ACTA ABP BIOCHIMICA POLONICA

Vol. 63, No 2/2016
229–233
http://dx.doi.org/10.18388/abp.2014_919

Review

# Structure and function relationships of proteins based on polar profile: a review

Carlos Polanco[1]✉, Thomas Buhse[2] and Vladimir N. Uversky[3,4,5]

[1]Department of Mathematics, Faculty of Sciences, Universidad Nacional Autónoma de México, C.P. 04510 D.F., México; [2]Centro de Investigaciones Químicas, Universidad Autónoma del Estado de Morelos, C.P. 62209 Chamilpa, Cuernavaca, Morelos, México; [3]Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, MDC07 Tampa, FL 33647, USA; [4]Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia; [5]Laboratory of Structural Dynamics, Stability and Folding of Proteins, Institute of Cytology, Russian Academy of Sciences, St. Petersburg, Russia

**Proteins in the post-genome era impose diverse research challenges, the main are the understanding of their structure-function mechanism, and the growing need for new pharmaceutical drugs, particularly antibiotics that help clinicians treat the ever- increasing number of Multidrug-Resistant Organisms (MDROs). Although, there is a wide range of mathematical-computational algorithms to satisfy the demand, among them the Quantitative Structure-Activity Relationship algorithms that have shown better performance using a characteristic training data of the property searched; their performance has stagnated regardless of the number of metrics they evaluate and their complexity. This article reviews the characteristics of these metrics, and the need to reconsider the mathematical structure that expresses them, directing their design to a more comprehensive algebraic structure. It also shows how the main function of a protein can be determined by measuring the polarity of its linear sequence, with a high level of accuracy, and how such exhaustive metric stands as a "fingerprint" that can be applied to scan the protein regions to obtain new pharmaceutical drugs, and thus to establish how the singularities led to the specialization of the protein groups known today.**

**Key words**: amino acids, polarity profile, pattern recognition, dynamical systems theory, atherosclerosis, selective antibacterial peptides, intrinsically disordered proteins

**Received**: 15 October, 2014; revised: 19 January, 2016; accepted: 23 February, 2016; available on-line: 08 April, 2016

## INTRODUCTION

In Proteomics, the Supervised learning (Larrañaga *et al*., 2006) essentially seeks to identify a regularity (Oestreicher, 2007) among a group of proteins with a particular characteristic or "training data", once this regularity is isolated, a mathematical-computational algorithm is built (Kitaev, 1997) to find the same regularity or the absence of it in other groups of proteins. The best scenario is when the desired regularity is evident in a group of proteins, however, that is not usually the case as the efficiency of algorithms is frequently low; this occurs particularly when trying to identify the primary function of a protein. Proteins do not normally have a unique action associated to them, i.e. only 1% of the peptides located in APD2 database (Wang *et al*., 2009) have a unique pathogenic action. The search of a non-evident regularity, such as the main function associated to a protein, cannot be done by finding similarities in the protein linear sequence, like sequence alignment algorithms do e.g. BLAST (Madden *et al*., 1996) or FASTA (GenBank, 2011). It requires strategies where it is possible to identify minimal regularities. Within the group of Supervised learning algorithms there are some that focus on relating chemical structure to biological activity, evaluating only one physico-chemical property and obtaining the best results in the identification of the main action or function of a protein, these algorithms are called Quantitative structure-activity relationship models (QSAR models) (Putz *et al*., 2011). The more than 80 QSAR algorithms known (Qureshi *et al*., 2014) use physico-chemical metrics involving the linear representation and/or the 3D structure of the protein and evaluate one or more properties simultaneously. What differentiates each QSAR model is the metrics they use, however, all of them produce a real value from a predetermined range, e.g. isoelectric point (Kosmulski, 2009) at 25°C for tungsten (VI) oxide WO3 in water: [0.2–0.5]. At first glance, the greater number of physico-chemical properties, lesser the number of "false positives", but this is not true here, there are QSAR models that include all known physico-chemical properties in their metrics (Yap, 2011), and yet the false positives still occur, with the percentage of efficiency not exceeding 80% in most of the models (Brendel *et al*., 1992). The probable cause is that when the result comes from a predetermined range, the completeness property of the real numbers is not considered, therefore, the combination of the physico-chemical properties does not add effectiveness to the algorithm, but it adds complexity to the computational implementation. A minimalist approach to the assessment of the physico-chemical properties can significantly improve the performance of a QSAR model, this approach consists of identifying the fundamental physico-chemical property influencing the studied phenomenon, and building a metric that expresses its dynamic and static behavior.

An example of this new family of QSAR models is polarity index method (Polanco *et al*., 2012), which assumes that the three-dimensional conformation of a protein defines its specific function and is the result of its electromagnetic balance. It also conjectures

✉e-mail: polanco@unam.mx

**Abbreviations**: MDROs, Multidrug-Resistant Organisms; QSAR models, Quantitative structure-activity relationship models

that this 3-D conformation is expressed in the linear sequence formed by its amino acids and that this balance can be measured through their polarity. For this purpose, amino acids are classified in four different groups: polar positively charged, polar negatively charged, polar neutral, and non-polar. If the amino acid sequence is read from N-terminal to C-terminal from left to right, moving one amino acid at a time and the 16 possible incidents are registered in an array, a comprehensive metric of the polar behavior of the protein will be obtained from this linear sequence. If this procedure is carried out with a training data, an array of polar incidents representative of that particular set will be generated. This array can be considered a "fingerprint" of the protein group studied and since this algorithm can simultaneously evaluate multiple proteins, it can be used for the polar classification of the existent protein groups (Boman, 1995), the exploration of peptide regions of a determined length, the construction of new pharmaceutical drugs from fully synthetic proteins, or in basic science, for discovering the profile of the first proteins from four billion years ago (Gaucher *et al*., 2010; Polanco *et al*., 2013; 2014; 2014b).

## FOUNDATION

The mathematical-computational algorithm called polarity index method (Polanco *et al*., 2012) had its foundations in the early studies this team did on polymerization of prebiotic proteins that had to be present 4 billion years ago (Gaucher *et al*., 2010), since it was not possible to use the current genetic code (Sharp, 1985), consisting of 20 amino acids, a random generation of amino acids from also randomly produced nucleotide triplets was used. It was observed that although the first amino acids did not correspond to the 20 amino acids known today, neither in number nor in type, it was possible to use the polar profile that was the result of the electromagnetic balance reached by each one of these amino acids, as this property was defined for all of them. This led to the construction of a polar equivalence (injective mapping) (Vinogradov, 1985) that allowed the comparison of the prebiotic proteins computationally built with those known today.

From this polar equivalence we obtained an array called "Polarity matrix" (Polanco *et al*., 2013b; Rabiner, 1989), that once normalized to one, represented the relative frequencies of all possible polar interaction; subsequently this matrix was expressed with a smooth curve. Here we present three groups that were identified by this method: (i) a set of lipoproteins related to atherosclerosis (Guyton & Klemp, 1989; Polanco *et al*., 2009), i.e. large VLDL, small dense LDL, and small DHDL subclasses (Koba *et al*, 2003), downloaded from UniProt Database (Magrane, 2011) (Fig. 1); (ii) the sets of natively unfolded proteins and natively folded proteins (Dunker *et al*., 2001; Polanco *et al*., 2015a; Uversky *et al*., 2000; 2002; 2008; 2008a; 2009; 2010; 2010a; 2010b; 2013; Wright & Dyson, 1999) from the supplementary material (Oldfield *et al*., 2005) (Fig. 2). Some of these proteins are known to produce serious pathological conditions, including the neurodegenerative diseases grouped under the term amyloidoses (Polanco *et al*., 2015a); and (iii) a set of selective cationic amphipathic antibacterial peptides (SCAAP) (Fig. 3) (Polanco & Samanie-

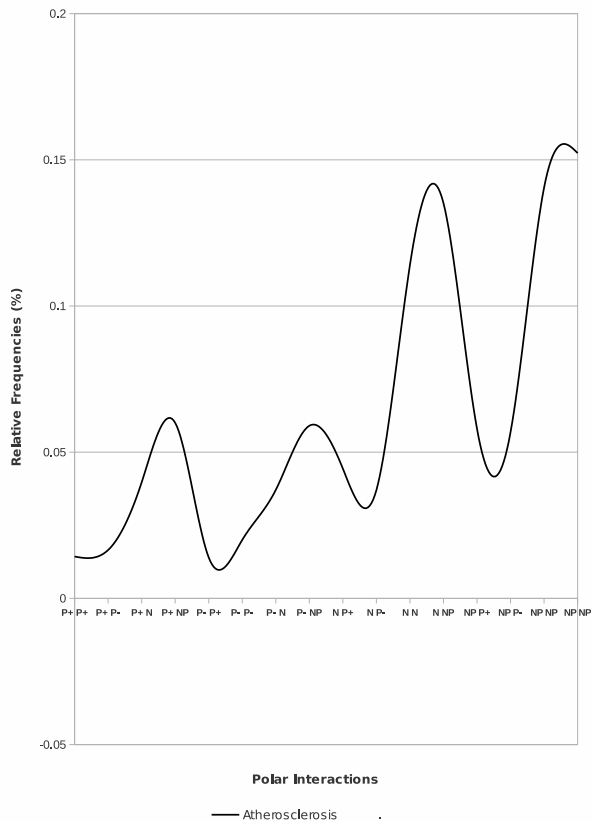


**Figure 1. Relative Frequency distribution for lipoproteins related to atherosclerosis (Polanco *et al.*, 2015). The X-axis represents the 16 polar interactions.**

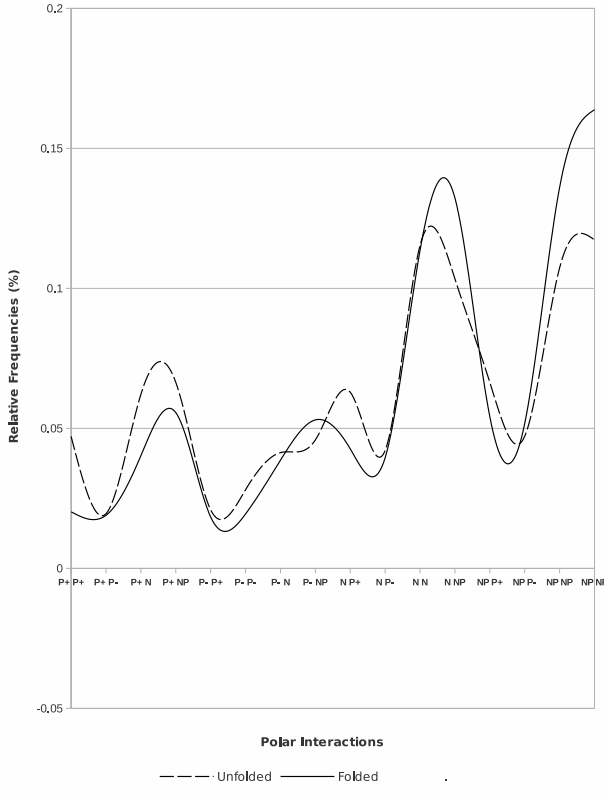


**Figure 2. Relative frequency distribution for the unfolded and folded proteins (Polanco *et al.*, 2015a). The X-axis represents the 16 polar interactions.**

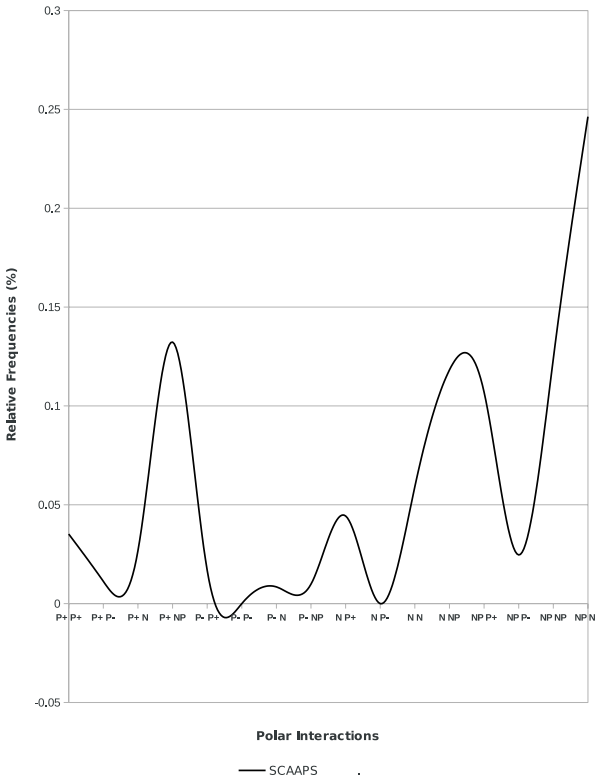


**Figure 3. Relative frequency distribution for the selective cationic amphipathic antibacterial peptides (SCAAP) (Polanco & Samaniego, 2009). The X-axis represents the 16 polar interactions.**

go, 2009) from (Polanco *et al*., 2013b). This small set of peptides is characterized by being highly toxic to bacterial membranes and present negligible toxicity to mammalian cells.

These groups of peptides and proteins (Figs. 1–3) do not have any coincidence in the minimum, maximum, or turning points. The intrinsically disordered protein group (Fig. 2) has similarities, however, there is a translation between the curves. The SCAAP group (Fig. 3), is substantially different from the others. The characteristics of the curves is typical for each group, and the proteins in each group are similar. This is the reason why the polarity profile is an effective discriminant for the functional (bacteria, fungi, virus, etc.) and structural groups (disordered proteins) studied.

The graphical representations used in different groups of peptides and proteins showed that the polarity matrix is neither symmetric nor antisymmetric (Munkres, 2000), we could verify that the inflection points (Munkres, 2000) located in the X-axis of the smooth curves characterize the group studied, i.e. the location of these points was an effective discriminant (80–90% in a double-blind statistical test), it was tested in more than 14 protein and peptide groups studied (Polanco & Samaniego, 2009; Polanco *et al*., 2012; 2013a; 2013b; 2014a; 2014c; 2014d; 2014e). However, we decided to choose the most accurate computational interpretation of the matrix and the analytical construction of the smooth curve presented a problem as the characteristic polynomial of the curve differed for each of the techniques used to obtain it.

If we consider these 16 polar interactions are the characteristics of the main action of a particular group of proteins, then the space where the protein is defined would be either $R^{16}$ (real field) or $C^8$ (complex field) (Munkres, 2000), and as the number of inflection points will always be minor to this number of incidents, the space where the discriminant property is defined will be a subspace of $C^8$. Furthermore, vector "x" whose 16 components are the elements of the polarity matrix, can be measured ||x|| (Munkres, 2000) therefore, every protein can be assigned to a group not only according to the similarities of their 16 components but also according to the length of the vector searched. From these considerations we can emphasize one aspect of the attributes of the discriminant — the feature that discriminates is identified by the singularities or inflection points (singularities degenerated), and not by the regularities or maximum-minimum points (singularities non-degenerated) observed in the smooth curve, this is how the physico-chemical property studied was identified (Polanco *et al*., 2012).

As mentioned before, the electromagnetic balance of the peptide or protein is classified into four polar groups (Pauling, 1960) closely related to the nature of the elements in all living matter, mainly formed by carbon (C), hydrogen (H), oxygen (O), and nitrogen (N). Therefore, the electromagnetic balance should be defined as a quantum electromagnetic balance i.e. the nature of the balance is not Newtonian, since this balance is the result of the energy exchange between the atoms and the particles in conjunction with the nucleus of the elements. At this level it can be explained as the polarity profile or electronegativity of the amino acid.

## MEDICAL IMPLICATIONS

The medical implications of Bioinformatics in the manufacture of new pharmaceutical drugs is incipient (Khan, 2011), mainly because the mathematical-computational algorithms known today do not include an exhaustive computational verification, this means they focus on the assessment of a property that is presumed to be an effective discriminant, and exclude the virtual recreation of the environment where the synthetic peptide acts. The complexity involved in simulating this virtual scenario is certainly high, currently this virtual scenario is replaced by the synthesis of peptides and their subsequent experimental testing in a laboratory. However, the new generation of Bioinformatics algorithms will have to provide a virtual scenario as well as a drastic reduction of the lab testing of the synthetic proteins produced by them. Furthermore, we think that the construction of such virtual scenario to test peptides should be a global initiative (Goodman, 2011) involving research groups in Bioinformatics from different countries, for two reasons: the complexity of the construction of the virtual scenario, and the standardization of the factors and variables that will have to be considered so the synthetic peptides are always evaluated under similar conditions. This initiative is important as it would prevent using bioinformatics algorithms as "filter algorithms", improving their efficiency, and bridging the gap between academic and industry institutions with regulatory agencies (Lesko, 2012).

In this work, we presented the results obtained with polarity index method for three groups of proteins that are a current topic in medicine: SCAAPs, intrinsically disordered proteins, and lipoproteins

related to atherosclerosis. The efficiency of the SCAAPs found in nature is high, however, there are two problems: the increasing difficulty to find them in other organisms and the high costs involved in their synthesis and experimental verification. Therefore, it is imperative to encourage the identification of SCAAPs, given the resurgence of MDROs and the epidemic outbreaks that turned pandemic during the last decade. The intrinsically disordered proteins have shown their association with neurodegenerative diseases known as Amyloidosis, which will have a high impact on the world population during the next decades; and the proteins related to atherosclerosis are associated with coronary artery disease, which is the first cause of death in the USA and in Europe it has been for decades a problem that impacts the health of workers.

## PERSPECTIVES

In humans 25 000–30 000 genes encode proteins so it is reasonable to consider existence of 500 thousand to one million different proteins, this is the result of two factors: a gene may express different proteins and they undergo post-translational changes (Crawford *et al*., 2004). Considering that a computational algorithm takes only one second to analyze the linear sequence of a protein, it will mean eleven days of continuous processing in case of a uniprocessor computer, or an hour in case of a 200 processor cluster (Niiler, 2001). The problem lies not in processing but in the effectiveness of the algorithm and, as it was noted before, the sum of all known algorithms applied to the same protein does not provide more effectiveness but it makes the analysis impractical due to the time-consuming processing. The hardware-software is not and will never be an impediment for the bioinformatics processes applied to Proteomics, but efforts should be aimed at duplicating massive storage capacity and simultaneously at reducing data processing time.

We think that in the near future, the approach to the metrics in new algorithms should be reconsidered to improve their effectiveness, using the known physico-chemical properties but changing their algebraic structure in such a way that they thoroughly inform the dynamic-static aspect of the property studied. As already mentioned, the physico-chemical property Polarity has been considered in many Bioinformatics algorithms (Qureshi *et al*., 2014), however, it was its comprehensive assessment that considers 16 possible polar interactions, which made the difference. To reconsider the approach does not mean to start from scratch, but to examine the most evaluated physico-chemical properties, and study them separately to avoid the over-expression of a property. This aspect in a minimalist approach means not only the expression of the physico-chemical property in the broadest sense, but also its isolation i.e. if a property defines what is sought it should not coexist with another property as this will distort the algorithm. Future algorithms should aim to be exhaustive but minimalist at the same time. A final aspect to consider during the design of these algorithms is that they should be embarrassingly parallel (Snir, 1998), this means programming should process the instructions or tasks of the algorithm simultaneously and computer programs should take into account the same outlook; it is worth mentioning that this technique is not new, its history goes back to 1950 (Wolinsky, 2007).

Finally, in our opinion it is essential to continue the exploration of the polar profile of the first proteins and the effect the bombarding of minimally biased amino acids had on them billion years ago, as the actual knowledge on proteins is negligible compared with the information this span of time can provide, particularly about the role the biases played during the forming of amino acids. On this topic it will be essential to implement broad prebiotic scenarios that allow the recreation of multiple variables from stochastic processes (Rabiner, 1989). In few decades, the design of new drugs will face a drastic reduction in experimental tests on animals (European Commission, 2014), this will involve the design of new algorithms not only according to the guidelines mentioned above but also consistently with the outline of computational biological scenarios that minimize the number of the synthetic proteins tested. The challenges are great and the financial implications considerable, but with the emergence of Multidrug-Resistant Organisms it is evident that it is the human race which is at stake and we have to be prepared to spare no efforts in that endeavor (Zuckerman *et al*., 2009).

### Availability

The test-files, and polarity index method program must be requested from 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, TB, VU, and JACG.

### Acknowledgments

The authors thank Concepción Celis Juárez whose suggestions and proof-reading have greatly improved the original manuscript, and we also acknowledge the Computer Science department at Instituto de Ciencias Nucleares at the Universidad Nacional Autónoma de México for support.

## REFERENCES

Boman HG (1995) Peptide antibiotics and their role in innate immunity. *Annu Rev Immunol* **13**: 61–92.

Brendel V, Bucher P, Nourbakhsh IR, Blaisdell BE, Karlin S (1992) Methods and algorithms for statistical analysis of protein sequences. *Proc Natl Acad Sci USA* **89**: 2002–2006.

Christian Oestreicher C (2007) A history of chaos theory. *Dialogues Clin Neurosci* **9**: 279–289.

Crawford GE, Holt IE, Mullikin JC, Tai D, Blakesley R, Bouffard G, Young A, Masiello C, Green ED, Wolfsberg TG, Collins FS (2004) Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc Natl Acad Sci USA* **101**: 992–997.

Delaye L, Becerra A, Lazcano A (2005) The last common ancestor: what's in a name? *Orig Life Evol Biosph* **35**: 537–554.

European Commission (2014) Digital Agenda for Europe. http://ec.europa.eu/digital-agenda/futurium/en/content/new-foods.

Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD,

Chiu W, Garner EC, Obradovic Z (2001) Intrinsically disordered protein. *J Mol Graph Model* **19**: 26–59.
Fox SW (1960) How did life begin. *Science* **132**: 200–208.
Gaucher EA, Kratzer JT, Randall RN (2010) Deep phylogeny—how a tree can help characterize early life on earth. *Cold Spring Harb Perspect Biol* **2**: a002238.doi: 10.1101/cshperspect.a002238.
GenBank (2011) The GenBank Submissions Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); **2011**. Formatting your Submission. http://www.ncbi.nlm.nih.gov/books/NBK53702/
Goodman AF (2011) Analysis, biomedicine, collaboration, and determinism challenges and guidance: wish list for biopharmaceuticals on the interface of computing and statistics. *J Biopharm Stat* **21**: 1140–1157. doi: 10.1080/10543406.2011.613361.
Guyton JR, Klemp KF (1989) The lipid-rich core region of human atherosclerotic fibrous plaques. Prevalence of small lipid droplets and vesicles by electron microscopy. *Am J Pathol* **134**: 705–717.
Khan SH, Ahmad F, Ahmad N, Flynn DC, Kumar R (2011) Protein-protein interactions: principles, techniques, and their potential role in new drug development. *J Biomol Struct Dyn* **28**: 929–938.
Kitaev AY (1997) Quantum computations: algorithms and error correction. *Russian Math Surveys* **6**: 1191–1249.
Koba S, Hirano T, Murayama S, Kotani T, Tsunoda F, Iso Y, Ban Y, Kondo T, Suzuki H, Katagiri T (2003) Small dense LDL phenotype is associated with postprandial increases of large VLDL and remnant-like particles in patients with acute myocardial infarction. *Atherosclerosis* **170**: 131–140.
Kosmulski M (2009) pH-dependent surface charging and points of zero charge. IV. Update and new approach. *J Colloid Interface Sci* **337**: 439–448. doi: 10.1016/j.jcis.2009.04.072.
Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armañanzas R, Santafé G, Pérez A, Robles V (2006) Machine learning in bioinformatics. *Brief Bioinform* **7**: 86–112.
Lesko LJ (2012) Drug research and translational bioinformatics. *Clinical Pharmacology & Therapeutics* **91**: 960–962. doi:10.1038/clpt.2012.45.
Madden TL, Tatusov RL, Zhang J (1996) Applications of network BLAST server. *Methods Enzymol* **266**: 131–141.
Magrane M. UniProt consortium (2011) UniProt Knowledgebase: a hub of integrated protein data Database bar009.
Miller SL (1953) A Production of amino acids under possible primitive earth conditions. *Science* **117**: 528–529.
Munkres J (2000) *Topology*. Pearson 2 edn, pp 537. ISBN-10: 0131816292.
Niiler E (2001) Supercomputer for proteomics. *Nat Biotechnol* **193**: 10.1038/85601.
Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, Dunker AK (2005) Comparing and combining predictors of mostly disordered proteins. *Biochemistry* **44**: 1989–2000.
Pauling L (1960) The Nature of the Chemical Bond and the Structure of Molecules and Crystals: *An Introduction to Modern Structural Chemistry*. Cornell University Press. pp 644, ISBN: 9780801403330, USA.
Polanco C, Samaniego JL (2009) Detection of selective cationic amphipatic antibacterial peptides by Hidden Markov models. *Acta Biochim Pol* **56**: 167–176.
Polanco C, Samaniego JL, Buhse T, Castañón González JA (2013) A toy model of prebiotic peptide evolution: the possible role of relative amino acid abundances. *Acta Biochim Pol* **60**: 175–182.
Polanco C, Samaniego JL, Buhse T, Mosqueira FG, Negron-Mendoza A, Ramos-Bernal S, Castanon-Gonzalez JA (2012) Characterization of selective antibacterial peptides by Polarity Index. *Int J Peptides* **2012**: 58502. http://dx.doi.org/10.1155/2012/585027.
Polanco C, Samaniego JL, Castañón-González JA, Buhse T (2014a) Polar profile of antiviral peptides from AVPpred Database. *Cell Biochem Biophys* **70**: 1469–1477. doi: 10.1007/s12013-014-0084-4.
Polanco C, Samaniego JL, Castañón-González JA, Buhse T, Arias-Estrada M (2014b) Computational model of abiogenic amino acid condensation to obtain a polar amino acid profile. *Acta Biochim Pol* **61**: 253–258.
Polanco C, Samaniego JL, Castañón-González JA, Buhse T, Sordo ML (2013a) Characterization of a possible uptake mechanism of selective antibacterial peptides. *Acta Biochim Pol* **60**: 629–633.
Polanco C, Samaniego JL, Castañón-González JA, Buhse T, Sordo ML (2013b) Detection of selective antibacterial peptides by the Polarity Profile method. *Acta Biochim Pol* **60**: 183–189.
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 (2014) Identification of proteins associated with amyloidosis by polarity index method *Acta Biochim Pol* **62**: 41–55.
Polanco C, Samaniego-Mendoza JL, Buhse T, Castañón-González JA, Leopold-Sordo M (2014b) Polar Characterization of Antifungal Peptides from APD2 Database. *Cell Biochem Biophys* **70**: 1479–1488. doi: 10.1007/s12013-014-0085-3.
Polanco C, Castañón-González JA, Uversky VM (Letter to the Editor) Buhimschi IA, Nayeri UA, Zhao G, Shook LL, Pensalfini A, Funai EF, Bernstein IM, Glabe CG, Buhimschi CS (2014e) Protein misfolding, congophilia, oligomerization, and defective amyloid processing in preeclampsia. *Sci Transl Med* **6**: 245ra92. doi: 10.1126/scitranslmed. 3008808.
Polanco C, Samaniego-Mendoza JL, Castañón-González JA, Buhse T (Letter to the Editor) Howard SJ, Hopwood S, Davies SC (2014d) Antimicrobial Resistance: A Global Challenge. *Sci Transl Med* doi:10.1126/scitranslmed. 3009315.
Polanco C, Castañón-González JA, Buhse T, Uversky VN, Zonana-Amkie R (2016) Classifying lipoproteins based on their polar profiles. *Acta Biochim Pol* 63: 233–239. http://dx.doi.org/10.18388/abp.2014_918.
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 (2015a) Identification of proteins associated with amyloidosis by polarity index method. *Acta Biochim Pol* **62**: 41–55. http://dx.doi.org/10.18388/abp.2014_755.
Putz MV, Lazea M, Putz AM, Duda-Seiman C (2011) Introducing Catastrophe-QSAR. Application on Modeling Molecular Mechanisms of Pyridinone Derivative-Type HIV Non-Nucleoside Reverse Transcriptase Inhibitors. *Int J Mol Sci* **12**: 9533–9569.
Qureshi A, Thakur N, Tandon H, Kumar M (2014) AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic Acids Res* **42**: D1147–D1153. doi: 10.1093/nar/gkt1191.
Rabiner LR (1989) A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE* **77**.
Rode BM (1999) Peptides and the origin of life. *Peptides* **20**:773–786.
Rudin W (1990) *Principles of Mathematical Analysis,* 3a edn, ISBN: 968-6046-82-8. McGraw-Hill, 1990.
Sharp PM (1985) Does the 'non-coding' strand code? *Nucleic Acids Res* **13**: 1389–1397.
Snir M (1998) MPI-The *Complete Reference*. 2nd edn. Cambridge, Massachusetts: MIT Press.
Uversky VN (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* **11**: 739–756.
Uversky VN (2009) Intrinsic disorder in proteins associated with neurodegenerative diseases. *Frontiers in Bioscience* **14**: 5188–5238.
Uversky VN (2010a) Mysterious oligomerization of the amyloidogenic proteins. *FEBS J* **277**: 2940–2953.
Uversky VN (2002) What does it mean to be natively unfolded? Eur J Biochem **269**: 2–12.
Uversky VN (2010b) The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome. *J Biomed Biotechnol* **2010**: 568068.
Uversky VN, Dunker AK (2010) Understanding protein non-folding. *Biochim Biophys Acta* **1804**: 1231–1264.
Uversky VN (2013) Intrinsic disorder-based protein interactions and their modulators. *Curr Pharm Des* **19**: 4191–4213.
Uversky VN, Fink AL (2004) Conformational constraints for the amyloid fibrillation: The importance of being unfolded. *Biochim Biophys Acta* **1698**: 131–153.
Uversky VN, Gillespie JR, Fink AL (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **41**: 415–427.
Uversky VN, Oldfield CJ, Dunker AK (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* **37**: 215–246. doi:10.1146/annurev.biophys.37.032807.125924.
Vinogradov IM (1985) Algebra, Mathematical Logic, Number Theory, Topology. Algebra, Mathematical Logic, Number Theory, Topology. American Mathematical Society. ISBN: 0821830961, 9780821830963, pp 266 http://books.google.com.mx/books?id=r-Say9 ZPucc cC.
Wang G, Li X, Wang Z (2009) APD2: the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Res* **37** ( Database issue): D933–D937.
Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* **293**: 321–331.
Woese CR (1979) A proposal concerning the origin of life on the planet earth. *J Mol Evol* **13**: 95–101.
Wolinsky H (2007) I, scientist. Will robots at the bench leave scientists free to think? *EMBO Rep* **8**: 720–722. doi: 10.1038/sj.embor.7401038.
Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* **32**: 1466–1474. doi: 10.1002/jcc.21707.
Zuckerman AJ, Banatvala JE, Schoub BD, Griffiths PD, Mortimer P, Mahy WJ (2009) *Emerging Virus Infections*. Brian W ed. John Wiley & Sons, Ltd. ISBN: 9780470741405. doi: 10.1002/97804707 41405.ch4.