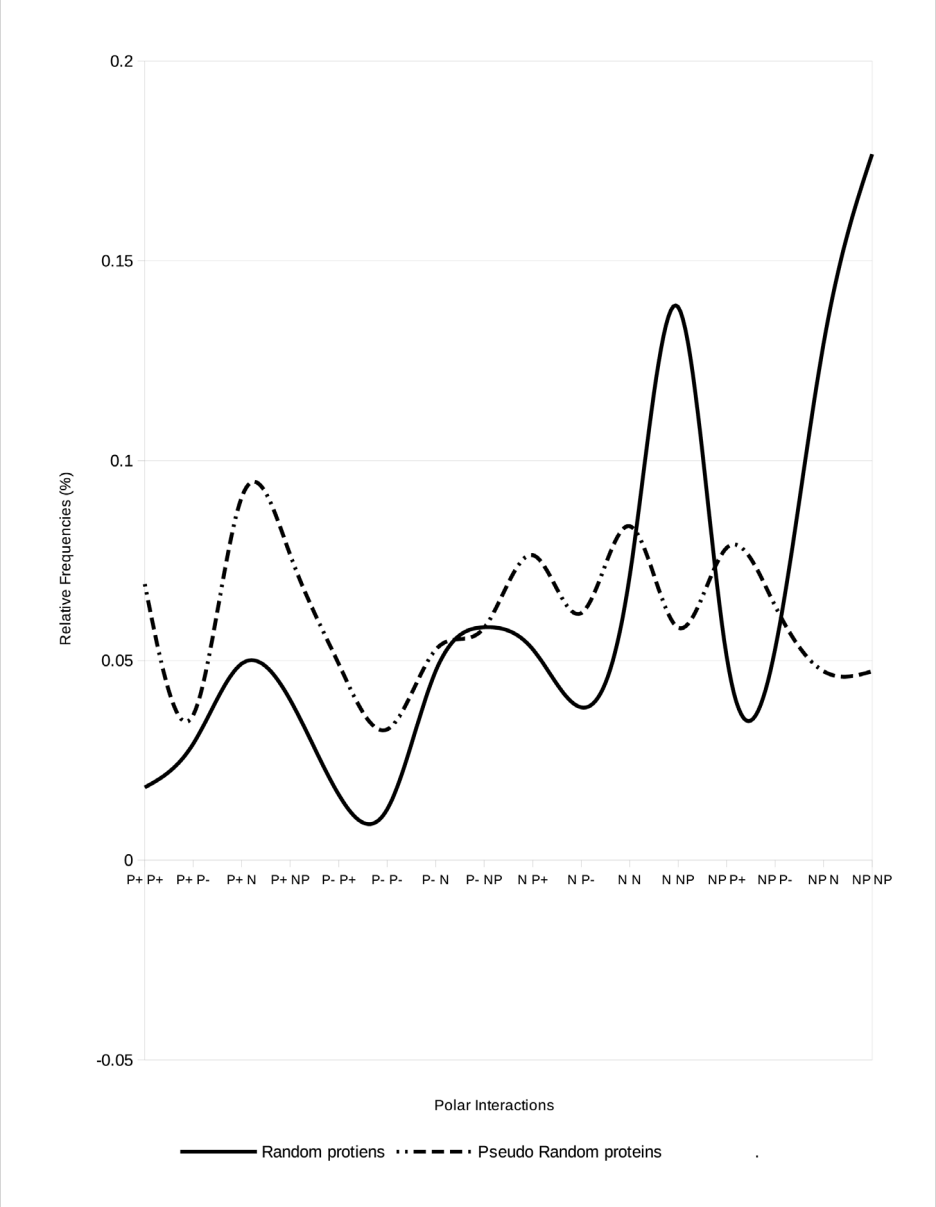

Fig. 1 Polar frequency distribution for the ***Random protein*** and ***Pseudo Random protein*** groups. The X-axis represents the 16 polarity interactions (Evaluation of electromagnetic profile section).

4. Discussion

There are ten times more matching fragments from Pseudo Random proteins in Uniprot Database [1] than matching fragments from Random proteins. The Pseudo Random proteins were built so the polarity profile influenced them, it was assumed that the "typical profile" of the proteins found in nature would resemble more the "Pseudo Random pattern" and less the "Random pattern", however, this was not so. Several studies characterizing different protein groups from their electromagnetic profile [3-19], extracted from diverse databases, showed that the Pseudo Random protein profile was the "typical profile". In fact, it was not expected to find such a high number of matching fragments from synthetic proteins, it was assumed that the "random profile" would be a new and rare protein group. Future work will be focused on searching the "random patterns" in all the proteins registered in Uniprot Database, in order to know how large the dissemination of this regularity is in the proteins found in nature.

5. Conclusions

The analysis of the regularities in the two groups of **synthetic proteins** built with a different level of randomness, whose fragments found in Uniprot Database showed high incidence, leads to the assumption that the **random pattern** is used by nature in the protein polymerization process.

Availability The scripts, supplementary figures/materials/data, and source codes are available as request to the corresponding author (polanco@unam.mx).

Conflict of Interests We declare that we do not have any financial and personal interest with other people or organizations that could inappropriately influence (bias) our work.

Author Contributions Theoretical conception and design: CP. Computational performance: CP. Data analysis: CP. Results discussion: CP.

Acknowledgments It is a pleasure to thank Concepción Celis Juárez whose suggestions, and proof-reading have greatly improved the original manuscript.

References

[1] Uniprot Consortium. Uniprot: a hub for protein information. Nucleic Acids Res. 43(Database issue)2015;:D204-12.
[2] Altschul SF, Gish W, Miller W, Myers EW. & Lipman DJ. Basic local alignment search tool J Mol Biol. 1990; 215:403-410.
[3] Polanco C, Buhse T, Castañón-González JA, Samaniego JL. Possible computational filter to detect proteins associated to influenza A subtype H1N1. Acta Biochimica Pol. 2014; 61:693-698.
[4] Polanco C, Buhse T, Samaniego JL, Castañón-González JA, Arias Estrada M. Computational model of abiogenic amino acid condensation to obtain a polar amino acid profile. Acta Biochim. Pol. 2014;61:253-258.
[5] Polanco C, Buhse T, Samaniego JL, Castañón-González JA. A toy model of prebiotic peptide evolution: the possible role of relative amino acid abundances. Acta Biochim. Pol. 2013;60:175-182.
[6] Polanco C, Buhse T, Samaniego Mendoza JL, Castañón-González JA. Detection of selective antibacterial peptides by the Polarity Profile method. Acta Biochim Pol. 2013;60:183-189.
[7] Polanco C, Castañón-González JA, Buhse T, Uversky VN, Zonana-Amkie R. Classifying lipoproteins based on their polar profiles. Acta Biochim Pol. 2016;3:235-241 DOI: 10.18388/abp.2014-918.
[8] Polanco C, Castañón-González JA, Uversky VN, Buhse T, Calva JJ. Electronegativity and intrinsic disorder of preeclampsia-related proteins. Acta Biochim Pol; 2016; [Head to print].
[9] Polanco C, Castañón-González JA, Uversky VN. (Letter to the Editor) Buhimschi IA, Nayeri UA, Zhao G, Shook LL, Pensalfini A, Funai EF, Bernstein IM, Glabe CG, Buhimschi CS. Protein misfolding, congophilia, oligomerization, and defective amyloid processing in preeclampsia. Sci Transl Med. 2014;6:245ra92 DOI:10.1126/scitranslmed. 3008808.
[10] Polanco C, Samaniego JL, Buhse T, Castañón González JA. Discrete dynamic system oriented on the formation of prebiotic dipeptides from Rode's experiment. Acta Biochim Pol. 2014;61:717-726.
[11] Polanco C, Samaniego JL, Buhse T, Mosqueira FG, Negron-Mendoza A, Ramos-Bernal S, Castanon-Gonzalez JA. Characterization of Selective Antibacterial Peptides by Polarity Index. International Journal of Peptides 2012;2012:58502, http://dx.doi.org/10.1155/2012/585027
[12] Polanco C, Samaniego JL, Castañón-González JA, Buhse T, Arias-Estrada M. Computational model of abiogenic amino acid condensation to obtain a polar amino acid profile. Acta Biochim Pol. 2014;61:253-258.
[13] Polanco C, Samaniego JL, Castañón-González JA, Buhse T, Sordo ML. Characterization of a possible uptake mechanism of selective antibacterial peptides. Acta Biochim Pol. 2013;60:629-633.
[14] Polanco C, Samaniego JL, Castañón-González JA, Buhse T. Polar Profile of Antiviral Peptides from AVPpred Database. Cell Biochem Biophys. 2014;70:1469-1477 DOI: 10.1007/ s12013-014-0084-4.
[15] Polanco C, Samaniego JL, Uversky VN, Castañón-González JA, Buhse T, Leopold-Sordo M, Madero-Arteaga A, Morales-Reyes A, Tavera-Sierra L, González-Bernal JA, Arias-Estrada M. Identification of proteins associated with amyloidosis by polarity index method. Acta Biochim Pol. 2015;62:41-55.
[16] Polanco C, Samaniego-Mendoza JL, Buhse T, Castañón-González JA, Leopold-Sordo M. Polar Characterization of Antifungal Peptides from APD2 Database. Cell Biochem Biophys. 2014; 70:1479-1488 DOI: 10.1007/s12013-014-0085-3.
[17] Polanco C, Samaniego-Mendoza JL, Castañón-González JA, Buhse T (Letter to the Editor) Howard SJ, Hopwood S, Davies SC. Antimicrobial Resistance: A Global Challenge Sci. Transl. Med. 2014 DOI:10.1126/scitranslmed.3009315.
[18] Polanco C. (2016a) Identification of antimicrobial peptides using eigenvectors. Acta Biochim Pol. 2016;63:483-91. DOI: 10.18388/abp.2015_993.
[19] Polanco C. (2016d) Polarity index in Proteins-A Bioinformatics Tool DOI: 10.2174/9781681082691116010l eISBN: 978-1-68108-270-7, ISBN:978-1-68108-269-1 Bentham Science Publishers Sharjah, U.A.E.
[20] Knuth DE. The Art of Computer Programming, Volume 2: Seminumerical Algorithms, Third Edition. Addison-Wesley, 1997. ISBN 0-201-89684-2. Section 3.2.1: The Linear Congruential Method, pp. 10-26.
[21] Slegel S. Estadística no paramétrica, aplicada a las ciencias Ed. Trillas, 9 ed. 1985 pp. 69-74.
[22] Dobzhansky T. Genetics and the origin of species (Columbia Univ. Press, New York); 2nd Ed., 1941; 3rd Ed., 1951.
[23] Guckenheimer J. Review: René Thom, Stabilité Structurelle et Morphogénèse, Essai d'une Théorie Générale des Modèles. Bull. Amer. Math Soc 1973;79:878-890 http://projecteuclid.org/euclid.bams/1183534961.
[24] Poincaré H. Sur l'équilibre d'une masses fluide animée d'un mouvement de rotation. Acta math. 1885;7:259-380.