Polar analysis of intrinsically disordered proteins: a review

C. Polanco[*,1], **J.A. Castañón González**[2], **J. Levy**[3] and **M. Arias Estrada**[4]

[1] Department of Mathematics, Faculty of Sciences, Universidad Nacional Autónoma de México, 04510 D.F., Mexico
[2] Departments of Critical Care Unit and Biomedical Research, Hospital Juárez de México, 07760 D.F., Mexico
[3] Faculty of Health Sciences, Universidad Anahuac, 52786 Estado de México, Mexico
[4] Department of Computational Sciences, Instituto Nacional de Astrofísica, Óptica y Electrónica, 72840 Puebla, Mexico

In nature, proteins exhibit mainly three states: folded, partially folded, and unfolded. These states are related to the level of "structure" that proteins have. An unfolded protein particularly, will show in its linear chain a random profile, either in a single segment or in separated segments jointed together. Recently it has been found that unfolded proteins adhere (or clump) to neurons, leading to severe neurodegenerative disorders in humans, grouped under the term Amyloidosis. Proteins that cause these disorders may well get their "random" profile by interacting with a group of proteins that act as "precursors". When these precursors interact with partially folded proteins, they "denature" them through biological process, transforming them into unfolded proteins. However, some partially folded proteins, under the same interaction, get an ordered state that do not clump on neurons. What physico-chemical profile differentiates unfolded proteins from neuron proteins? What physico-chemical profile features the precursors that transform a partially folded protein into a folded or unfolded protein? In this review a mathematical-computational method, called Polarity Index Method, is applied to describe the physico-chemical profile of these protein groups, using only Polarity (according to L. Pauling 1955), as metric. It also reviews the main *ad hoc* computational architectures for the method that will enhance its performance since it is computational expensive, so in further stages, it can be used to analyse protein regions with specific pathogenic or structural action.

Keywords: intrinsically disordered proteins; folded proteins; partially folded proteins; unfolded proteins

1. General remarks

The Intrinsically Disordered Proteins (IDP) [1] are proteins that do not have a stable tertiary structural conformation [2]. These proteins are classified, according to the degree of disorder in their tertiary structure into: folded proteins [3]; unfolded proteins [3]; and an intermediate group, called partially folded proteins [4]. Although for decades a stable tertiary structure has been synonymous with protein functionality [5], the IDPs challenge this paradigm because they are featured by an unstable tertiary structure, and have the characteristic to clump into neurons [6], leading to important neurological disorders grouped under the term Amyloidosis [7]. However, there are cases where IDPs adopt a stable tertiary structure after interacting with precursor proteins [8]. In order to elaborate on the understanding of the IDPs, the matematical-computacional method called polarity index method (PIM) [9-11], was oriented to determine the polarity profile of proteins associated with the IDPs extracted from Oldfield et al., 2005 [4]. Then, this profile was compared with the groups of human neuronal, and non-human neuronal proteins from UNIPROT Database (UNIPROT) [12], and the subset of amyloid precursor proteins from AMYPDB Database (AMYPDB) [8]. These proteins are featured by having the longest sub-sequences in each of the 33 families of precursor proteins in human beings [8]. This work was based on an important statement made by Pauling l. 1955 [2], related to the different values of the electronegativity in the primary elements of matter, that can be "reflected" in the amino acids that constitute a protein. To conduct this evaluation, the polarity profile for the IDPs was determined with the PIM. The results showed that their profile is different from the neuron protein profile and the profiles of the IDP subgroups. Since this polar profile required the evaluation of large groups of proteins in databases, it was convenient to introduce computational adjustments to enable the analysis of massive information. For example, in order to explore all 13aa in length proteins, it is necessary to evaluate 20^{13} proteins. This large number of proteins needs parallel processing [13], as in a personal computer (PC) it is not possible to evaluate all the proteins in a reasonable time; therefore, it was necessary to use parallel computing technologies to substantially reduce the processing time from hours to milliseconds.

2. Amyloidosis: state of the art

Although a highly effective intracellular mechanism exist to chaperone proteins while they are being synthetized and secreted to eliminate proteins that misfold and to ensure that they achieve correct tertiary conformation and function; some genetic mutations, incorrect processing and other factors may favor misfolding with loss of normal protein function with subsequent intracellular or extracellular aggregation. In the Amyloidosis the aggregates are extracellular, and the misfolded protein subunits assume a common antiparallel, beta pleated sheet structural conformation that leads to the formation of higher order oligomers and then of fibrils with unique staining properties. Ultimately, the fibrillary protein deposits are likely to interfere with normal organ function. Amyloidosis defined by the biochemical nature of the protein composing the fibril deposits, are classified according to whether they are systemic or localized, inherited or

acquired and their clinical patterns. Diagnosis and treatment of Amyloidosis rest upon the histopathologic identification of Amyloid deposits and immunohistochemical, biochemical or genetic determination of amyloid type.

3. Intrinsically disordered proteins

Intrinsically or natively disordered proteins (IDP) is a group of proteins that lack ordered three dimensional structure, they are classified according to the degree of disorder in their protein regions. Among the IDPs, the natively unfolded proteins (NUP), show the highest level of disorder, and they are strongly associated with Amyloidosis. Thirty percent of the proteins in Eukaryotes are NUP [14], they facilitate the recognition, interaction, transcription and signaling between proteins, but they are also involved in the development of neurodegenerative disorders.

3.1 Neuron proteins

Protein misfolding has great impact at the neurological level causing a variety of illnesses associated with Amyloid, when it accumulates in the Central Nervous System (CNS) it produces Neurodegeneration [15]. One of the best known examples is Alzheimer's disease (AD), which is estimated to have affected more than 30 million people [16]. It has been found that at least seven neuro proteins are directly related to neurodegenerative diseases.

A-beta protein precursor. The precursor to amyloid beta (A-beta) is a protein with multiple functions, it is found in high concentrations in the CNS. The A-beta is transported through the axons and it is accumulated in the pre-synaptic terminal space and nerve growth cones. It is largely related to the AD, but it can also be found in other neurodegenerative diseases like the Down syndrome and the hereditary cerebral hemorrhage associated with Amyloidosis (Dutch) [17-20].

ABriPP/AdanPP. The BriPP is a transmembrane protein, when it presents an autosomal dominant mutation in the BRI gene, it causes the formation of the ABriPP [21, 22] amyloid precursor. This precursor is deposited in the CNS in form of amyloid, generating an amyloid arthropathy that causes a CNS atrophy; this disease is called Familial British dementia and it usually occurs in the fifth decade of life [21]. The clinical features of this disease are progressive dementia, spastic palsy and cerebellar ataxia [23]. The ABriPP plaques can also be found in the parenchyma of the myocardium, pancreas, and skeletal muscle [21-23].

Alpha Synuclein. The alpha synuclein is located initially in the pre-synaptic terminal regions of neurons and their core, these proteins have different transport properties both intracellular and extracellular [24]. When the alpha synuclein folds incorrectly in oligomeric formations it causes aggregation [25-27].

Cystatin C. The Cystatin C protein (CysC) is abundantly located in the CNS and it is expressed by different cells of the nervous system, neurons, astrocytes and microglial [28] cells. Most of the CysC found in the brain fluids is produced by the choroid plexus [29]. It has been observed that in neurodegenerative pathologies CysC diminishes in the cerebrospinal fluid, therefore, it has been used as biomarker for diseases such as Amyotrophic lateral sclerosis (ALE), and Alzheimer's disease, as it is significantly reduced in these pathologies compared to people without [28] neurodegenerative diseases.

Huntingtin. The brain is rich in proteins associated with Huntingtin 1. This protein is involved in the intracellular traffic of neurons in different molecules that are fundamental for their physiological functions [31]. The Clinical presentation of HD indicates multiple psychiatric, neuromotor, and dementia disorders.

Prion protein. Prion diseases are progressive neurodegenerative disorders that are inevitably fatal [32]. These proteins are found in different human and animal diseases, the main conditions are [33, 34]: Creutzfel-Jakob disease, a new variant Creutzfel-Jakob disease, Grestmann-Straussler-Scheinker disease, fatal familial insomnia, Kuru, and bovine spongiform encephalopathy [33, 36].

Tau. Tau proteins belong to the Microtubule Associated Protein group (MAP), they give stability and dynamism to microtubules and are important in the formation of dendrites and axons [37, 38]. Their misfolding in the CNS generates pathological processes that create instability on the neurons cytoskeleton, preventing them from performing their functions, and causing neuronal death [37-39].

3.2 Precursors

Protein precursors of amyloid plaques (amyloid), are proteins with multiple and different physiological functions [40]. They have large structural and functional differences; if they fold incorrectly they can form insoluble structures incapable of proteolysis [40-42]. The pathological conformation of these proteins is obtained when a predominant antiparallel structure, called antiparallel fibril beta sheets, is formed. This pathological folding occurs simultaneously with physiological folding [41].

4. The polarity index method

The PIM conducts a comprehensive scrutiny of polarity in peptides and proteins, from the polarity incidences of the amino acids that form their linear sequence [2]. The method records these incidences, readying the protein sequence, from left to right, taking the amino acids in pairs sequentially, and registering them in an incidence matrix. The PIM is based on the 20 known amino acids that are grouped in the following categories [2]: acidic-polar P- = {D, E}, basic-polar P+ = {H, K, R}, non-polar NP = {A, F, I, L, M, P, V, W}, and neutral-polar N = {C, G, N, Q, S, T, Y}. From this classification, the protein sequence formed by amino acids is translated into a numerical sequence with the rule: {P+, P-, N, NP} → {1, 2, 3, 4}. For instance, if the sequence of a protein is formed with these amino acids VNWKKVLGKIIKVAK, the numerical sequence will be 434114431441441. With this conversion, it will be possible to build an incidence matrix with the amino acid pairs read from left to right, where (i,j) = (row, column). In this example, the interaction (4.4) occurs three times, while the interaction (2.2) does not occur at all. It is most likely that each protein has a different incidence matrix, however, if the matrices of a group of proteins with similar function (virus, mammalian cell etc.) are compared, there will be regularity in the profile of each group [9-11]. This approximation model has a high level of efficiency (> 75%), to predict correct protein action. The proposed metric comprehensively evaluates a single physico-chemical property, since it takes the 16 possible polarity interactions into account; and it is highly likely that this algebraic approach is what confers the method [15] its efficiency.

5. Electromagnetic balance

The electronegativity [2], understood as the number of the external electrons in the balance that are shared by two elements, it is a metric that the PIM extends to the constitutive amino acids of a protein. Thus the amino acid interaction, read in the gene of a genome, is a sample of the electromagnetic balance that the protein has in 3D space. The PIM is based on the statement that the electromagnetic balance of the amino acids permeates the linear sequence of the gene in the genome, and makes possible the prediction of the protein function. Is it possible to preserve the electromagnetic balance of a protein, despite changing the amino acid sequence of the gene? With the PIM it has been observed that it is possible. Some changes do not alter the electromagnetic balance of the protein, however, there are cases where a single variation, completely changes its polarity profile, and therefore, its function.

6. Results

The PIM effectively discriminates the groups of folded, partially folded and unfolded proteins, as well as the human neuronal and non-human neuronal protein groups. (Table 1). According to its results, the precursor protein group is partially folded (Table 1).

Table 1 Structural groups.

Groups → ↓	**Folded**	**Partially folded**	**Unfolded**	**Human neuronal proteins**	**Non-human neuronal proteins**	**Precursors**
Folded	**70**	19	6	**44**	**42**	10
Partially folded	33	**71**	32	32	29	**40**
Unfolded	6	4	**72**	9	8	4
Precursors	2	2	2	1	1	**73**

7. Stochastic models

Deterministic models [44] are those where the variables that influence the experiment are known with certainty all the time, i.e., when the model is analyzed, the information of all variables are always available. In contrast, stochastic models [44] have an unknown random variable, providing a degree of uncertainty. The discrete stochastic models, whose immediate future depends on the present moment, are called stochastic Markov models or stochastic models that verify the Markov conjecture [45].

8. Clustering approach

The polarity index method introduced in section 4, is a compact way to represent the sequence and relationship of proteins in the amino acid. Previous work has shown [9-11] a large efficiency to identify the level of structure of the protein that can be further improved. On one hand, the polarity index method is an efficient way to describe the actual physics of the protein sequence, but probably still incomplete. Nevertheless, it is a good trade-off to represent the amino acids and analyze large data sets of them using current state of the art pattern recognition, data mining algorithms and computational resources. The polarity index provides a simplified (but effective) representation of the amino acid, an exhaustive representation would require to model at the atom level the actual molecular structure of the amino acid, and the electromagnetic forces among all atoms but, it would be too computational expensive and probably a long way to discover additional properties that define the behavior of each amino acid.

On the other hand, the polarity index method can be explored in new directions to introduce advanced clustering algorithms such as data mining algorithms to explore the limits of the representability of the method, that is, how accurate the polarity index method represents the actual molecular and electromagnetic properties that give each amino acid its unique characteristics. In this direction there is an open field to explore, in particular the kind of clustering methods that can be applied, and further fine tuning the way the index is being generated.

Clustering can be conducted in two wide techniques: the first one is to explore, using the traditional pattern recognition methods, to find clusters of similar properties encoded in the polarity index matrix, in a multi-dimensional space (where the number of proteins in the sequence is the number of dimensions). There are mature algorithms to do some basic clustering like K-means based algorithms and other supervised and not supervised methods that could help discover hidden clusters and patterns that can better explain the behavior of the amino acids and obtain insights on a possible generalization during the classification. Furthermore, there seems to be some parts of the sequence that are found in other sequences that could give additional insights of the amino acid level of structure, that could help better define the sequence description.

A second line to explore is to consider the polarity index unfolded matrix, as a "word", which gives a unique representation to the amino acid and, that could be mined from all the possible amino acid combinations to find semantic relations between the incidence of the polarity index. Techniques like semantics, well advanced in text data mining applications, could help discover hidden relationships that go beyond the first neighbor relationship that the polarity index encodes and find other insights on which structures or subsequences of the amino acid are responsible for the level of structure.

9. Discussion

In recent decades, the work in bioinformatics have made important contributions to the understanding of Amyloidosis, specifically the information in databases of proteins associated with these neurological disorders and the application of software aimed at their identification and characterization. In this work, the PIM has proved to be an effective discriminant of the IDP sub-classifications, however, the method can be enhanced by: (i) converting it into an unsupervised self-learning method that identifies systematically IDPs; (ii) setting up a website where users can test protein groups in FASTA format; and (iii) implementing parallel processing in clusters or supercomputers, that enable the exploration of specific protein regions of fixed “n” length. It is also important to understand the reasons of the effectiveness of polarity, as the only physico-chemical property required; a contribution to this understanding might be the exhaustive measurement given by the matrix. Nevertheless, other properties will have to be examined, and new algebraic structures will have to be found to provide comprehensive measurements.

10. Future trends

From the mathematical-computational point of view, there are methods that relate the chemical structure with the main activity associated with IDPs. These are classified into two main groups: supervised and non-supervised algorithms [46], both with varying degrees of complexity [47]. What is the trend for the next decades? Bioinformatics algorithms will probably become non-supervised and completely parallelized, to process large amounts of data. After this stage, these algorithms will probably be used in robots capable of doing the same identification inside the organism. It is also plausible that those robots will be equipped with biological processes, to dissolve the protein clumps characteristic of IDPs.

Acknowledgements The authors acknowledge the Computer Science department at Institute for Nuclear Sciences at the Universidad Nacional Autónoma de México for support.

References

[1] Dunker AK, *et al*. Intrinsically disordered protein. J Mol Graph Model 2001;19:26-59.
[2] Pauling L. General Chemistry [Química General] 3rd edition W. H. Freeman & Company Publishers, 1955, pp. 227,621.
[3] Dobson CM. Protein folding and misfolding. Nature 2003;426:884-90.
[4] Oldfield CJ, Cheng Y, Cortese MS,Brown CJ, Uversky VN, Dunker AK. Comparing and combining predictors of mostly disordered proteins. Biochemistry 2005;44:1989−2000.
[5] Pal D, Eisenberg D. Inference of protein function from protein structure. Structure 2005;13:121-30.
[6] Mantha AK, Sarkar B, Tell G. A short review on the implications of base excision repair pathway for neurons: relevance to neurodegenerative diseases. Mitochondrion. 2014;16:38-49. DOI: 10.1016/j.mito.2013.10.007
[7] Loizos S, Shiakalli Chrysa T, Christos GS. Amyloidosis: review and imaging findings. Semin Ultrasound CT MR. 2014;35:225-39. DOI: 10.1053/j.sult.2013.12.006.
[8] Pawlicki S, Le Béchec A, Delamarche C. AMYPdb: a database dedicated to amyloid precursor proteins. BMC Bioinformatics 2008;9:273. DOI: 10.1186/1471-2105-9-273.
[9] Polanco C, Buhse T, Samaniego JL, Castañón-González JA. Detection of selective antibacterial peptides by the Polarity Profile method. Acta Biochim Pol. 2013;60:183-9.
[10] Polanco C, Castañón González JA, Uversky V. Profiling TSOP proteins by Polarity Index Method. Comment on: Dobson A. Mathematical models for emerging disease. Science 2015;347:551-555.
[11] Polanco C, *et al*. Characterization of a possible uptake mechanism of selective antibacterial peptides. Acta Biochim Pol 2013;60:629-33.
[12] Magrane M, UniProt Consortium. UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011. Available from: http://www.uniprot.org/
[13] Ostrovsky B, Smith MA, Bar-Yam Y. Applications of parallel computing to biological problems. Annu Rev Biophys Biomol Struct 1995;24:239-67.
[14] Uversky VN. Natively unfolded proteins: a point where biology waits for physics. Protein Sci 2002;11:739-56.
[15] Muller T, Meyer HE, Egensperger R, Marcus K. The amyloid precursor protein intracellular domain (AICD) as modulator of gene expression, apoptosis, and cytoskeletal dynamics-relevance for alzheimer's disease. Prog Neurobiol. 2008;85:393-406.
[16] Janson J, Ashley RH, Harrison D, McIntyre S, Butler PC. The mechanism of islet amyloid polypeptide toxicity is membrane disruption by intermediate-sized toxic amyloid particles. Diabetes. 1999;48:491-498.
[17] Mattson MP. Cellular actions of beta-amyloid precursor protein and its soluble and fibrillogenic derivatives. Physiological Reviews. 1997;77:1081-1132.
[18] Bandyopadhyay S, Cahill C, Balleidier A, et al. Novel 5' untranslated region directed blockers of iron-regulatory protein-1 dependent amyloid precursor protein translation: Implications for down syndrome and alzheimer's disease. PLoS ONE. 2013;8:1-14.
[19] Pilotto A, Padovani A, Borroni B. Clinical, biological, and imaging features of monogenic alzheimer's disease.BioMed Research International. 2013; pp9 http://dx.doi.org/10.1155/2013/689591-9 .
[20] Wilhelmus MMM, Boelens WC, Kox M, et al. Small heat shock proteins associated with cerebral amyloid angiopathy of hereditary cerebral hemorrhage with amyloidosis (dutch type) induce interleukin-6 secretion. Neurobiol Aging. 2009;30:229-240.
[21] Holton JL, Ghiso J, Lashley T, et al. Familial british dementia (FBD): A cerebral amyloidosis with systemic amyloid deposition. Neuropathology & Applied Neurobiology. 2002;28:148-148.
[22] Schwab C, Hosokawa M, Akiyama H, McGeer PL. Familial british dementia: Colocalization of furin and ABri amyloid. Acta Neuropathol. 2003;106:278-284.
[23] Vidal R, Revesz TE. A decamer duplication in the 3' region of the BRI gene originates an amyloid peptide that is associated with dementia in a danish kindred. Proc Natl Acad Sci U S A. 2000;97:4920-4925.
[24] Fonseca T, Villar-Piqué A, Outeiro T. The interplay between alpha-synuclein clearance and spreading. Biomolecules. 2015;5:435-471.
[25] Illes-toth E, Ramos M, Cappai R, Dalton C, Smith D. Structural characterisation of high order alpha-synuclein oligomers capable of inducing intracellular aggregation. Biochem J. 2015.
[26] Nikisha C, et al. "Characterization of HTT Inclusion Size, Location, and Timing in the zQ175 Mouse Model of Huntington´s Disease: An In Vivo High-Content Imaging Study." Ed. David R Borchelt. PLoS ONE 10.4 (2015): e0123527. PMC. Web. 12 May 2015.
[27] Roberts H, Brown D. Seeking a mechanism for the toxicity of oligomeric a-synuclein. Biomolecules. 2015;5:282-305.
[28] Kaur G, Levy E. Cystatin C in alzheimer's disease. Frontiers in Molecular Neuroscience. 2012;5.
[29] Kaeser SA, Herzig MC, Coomaraswamy J, et al. Cystatin C modulates cerebral ß-amyloidosis. Nat Genet. 2007;39:1437-1439.
[30] Yang G, Yang M, Lim Y, et al. Huntingtin associated protein 1 regulates trafficking of the amyloid precursor protein and modulates amyloid beta levels in neurons. J Neurochem. 2012;122:1010-1022.
[31] Borchelt DR, Carty N, Berson N, et al. Characterization of HTT inclusion size, location, and timing in the zQ175 mouse model of huntington´s disease: An in vivo high-content imaging study. PLOS ONE. 2015;10:e0123527.
[32] Weissmann C. Molecular genetics of transmissible spongiform encephalopathies. J Biol Chem. 1999;274:3-6.
[33] Belay ED. Transmissible spongiform encephalopathies in humans, Annual Review of Microbiology. October 1999;53:283.
[34] Prusiner SB. Novel proteinaceous infectious particles cause scrapie. Science. 1982;216:136-144.
[35] Masters CL, Beyreuther K. Tracking turncoat prion proteins. Nature. 1997;388:228.
[36] Coria BF, Alberca R, lópez pousa YS. Demencias por priones. en:. enfermedad de alzheimer y otras demencias. 2002nd ed. España: Médica Panamericana.
[37] Erduy J. Las Taupatías. *Revista Mexícana de Neurociencias*, http://revmexneuroci.com/wp-content/uploads/2014/07/Nm00 23-07.pdf. 2002;3, pp.165-167.

[38] Fellous A, Francon J, Lennon AM, Nunez J. Microtubule assembly in vitro. purification of assembly-promoting factors. European Journal of Biochemistry. 1977;78:167-174.
[39] Couchie D, Mavilia C, Georgieff IS, Liem RK, Shelanski ML, Nunez J. Primary structure of high molecular weight tau present in the peripheral nervous system. Proc Natl Acad Sci U S A. 1992;89:4378-4381.
[40] Merlini G, Bellotti V. Molecular mechanisms of amyloidosis. N Engl J Med. 2003;349:583-596.
[41] Polanco C, SamaniegoJL, Uversky V.N, Castañón-González J.A, et al. Identification of proteins associated with amyloidosis by polarity index method. Acta Biochim Pol. 2015;62:41-55.
[42] Stoppini M, Obici L, Lavatelli F, et al. Proteomics in protein misfolding diseases. Clinical Chemistry & Laboratory Medicine. 2009;47:627-635.
[43] Thakur N, Qureshi A, Kumar M. AVPpred: collection and prediction of highly effective antiviral peptides. Nucleic Acids Res 2012. W199–204.
[44] Shmulevich I, Aitchison JD. Deterministic and stochastic models of genetic regulatory networks. Methods in enzymology 2009;467:335-356. DOI:10.1016/S0076-6879(09)67013-0.
[45] Eddy SR. Hidden Markov models, Current Opinion in Structural Biology. 1996;6:361-365, ISSN 0959-440X, http://dx.doi.org/10.1016/S0959-440X(96)80056-X.
[46] Pan W, Shen X, Liu B. Cluster Analysis: Unsupervised Learning via Supervised Learning with a Non-convex Penalty. Journal of machine learning research : JMLR 2013;14:1865-.
[47] Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to Algorithms, (3rd ed.). The MIT Press, 1995.

