ACTA ABP BIOCHIMICA POLONICA

Vol. 62, No 1/2015
41–55
http://dx.doi.org/10.18388/abp.2014_755

Regular paper

# Identification of proteins associated with amyloidosis by polarity index method

Carlos Polanco[1]✉, José Lino Samaniego[1], Vladimir N. Uversky[2,3,4], Jorge Alberto Castañón-González[1], Thomas Buhse[5], Marili Leopold-Sordo[1], Alejandro Madero-Arteaga[1], Alicia Morales-Reyes[6], Lourdes Tavera-Sierra[7], Jesus A. González-Bernal[6] and Miguel Arias-Estrada[6]

[1]Facultad de Ciencias de la Salud, Universidad Anáhuac, Anáhuac, México; [2]Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, USA; [3]Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia; [4]Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino, Moscow Region, Russia; [5]Centro de Investigaciones Químicas, Universidad Autónoma del Estado de Morelos, Cuernavaca, Morelos México; [6]Departamento de Ciencias Computacionales, Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, México; [7]Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México, México

**There is a natural protein form, insoluble and resistant to proteolysis, adopted by many proteins independently of their amino acid sequences *via* specific misfolding-aggregation process. This dynamic process occurs in parallel with or as an alternative to physiologic folding, generating toxic protein aggregates that are deposited and accumulated in various organs and tissues. These proteinaceous deposits typically represent bundles of β-sheet-enriched fibrillar species known as the amyloid fibrils that are responsible for serious pathological conditions, including but not limited to neurodegenerative diseases, grouped under the term amyloidoses. The proteins that might adopt this fibrillar conformation are some globular proteins and natively unfolded (or intrinsically disordered) proteins. Our work shows that intrinsically disordered and intrinsically ordered proteins can be reliably identified, discriminated, and differentiated by analyzing their polarity profiles generated using a computational tool known as the polarity index method (Polanco & Samaniego, 2009; Polanco *et al*., 2012; 2013; 2013a; 2014; 2014a; 2014b; 2014c; 2014d). We also show that proteins expressed in neurons can be differentiated from proteins in these two groups based on their polarity profiles, and also that this computational tool can be used to identify proteins associated with amyloidoses. The efficiency of the proposed method is high (i.e. 70%) as evidenced by the analysis of peptides and proteins in the APD2 database (2012), AVPpred database (2013), and CPPsite database (2013), the set of selective antibacterial peptides from del Rio *et al*. (2001), the sets of natively unfolded and natively folded proteins from Oldfield *et al*. (2005), the set of human revised proteins expressed in neurons, and non-human revised proteins expressed in neurons, from the Uniprot database (2014), and also the set of amyloidogenic proteins from the AmyPDB database (2014).**

**Key words**: Polarity index method; natively unfolded proteins; intrinsically disordered proteins; natively folded proteins; neurons; amyloidosis; amyloid; amyloidogenic protein

**Received**: 13 March, 2014; revised: 25 June, 2014; accepted: 23 November, 2014; available on-line: 12 February, 2015

## INTRODUCTION

The amyloidoses are a large group of protein conformational diseases in which pathological intracellular or extracellular protein aggregation takes place largely because of the protein misfolding events leading to specific partially folded species with a strong propensity to acquire more than one conformation. Although certain group of proteins, known as natively unfolded or intrinsically disordered proteins, require a high degree of structural "disorder" or structural plasticity in their native state to favor interactions with specific ligands (Dunker *et al*., 2001; Uversky *et al*., 2000; Uversky, 2013; Wright & Dyson, 1999), they also poses a delicate balance in which the hazy border between risky self-aggregation and sophisticated function is easily crossed (Uversky *et al*., 2008a; Uversky, 2009a; Uversky, 2010). In contrast to the classic notion that foldable proteins require well-defined globular structure to be functional, genomic and proteomic analyses revealed that functional proteins without unique 3D structure are common, and the abundance of these proteins correlates directly with the complexity of organisms, with this property being present in at least 2% of archaeal, 4% of eubacterial, and 33% of eukaryotic proteins (Hansen *et al*., 2006; Uversky, 2010b; Xue *et al*., 2010; Xue *et al*., 2010a; Xue *et al*., 2012). Therefore, protein intrinsic disorder can be considered as an evolutionarily conserved phenomenon, which is related to some important biological functions. In fact, this structural property provides significant functional advantages, as the intrinsically disordered regions may enable enhanced rates of self-assembly processes of viruses and bacterial groups, and play a regulatory role in adding new components in the process of cell growth. Many different types of proteins have been recognized as the causative agents of amyloid diseases, despite having wide and heterogeneous structures and functions, all of them generate morphologically similar amyloid fibrils (Uversky & Fink, 2004; Xing & Higuchi, 2002). The amyloid fibrils are insoluble, rigid and measuring on average 7.5 to 10 μm in length, and can be derived from specific

✉e-mail: polanco@unam.mx

amyloidogenic regions located within globular proteins, unstructured peptides (Uversky *et al.*, 2008), intrinsically disordered proteins, and mostly unfolded fragments of foldable proteins.

In humans, some proteins such as apolipoproteins I and II, are classified as amyloidogenic proteins. Apolipoproteins require a high degree of structural disorder or plasticity to fulfill their biologic function and at the same time to avoid aggregation. For example, lipid-free apolipoprotein behaves as an intrinsically disordered protein but folds to a more ordered structure when lipids are taken up (Andreola *et al.*, 2006). In amyloidoses, multimodal external factors (such as pH, oxidation, toxicants, temperature, etc.) converge independently or simultaneously to destabilize the 3D structure of an ordered protein or affect the conformational ensemble of intrinsically disordered proteins to induce a transition from the native (folded or intrinsically disordered) to partially structured form allowing alternative spatial arrangements of the same polypeptide. However, besides these external features, there are several intrinsic factors that play a role in protein structural stability, with strategically distributed charged residues known to act as efficient modulators of the aggregation process by providing repulsive forces that guard against a pathological conformation (Chiti *et al.*, 1999).

The risk of unwanted protein aggregation, which poses toxic threats to the cells, is minimized by naturally selected sequences of globular proteins that confer the properties of high stability and fast folding kinetics, both of which minimize the concentration of easily aggregating, partially folded proteins. However, despite the evolutionary controlled protection against unwanted aggregation, the misfolded proteins with pathogenic potential can be formed in different ways, e.g. there are proteins that have an intrinsic propensity to assume a pathological conformation (e.g., transthyretin in senile amyloidosis), others acquire pathological conformation when their concentration exceeds a specific threshold (e.g., β2 microglobulin in chronic amyloidosis), or by a replacement in the amino acid sequence of a protein (hereditary amyloidoses), or by a proteolytic degradation of the precursor protein, as is the case of the β-amyloid precursor protein (APP) in Alzheimer's disease.

It is in this scenario that the present work introduces the use of a *Quantitative Structure Activity Relationship* (QSAR) method called Polarity index (Polanco & Samaniego, 2009; Polanco *et al.*, 2012; 2013; 2013a; 2014; 2014a; 2014b; 2014c; 2014d), which from reading the linear sequence of the peptide, identifies whether or not a peptide belongs to any of the next groups: natively unfolded proteins, folded proteins, and amyloidogenic proteins. It also allows to study the relationship of these proteins with neuronal proteins, both human and non-human. The method analyzes comprehensively, the static and dynamic aspect of the peptide, under consideration of a single physico-chemical property of a polypeptide: its polarity. This can be a competitive advantage if we consider other methods, such as CATH: Protein Structure (Sillitoe e*t al.*, 2013), and PSIPRED: Protein Sequence Analysis Workbench (McGuffin e*t al.*, 2000), which are a combination of prediction algorithms: structural comparison (Redfern *et al.*, 2007) and hidden-Markov model (HMM)-based methods (Sillitoe *et al.*, 2005).

The mathematical-computational method has been previously used for the identification of the antimicrobial peptides (Izadpanah & Gallo, 2005) from the Antimicrobial peptide database (APD2), the antiviral peptides (Real *et al.*, 2004) from the AVPpred database, and the set of cell penetrating peptides from the CPPsite database. The algorithm presented here is based on measuring only the polarity or electronegativity of a peptide, being understood by this measure, the construction of an incidence matrix of polar interactions in the peptide from its linear sequence (Polanco & Samaniego, 2009; Polanco *et al.*, 2012; 2013; 2013a; 2014; 2014a; 2014b; 2014c; 2014d). To achieve this, the method considers 20 amino acids classified in four polarity groups, P+, P-, N, and NP (which stay for polar positively charged, polar negatively charged, polar neutral, and non-polar), and counts for the impact of the interactions between the two amino acids. The computational tool here presented, reads the linear sequence of a peptide from N-terminus to C-terminus (not from C-terminus to N-terminus, because the incidences matrix would be different, see Methods and Materials section), moving one amino acid to the right at a time, and records these incidents in a matrix where the rows and columns correspond to the four polar groups. This generates a profile, which so far was proven to be an effective discriminant to identify proteins and peptides with strong pathogenic action. The following groups of proteins were studied in this work: natively unfolded and folded proteins from Oldfield *et al.* (Table 6), proteins of human neurons from the Uniprot database (Table 6), non-human neuronal proteins from the Uniprot database (Table 6), and amyloidogenic proteins from the AmyPDB database (Table 6). These sets were selected with the intention of finding any structural polarity-based differences, between the proteins that are expressed in neurons and those that are actively involved in amyloidosis. For this reason the classification included proteins that are expressed in neurons of various organisms and proteins expressed only in human neurons. In addition, the discriminative efficiency of this approach was evaluated (Table 7) by showing that the proposed computational tool can efficiently classify almost all antibacterial peptides located in the APD2 database, the antiviral peptides from the AVPpred database, the set of 30 selective antibacterial peptides from del Rio *et al.* (2001), the cell penetrating peptides type: non-endocytic, endocytic, and unknown pathway, from the CPPsite database, and the proteins that are expressed in human neurons, and in non-human neurons from the UniProt database.

## METHODS AND MATERIALS

Polarity index method was previously published by this group (patent-pending) (Polanco & Samaniego, 2009; Polanco *et al.*, 2012; 2013; 2013a; 2014; 2014a; 2014b; 2014c; 2014d). However, in order to identify proteins associated with amyloidosis, the following modifications were made to the program. For this purpose, the classification of Timberlake (Timberlake, 1992) was used, which is the simplest known approach, classifying the amino acids as: 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}. Notice that the amino acid G has been considered in the neutral-polar group. We adopted this classification for being a general classification and much oriented towards the polar profile. We do not opt for the Koolman & Rohm classification, because they subdivide the four groups, to get seven subgroups (Koolman & Rohm, 1996). We also did not use other classifications (Devlin, 1992).

**Table 1. Polarity matrix P[*i,j*].**

| | P+ | P– | N | NP |
|---|---|---|---|---|
| **P+** | 0.0403729603 | 0.0193006992 | 0.0546386950 | 0.0600466207 |
| **P-** | 0.0190209784 | 0.0336596742 | 0.0429836847 | 0.0514685325 |
| **N** | 0.0552913770 | 0.0456876457 | 0.1129137501 | 0.1103962734 |
| **NP** | 0.0607925393 | 0.0491375290 | 0.1154312342 | 0.1241025627 |

Polarity matrix **P**[*i,j*] built with the natively unfolded proteins group (Table 7).

Previous versions of the method were published (Polanco & Samaniego, 2009; Polanco *et al.*, 2012; 2013; 2013a; 2014; 2014a; 2014b; 2014c; 2014d), but here we consolidated an improved version that outperforms previous work, for that reason, Section "Example" introduces an example as the basis of the full method explanation presented in section "Polarity Index Method-Modifications".

Eschenfeldt & Berger, 1986) belongs to the category of natively unfolded proteins, according to polarity index method, it is necessary to follow the next steps:

1. Convert the above sequence to its numeric equivalent according to the following rule of equivalence: The amino acids: H, K, and R are replaced by the number "1"; the amino acids: D, and E are replaced by number "2"; the amino acids: C, G, N, Q, S, T, and Y are replaced by number "3"; finally the amino acids: A, F, I, L, M, P, V, and W are replaced by number "4". Note that the four numerical equivalents {1, 2, 3, and 4} correspond to the four polar groups: [**P+**], [**P–**], [**N**], and [**NP**], listed in the same order. The numeric equivalence of the aforementioned sequence is: 43244423332433124 12112442242331244433342322332324232422222233222 22222323222232222242343311442222222423113132222.

2. Read the resulting numerical sequence, from N-terminus to C-terminus, moving one position at a time.

**Table 2. Polarity Index Method testing (natively unfolded proteins)**

| **Position** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Vector **(Q[*i,j*] + P[*i,j*])** of study. | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ | $x_{10}$ | $x_{11}$ | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ |
| Polar interaction 16 is **not** present from 8th to 13th, or 15th or 16th positions. | | | | | | | | × | × | × | × | × | × | | × | × |
| Polar interaction 15 is **not** present in 7th, or from 10th to 16th positions. | | | | | | | × | | | × | × | × | × | × | × | × |
| Polar interaction 14 is **not** present in 1st or 3rd positions | × | | × | | | | | | | | | | | | | |
| Polar interaction 13 is **not** present in 4th or 15th or 16th positions. | | | | × | | | | | | | | | | | × | × |
| Polar interaction 12 is **not** present 3rd, from 9th to 11th, and from 13th to 16th positions. | | | × | | | | | | × | | × | | × | × | × | × |
| Polar interaction 11 is **not** present from 6th to 8th, and from 14th to 16th positions. | | | | | | × | × | | | | | | × | × | | |
| Polar interaction 10 is **not** present in 1st, 2nd, and 15th positions. | × | × | | | | | | | | | | | | | × | |
| Polar interaction 9 is **not** present in 1st, 2nd, 14th, and 16th positions. | × | × | | | | | | | | | | | | × | | × |
| Polar interaction 8 is **not** present in 1st, 2nd, 14th, and 16th positions. | × | × | | | | | | | | | | | | × | | × |
| Polar interaction 7 is **not** present in 1st, 3rd, 4th, and 7th positions. | × | | × | × | | | × | | | | | | | | | |
| Polar interaction 6 is **not** present in 2nd, 3rd, 5th, 7th, and 8th positions. | | × | × | | × | | × | × | | | | | | | | |
| Polar interaction 5 is **not** present in 7th position. | | | | | × | | | | | | | | | | | |
| Polar interaction 4 is **not** present in 5th, and 15th positions. | | | | | × | | | | | | | | | | × | |
| Polar interaction 3 is **not** present in 1st, and 3rd, and 14th positions. | × | | × | | | | | | | | | | | × | | |
| Polar interaction 2 is **not** present from 1st to 3rd, and from 5th to 7th, and 9th, and 12th positions. | × | × | × | | × | × | × | | × | | | × | | | | |
| Polar interaction 1 is **not** present in 2nd, 4th, and 8th positions. | | × | | × | | | | × | | | | | | | | |

Natively unfolded proteins testing by polarity index method. (×): The polar interaction is not present in the position.

## Example

Here we provide a detailed description of an illustrative example showing how the main action of a peptide/protein is identified. To find out if the protein described by sequence MSDAAVDTSSEITTKDLKEKKEVVEEAENGRDAPANGNAENEENGEQEADNEVDEEEEEGGEEEEEEEEGDGEEEDGDEDEEAESATGKRAAEDDDDDVDTKKQKTDEDD (see Appendix A, #1: Each pair is considered as an element (*i,j*) of matrix **Q**[*i,j*]. For this example, the first pair is (*i,j*) = (*4,3*), the second pair will be (*i,j*) = (*3,2*), and so on until the last pair (*i,j*) = (*2,2*) is reached. Note that the pairs (*i,j*) correspond to a square matrix of order 4, that we named matrix **Q**[*i,j*], and where element *i* represents the row, and *j* the column of matrix **Q**[*i,j*]. Note that, if the reading order had been changed, i.e. from N-terminus to C-terminus, the matrix **Q**[*i,j*] would have been different.

**Table 3. Polarity matrix P[*i,j*].**

| | P+ | P– | N | NP |
|---|---|---|---|---|
| **P+** | 0.0196416602 | 0.0187026914 | 0.0405097269 | 0.0556864999 |
| **P-** | 0.0180894900 | 0.0184727404 | 0.0386701152 | 0.0534828007 |
| **N** | 0.0426559374 | 0.0399156846 | 0.1148414314 | 0.1327776164 |
| **NP** | 0.0544792563 | 0.0518156551 | 0.1365909725 | 0.1618089527 |

Polarity matrix **P**[*i,j*] built with the natively **folded** proteins group (Table 7).

3. Count the incidents of every (*i,j*) pair in matrix **Q**[*i,j*]. In this way matrix **Q**[*i,j*] represents the incidents of the numerical sequence in study. Note that pair (*i,j*) = (*1,1*), will have at the end the value of 3, and pair (*i,j*) = (*2,3*) will have the value of 12 (matrix not shown).

4. Repeat steps 2 and 3 but instead of taking only the sequence studied, take the group of peptides/proteins with the characteristics searched of interest and express the incidents in a matrix called **P**[*i,j*], this time to identify the natively unfolded protein group. As this group is formed by 51 proteins (Appendix A), once it finishes counting the incidents in the first peptide/protein it will carry on counting the incidents in the next peptide/protein until completing the group.

5. Normalize to unity matrices **Q**[*i,j*] (matrix peptide in study, matrix not shown), and **P**[*i,j*] (data training, Table 1).

6. Weight matrix **Q**[*i,j*] with matrix **P**[*i,j*], to form a new matrix (**Q**[*i,j*] + **P**[*i,j*]). Finally linearize matrix (**Q**[*i,j*] + **P**[*i,j*]). As a result, matrix (**Q**[*i,j*] + **P**[*i,j*]) becomes a vector (**Q**[*i,j*] + **P**[*i,j*]) of 16 elements, i.e. {6, 7, 10, 8, 11, 14, 16, 9, 15, 2, 1, 12, 3, 13, 5, 4}. Note that we obtain *n* vectors, where *n* is the number of peptides in study.

7. Compare the vector with rules in Table 2. For this example, all the rules are accepted, and therefore this protein is considered as a natively unfolded protein candidate.

**Table 4. Polarity Index Method testing (natively folded proteins)**

| **Position** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Vector **(Q[*i,j*] + P[*i,j*])** of study. | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ | $x_{10}$ | $x_{11}$ | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ |
| Polar interaction 12 is present from 1st to 4th positions. | ✔ | ✔ | ✔ | ✔ | | | | | | | | | | | | |
| Polar interaction 5 is **not** present in 11th 14th positions. | | | | | | | | | | | × | | | × | | |
| Polar interaction 16 is **not** present from 5st to 16rd positions. | | | | | × | × | × | × | × | × | × | × | × | × | × | × |
| Polar interaction 15 is **not** present from 5st to 16rd positions. | | | | | × | × | × | × | × | × | × | × | × | × | × | × |
| Polar interaction 12 is **not** present from 5st to 16rd positions. | | | | | × | × | × | × | × | × | × | × | × | × | × | × |
| Polar interaction 14 is **not** present 1st, 2nd, and from 14th to 16rd positions. | × | × | | | | | | | | | | | × | × | × | × |
| Polar interaction 13 is **not** present from 1st to 3rd, and from 13th to 16th positions. | × | × | × | | | | | | | | | | × | × | × | × |
| Polar interaction 16 is **not** present from 1st to 3rd, and from 13th to 16th positions. | × | × | × | | | | | | | | | | × | × | × | × |
| Polar interaction 9 is **not** present from 1st to 4th , and 16th positions. | × | × | × | | | | | | | | | | | | | × |
| Polar interaction 8 is **not** present from 1st to 3rd , and from 14th to 16th positions. | × | × | × | | | | | | | | | | | × | × | × |
| Polar interaction 11 is **not** present from 9th to 11th, and from 14th to 16th positions. | | | | | | | | | × | × | × | | | × | × | × |
| Polar interaction 7 is **not** present from 1st to 4th , and 16th positions. | × | × | × | × | | | | | | | | | | | | × |
| Polar interaction 3 is **not** present from 1st to 4th positions. | × | × | × | × | | | | | | | | | | | | |
| Polar interaction 6 is **not** present from 1st to 8th positions. | × | × | × | × | × | × | × | × | | | | | | | | |
| Polar interaction 5 is **not** present from 1st to 8th positions. | × | × | × | × | × | × | × | × | | | | | | | | |
| Polar interaction 2 is **not** present from 1st to 8th positions. | × | × | × | × | × | × | × | × | | | | | | | | |
| Polar interaction 5 is **not** present 10th position. | | | | | | | | | | × | | | | | | |
| Polar interaction 4 is **not** present from 1st to 3rd , and from 14th to 16th positions. | × | × | × | | | | | | | | | | | × | × | × |
| Polar interaction 1 is **not** present from 1st to 5th , and 9th and 10th positions. | × | × | × | × | × | | | | × | × | | | | | | |

Natively folded proteins testing by polarity index method. (✔): The polar interaction is present in the position. (×): The polar interaction is not present in the position.

**Table 5. Polarity matrix P[*i,j*].**

| | P+ | P– | N | NP |
|---|---|---|---|---|
| **P+** | 0.0171457380 | 0.0181543119 | 0.0418557748 | 0.0577407964 |
| **P-** | 0.0186585989 | 0.0234493185 | 0.0337871909 | 0.0529500768 |
| **N** | 0.0453857780 | 0.0355521925 | 0.1177508831 | 0.1243066043 |
| **NP** | 0.0539586470 | 0.0519415028 | 0.1313666105 | 0.1722138226 |

Polarity matrix **P**[*i,j*] built with the natively amyloidogenic proteins group (Table 7.

8. If this same sequence is verified with matrix **P**[*i,j*] from Table 3 corresponding to the set of natively **folded** proteins (steps 2–3), and **P**[*i,j*] from Table 5 corresponding to the set of amyloidogenic proteins (steps 2–3), the method will find that it is not accepted in neither of these two groups of proteins.

## Polarity Index Method-Modifications

The polarity index method (Polanco & Samaniego, 2009; Polanco *et al*., 2012; 2013; 2013a; 2014; 2014a; 2014b; 2014c; 2014d), essentially extracts a polarity profile, in the most comprehensive form that we think is possible, from a linear sequence of the peptide/protein, where a count of 16 possible polar interactions is carried out based on the 20 amino acids classified in 4 polarity groups. This count is done by reading pair incidents of amino acids that are observed when slicing the query sequence from N-terminus to C-terminus.

Here we describe the modifications to the original polarity index method (Polanco & Samaniego, 2009; Polanco *et al*., 2012; 2013; 2013a; 2014; 2014a; 2014b; 2014c; 2014d), to identify natively unfolded proteins (identified in red color), natively folded proteins (identified in blue color), and amyloidogenic proteins (identified in green color).

(A) Natively unfolded proteins

Building matrix **P**[*i,j*] with the entire protein set of natively unfolded proteins. When polarity matrix **P**[*i,j*] was concluded, it was normalized to unity (Table 1), and the matrix **Q**[*i,j*] contained the profile of incidents for each sequence under study (Table 7).

Polarity index method qualified as the **natively unfolded proteins** candidates, those proteins in vector (**Q**[*i,j*] + **P**[*i,j*]) that complied with the following rules expressed in Table 2.

(B) Natively folded proteins

Building matrix **P**[*i,j*] with the entire protein set of natively **folded** proteins. When polarity matrix **P**[*i,j*] was completed, it was normalized to unity (Table 3), and the matrix **Q**[*i,j*] contained the profile of incidents for each sequence under study (Table 7).

Polarity index method qualified as the natively folded proteins candidate, those proteins in the vector (**Q**[*i,j*] + **P**[*i,j*]) that complied with the rules in Table 4.

(C) Amyloidogenic proteins

Building matrix **P**[*i,j*] with the entire protein set of amyloidogenic proteins. When polarity matrix **P**[*i,j*] was completed, it was normalized to unity (Table 5), and the

**Table 6. Polarity Index Method testing (amyloidogenic proteins)**

| **Position** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Q[*i,j*] + P[*i,j*]** vector of study. | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ | $x_{10}$ | $x_{11}$ | $x_{12}$ | $x_{13}$ | $x_{14}$ | $x_{15}$ | $x_{16}$ |
| Polar interaction 16 is **not** present in 2nd and from 4th to 16th positions. | | × | | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Polar interaction 15 is **not** present from 5th to 16th positions. | | | | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Polar interaction 14 is **not** present in 9th, and from 1st to 4th, 13th to 16th positions. | × | × | × | × | | | | | × | | | | × | × | × | × |
| Polar interaction 13 is **not** present from 1st to 4th , 8th, and from 10th to 16th positions. | × | × | × | × | | | | × | | × | × | × | × | × | × | × |
| Polar interaction 1 is **not** present from 1st to 13th positions. | × | × | × | × | × | × | × | × | × | × | × | × | × | | | |
| Polar interaction 12 is **not** present 1st , and 5th to 16th positions. | × | | | | × | × | × | × | × | × | × | × | × | × | × | × |
| Polar interaction 11 is **not** present 5th, and 7th to 16th positions. | | | | | × | | × | × | × | × | × | × | × | × | × | × |
| Polar interaction 10 is **not** present in 14th, and from 1st to 7th positions. | × | × | × | × | × | × | × | | | | | | | × | | |
| Polar interaction 9 is **not** present from 1st to 5th, 11th, and from 14th to 16th positions. | × | × | × | × | × | | | | | | × | | | × | × | × |
| Polar interaction 8 is **not** present from 1st to 4th, 10th, and from 12th to 14th positions. | × | × | × | × | | | | | | × | | × | × | × | | |
| Polar interaction 7 is **not** present from 1st to 8th positions. | × | × | × | × | × | × | × | × | | | | | | | | |
| Polar interaction 1 is **not** present from 1st to 13th positions. | × | × | × | × | × | × | × | × | × | × | × | × | × | | | |

Natively amyloidogenic proteins testing by polarity index method. (✔): The polar interaction is present in the position. (×): The polar interaction is not present in the position.

**Table 7. Test databases.**

| # | Database | Classification | Reference |
|---|---|---|---|
| 1 | APD2 | The peptides were classified as unique or multiple action peptides according to the following criteria: (i) *Unique*. A peptide that is only located in a subgroup of the APD2 database, and (ii) *Multiple*. A peptide that is located in two or more subgroups of this database. From all 3636 peptides studied and classified in this database, we found the following peptides with multiple action on: 149 Gram – ONLY, 1711 Gram +/Gram - ONLY, 315 Gram + ONLY, 141 cancer cells, 744 fungi, 244 mammalian cells, 39 chemotaxis; and 1059 with single action on: 111 Gram – ONLY, 213 Gram + ONLY, 518 Gram +/Gram - ONLY, 20 cancer cells, 88 fungi, 88 HIV, 11, and mammalian cells, from the database accessed on March 11, 2012. | Wang & Wang, 2009 |
| 2 | CPPsite | 520 cell penetrating peptides were classified from the database by their uptake mechanism of which 22 peptides exhibited an *endocytic* pathway, 93 a *non-endocytic* pathway, and 405 an unknown pathway. The database presents a record of amino acids with lowercase letters, some inconsistency in the legends of the uptake mechanism, and sometimes duplicated sequences. All inconsistencies were handled as unknown pathway and did not represent more than 7% of the total records from the database accessed on March 11, 2013. | Gautam *et al.*, 2012 |
| 3 | Oldfield *et al.* | 148 proteins: 51 natively unfolded proteins, and 97 natively folded proteins. | Oldfield *et al.*, 2005 supplementary material |
| 4 | Uniprot | 755 human revised proteins expressed in neurons, and 2879 non-human revised proteins expressed in neurons, from the database accessed on March 11, 2014. | Magrane & Uniprot, 2011 |
| 5 | AmyPDB | 15 of 1705 proteins originally classified in several amyloid protein families: α-Fibrinogen, α-Synuclein, Synelfin, Amyloid Precursor Protein (APP), Apolipoprotein A-1 (ApoA1), Atrial Natriuretic Factor (ANF), β2 Microglobulin (Beta2M), Bri2, C Protein (SP-C), Calcitonin (CT), Cystatin C, Gelsolin, Het-S, Huntingtin (htt), Immunoglobulins, Insulin, Islet Amyloid Polypeptide (IAPP), Amylin, Lactadherin, Lactoferrin, lactotransferrin, Lysozyme, Microcin E492, New 1, Parkin, Prion Protein (PrP), Prolactin (PRL), Rnq 1, Serpin, Serum amyloid A (SAA), Sup35, or eRF2, or eRF3, Tau, Transthyretin (TTR), Ure2, or Ure2p, stored in AmyPDB database (Pawlicki et al., 2008), and restricted to: (i) Amyloid formed in vivo (the precursor protein, or a specific sub-segment, forms fibrils in human), and (ii) Amyloid formed in vitro (the polypeptide forms fibrils under experimental conditions), from the database accessed on March 11, 2014. | Pawlicki *et al.*, 2008 |
| 6 | del Rio *et al.* | 30 selective Cationic Amphipathic Antibacterial Peptides (SCAAP). | del Rio *et al.*, 2001 Table 2 and Table 2A, Polanco *et al.*, 2014 |
| 7 | AVPpred | From Thakur *et al.* work (2012) we took 60 validated and experimental peptides from 1245 antiviral peptides. Those peptides were evaluated with 25 physico-chemical properties (Thakur *et al.*, 2012), from the database accessed on March 11, 2013. | Thakur *et al.*, 2012 |

Description of peptides and proteins used to verify the efficiency of the polarity index method.

resulting matrix $\mathbf{Q}[i,j]$ contained the profile of incidents for the sequence under study (Table 7).

Polarity index method qualified as the **a**myloidogenic protein candidates, those proteins in the vector ($\mathbf{Q}[i,j]$ + $\mathbf{P}[i,j]$) that complied with the rules in Table 6.

### Trial Data Preparation

A comprehensive and differentiated set of peptides and proteins was designed to test the groups studied (Table 7). Each group was classified for its multiple or unique action (entry # 1, Table 7). In the remaining cases (entries # 2–7, Table 7), we checked the experimental qualifications given in each database.

### Linear matches

All the proteins and peptides listed in the databases described above (see Section 2.3) were used to find some peculiar amino acid patterns in their sequences. By means of a data mining algorithm based on graphs named Subdue system, (Cook & Holder, 1994; Kukluk *et al.*, 2007; You *et al.,* 2006), we searched for matches of dipeptides, tripeptides, and so on, up to stretches of ten amino acids in length.

### Test Plan

The discriminative efficiency of the polarity index method is determined from calculating two factors: (i) the percentage of success in the identification of the target group, and (ii) the percentage of mistakes in the identification of the other groups. In this sense, the method must be efficient in identifying the target group and simultaneously rejecting those candidates which are not part of this target group.

### Graphics

The polarity matrices of each group studied (Tables 1, and 3), expressed in relative frequency distribution, are interpreted in terms of smoothed curves. Graphs presented in Figs. 1 and 2 can be compared evaluating only two states:

Profiles are considered similar when all their concavities, turning points and points of maximum and minimum match for the 16 polar interactions.

Profiles are considered as dissimilar when the compared curves do not match and differ from each other in their concavities, inflection points and points of maximum or minimum for the 16 possible interactions.

It is important to emphasize that the comparison of these three groups is interpreted with smoothed curves and not with histograms, as the purpose is only to identify their concavities and the maximum or minimum inflection points in the 16 possible interactions (Section "Natively folded proteins"). These graphs provide better and more understandable information on the role of polarity as the main profile to identify the key function of a peptide or protein. The polar interactions (X-axis, Figs. 1 and 2) indeed form a discrete set, the only dense set is the group of real numbers, however, the level of discretization in the X-axis set can be considered a continuum as there are no intermediate elements, so the relation between polar interactions (X-axis) and their relative frequency can be expressed with a smoothed curve.

**Table 8. Polarity index matches by pathogenic action (Natively unfolded proteins).**

| Data-base | APD2 | APD2 | APD2 | APD2 | APD2 | APD2 | APD2 | APD2 | APD2 | AmyPDB |
|---|---|---|---|---|---|---|---|---|---|---|
| Total Hits | Anti-Gram+ ONLY peptides | Anti-Gram– ONLY peptides | Anti-Gram+/ Gram– peptides | Antifungal peptides | Anti-chemotaxis peptides | Anti-parasites peptides | Anti-Cancer cells peptides | Anti-mammalian cells peptides | Anti-HIV | Amyloid proteins |
| Unique action | 45<br>213 | 28<br>111 | 99<br>518 | 11<br>88 | 0<br>0 | 0<br>9 | 9<br>20 | 2<br>11 | 1<br>88 | 5<br>15 |
| Multiple action | 70<br>315 | 37<br>149 | 347<br>1711 | 144<br>744 | 12<br>39 | 8<br>47 | 9<br>20 | 2<br>11 | 0<br>0 | 0<br>0 |
| Data-base | del Rio | AVPpred | CPPsite | CPPsite | CPPsite | Oldfield | Oldfield | Uniprot | Uniprot | % |
| Total Hits | Selective Cationic Amphipatic anti-bacterial peptides | Antiviral peptides | Cells penetrating peptides Non-endocytic pathway | Cells penetrating Endocytic pathway proteins | Cells penetrating Unknown pathway proteins | Natively unfolded proteins | Natively folded proteins | Human neuronal proteins | Non human neuronal proteins | |
| Unique action | 2<br>30 | 9<br>60 | 17<br>93 | 2<br>22 | 0<br>0 | 37<br>51 | 23<br>97 | 278<br>755 | 0<br>0 | 73 |
| Multiple action | 0<br>0 | 0<br>0 | 0<br>0 | 0<br>0 | 85<br>405 | 0<br>0 | 0<br>0 | 0<br>0 | 1077<br>2879 | |

Matches found by Polarity Index method for natively unfolded proteins in both unique and multiple action peptide groups. Unique action: Peptides with pathogenic action against only one group. Multiple action: Peptides with pathogenic action against two or more groups. (%): Percentage hits/total peptides. Database: Sets described in Table 7.

### Polarity matrix

It is worth mentioning that the square matrix $\mathbf{P}[i,j]$ is neither symmetric nor skew-symmetric. A previous work on the characterization of SCAAP evidenced this fact (Polanco & Samaniego, 2009; Polanco *et al.*, 2012; 2013; 2013a; 2014; 2014a; 2014b; 2014c; 2014d), and a similar work using the elements of this matrix related to the formation of copolymers can be found in (Mosqueira *et al.*, 2012).

### Rules Polarity index method

The rules in Tables 2, 4 and 6 are the result of the inspection of the n vectors ($\mathbf{Q}[i,j] + \mathbf{P}[i,j]$) obtained in Section "Example", entry 7, that search the incidents

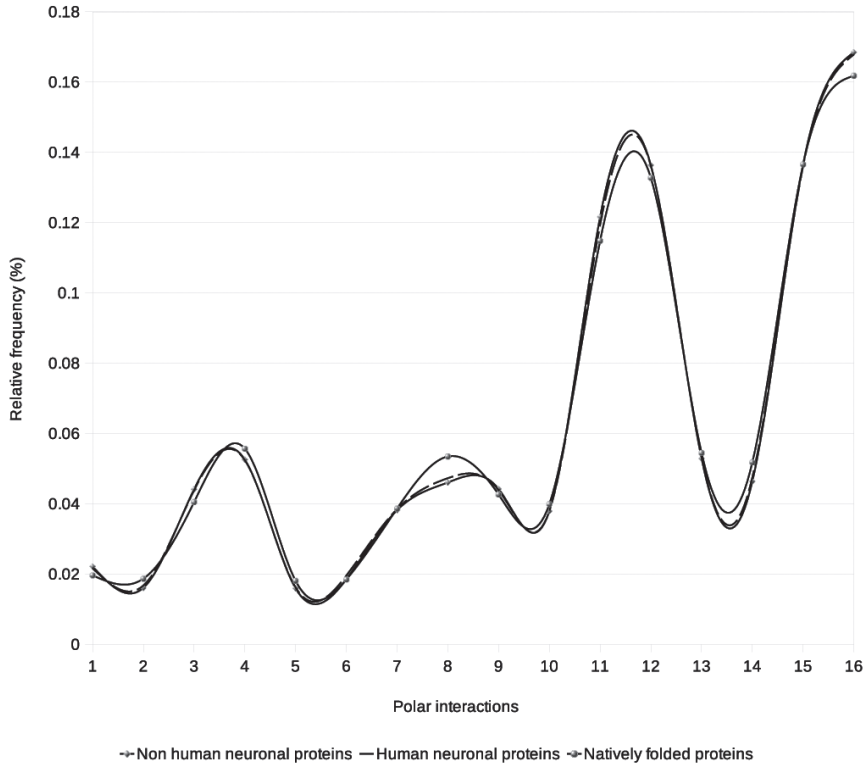


**Figure 1. Comparison of polar group distribution. X-axis corresponds to the 16 polar interactions.**
Human neuronal proteins: Set of sequences expressed in neurons located only in humans (Table 7). Non human neuronal proteins: Set of sequences expressed in neurons located in all living organisms, excluding human beings (Table 7). Natively folded proteins: Set of natively folded proteins (Table 7).

**Table 9. Polarity index matches by pathogenic action (Natively folded proteins).**

| Database | APD2 | APD2 | APD2 | APD2 | APD2 | APD2 | APD2 | APD2 | APD2 | AmyPDB |
|---|---|---|---|---|---|---|---|---|---|---|
| Total Hits | Anti-Gram+ ONLY peptides | Anti-Gram– ONLY peptides | Anti-Gram+/ Gram– peptides | Antifungal peptides | Anti-chemotaxis peptides | Anti-parasites peptides | Anti-Cancer cells peptides | Anti-mammalian cells peptides | Anti-HIV | Amyloid proteins |
| Unique action | 36<br>213 | 13<br>111 | 57<br>518 | 9<br>88 | 0<br>0 | 1<br>9 | 0<br>20 | 3<br>11 | 5<br>88 | 5<br>15 |
| Multiple action | 48<br>315 | 24<br>149 | 233<br>1711 | 82<br>744 | 8<br>39 | 3<br>47 | 13<br>141 | 45<br>244 | 0<br>0 | 0<br>0 |
| Database | del Rio | AVPpred | CPPsite | CPPsite | CPPsite | Oldfield | Oldfield | Uniprot | Uniprot | % |
| Total Hits | Selective Cationic Amphipatic antibacterial peptides | Antiviral peptides | Cells penetrating peptides Non-endocytic pathway | Cells penetrating Endocytic pathway proteins | Cells penetrating Unknown pathway proteins | Natively unfolded proteins | Natively folded proteins | Human neuronal proteins | Non human neuronal proteins | |
| Unique action | 3<br>30 | 3<br>60 | 3<br>93 | 0<br>22 | 0<br>0 | 10<br>51 | 69<br>97 | 431<br>755 | 0<br>0 | 72 |
| Multiple action | 0<br>0 | 0<br>0 | 0<br>0 | 0<br>0 | 53<br>405 | 0<br>0 | 0<br>0 | 0<br>0 | 1571<br>2879 | |

Matches found by Polarity Index method for natively folded proteins in both unique and multiple action peptide groups. Unique action: Peptides with pathogenic action against only one group. Multiple action: Peptides with pathogenic action against two or more groups. (%): Percentage hits/total peptides. Database: Sets described in Table 7.

(or lack of them) in each of the 16 possible polar interactions for each of the 16 positions. As a result the number of possible options is much less than $2^{16}$ in all cases.

### Statistical tests

The purpose of the statistical tests is to verify if the position of the inflection points is a bias for the groups of the Figs. 1 and 2, for that the test considers the matrices $\mathbf{P}[i,j]$ of the groups compared graphically in those figures. The variable being evaluated is the position of the inflection points. The statistical test used is the Kolmogorov-Smirnov test (Siegel, 1970) with $\alpha = 0.01$.

## RESULTS

The application of the polarity index method to identify the target protein groups described in the Methods and Materials section included the verification of its discriminant ability and the graph similarity analysis (Section "Linear matches"), showing the following efficiency results:

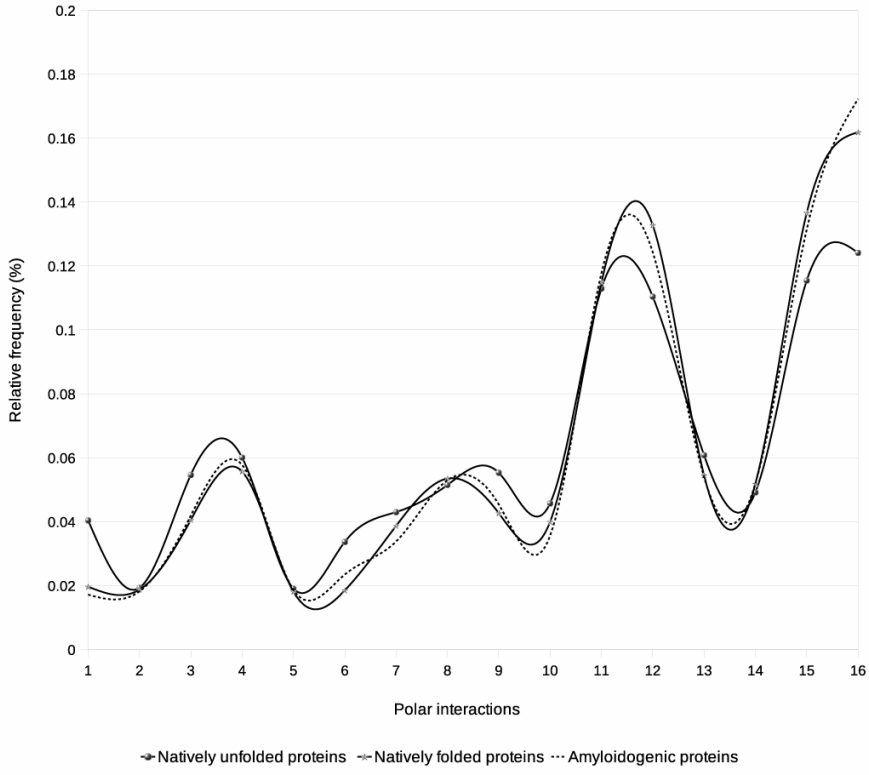


**Figure 2. Comparison of polar group distribution. X-axis corresponds to the 16 polar interactions.**
Natively folded proteins: extracted from Oldfield *et al*. (Table 7). Natively unfolded proteins: Set of natively unfolded proteins extracted from Oldfield *et al*. (Table 7), and Amyloidogenic proteins extracted from Pawlicki *et al*. (Table 7).

**Table 10. Polarity index matches by pathogenic action (amyloidogenic proteins).**

| Database | APD2 | APD2 | APD2 | APD2 | APD2 | APD2 | APD2 | APD2 | APD2 | AmyPDB |
|---|---|---|---|---|---|---|---|---|---|---|
| Total Hits | Anti-Gram+ ONLY peptides | Anti-Gram– ONLY peptides | Anti-Gram+/ Gram– peptides | Antifungal peptides | Anti-chemotaxis peptides | Anti-parasites peptides | Anti-Cancer cells peptides | Anti-mammalian cells peptides | Anti-HIV | Amyloid proteins |
| Unique action | 10<br>213 | 0<br>111 | 22<br>518 | 0<br>88 | 0<br>0 | 0<br>9 | 0<br>20 | 0<br>11 | 1<br>88 | 11<br>15 |
| Multiple action | 15<br>315 | 1<br>149 | 58<br>1711 | 16<br>744 | 1<br>39 | 2<br>47 | 5<br>141 | 10<br>244 | 0<br>0 | 0<br>0 |
| Database | del Rio | AVPpred | CPPsite | CPPsite | CPPsite | Oldfield | Oldfield | Uniprot | Uniprot | % |
| Total Hits | Selective Cationic Amphipatic antibacterial peptides | Antiviral peptides | Cells penetrating peptides Non-endocytic pathway | Cells penetrating Endocytic pathway proteins | Cells penetrating Unknown pathway proteins | Natively unfolded proteins | Natively folded proteins | Human neuronal proteins | Non human neuronal proteins | |
| Unique action | 0<br>30 | 2<br>60 | 2<br>93 | 0<br>22 | 0<br>0 | 2<br>51 | 17<br>97 | 100<br>755 | 0<br>0 | 74 |
| Multiple action | 0<br>0 | 0<br>0 | 0<br>0 | 0<br>0 | 13<br>405 | 0<br>0 | 0<br>0 | 0<br>0 | 346<br>2879 | |

Matches found by Polarity Index method for amyloidogenic proteins in both unique and multiple action peptide groups. Unique action: Peptides with pathogenic action against only one group. Multiple action: Peptides with pathogenic action against two or more groups. (%): Percentage hits/total peptides. Database: Sets described in Table 7.

**Table 10. Similarities among groups.**

| # | Pubmed | AmyPDB database | Polarity index method | Polarity index method | Polarity index method | Reference |
|---|---|---|---|---|---|---|
| | | | Amyloidogenic proteins | Natively unfolded proteins | Natively folded proteins | |
| 1 | 2881207 | A4-HUMAN | ✔ | ✔ | ✔ | Kang *et al.*, 1987 |
| 2 | 6203042 | ANF_HUMAN | ✔ | × | ✔ | Oikawa *et al.*, 1984 |
| 3 | 6406984 | APOA1_HUMAN | ✔ | × | ✔ | Shoulders *et al.*, 1983 |
| 4 | 3312414 | B2MG_HUMAN | ✔ | × | ✔ | Güssow *et al.*, 1987 |
| 5 | 3495457 | CYTC_HUMAN | × | × | × | Abrahamson *et al.*, 1987 |
| 6 | 3020431 | GELS_HUMAN | ✔ | × | ✔ | Kwiatkowski *et al.*, 1986 |
| 7 | 2651160 | IAPP_HUMAN | × | × | × | Mosselman *et al.*, 1989 |
| 8 | 10391242 | ITM2B_HUMAN | ✔ | × | ✔ | Vidal *et al.*, 1999 |
| 9 | 8639264 | MFGM_HUMAN | × | ✔ | × | Couto *et al.*, 1996 |
| 10 | 3755672 | PRIO_HUMAN | × | ✔ | × | Kretzschmar *et al.*, 1986 |
| 11 | 6260780 | PRL_HUMAN | ✔ | ✔ | ✔ | Cooke *et al.*, 1981 |
| 12 | 3312414 | Q540F8_HUMAN | ✔ | × | ✔ | Güssow *et al.*, 1987 |
| 13 | 3312414 | Q6IAT8_HUMAN | ✔ | × | ✔ | Güssow *et al.*, 1987 |
| 14 | 3839415 | SAA_HUMAN | ✔ | × | ✔ | Sipe *et al.*, 1985 |
| 15 | 6093805 | TTHY_HUMAN | ✔ | ✔ | ✔ | Mita et al., 1984 |

Amyloidogenic proteins identified by polarity index method from the AmyPDB database (Pawlicki *et al.*, 2008). PUBMED: National Center for Biotechnology Information, U.S. National Library of Medicine http://blast.ncbi.nlm.nih.gov/ in database: Swiss-Prot (swissprot), accessed March 11, 2014. AmyPDB database: Identification in AmyPDB database (Table 6). Polarity index method: (×): Protein **not** accepted by polarity index method in this set of proteins. (✔): Protein accepted by polarity index method in this set of proteins.

The group of natively unfolded proteins (73%, Table 8); the group of natively folded proteins (70%, Table 9), and the group of amyloidogenic proteins (74%, Table 10). For these three groups the method also showed an efficiency of 72% discriminating false positives (Tables 8–10).

The polarity profiles of the protein groups: (i) human neuronal proteins, (ii) non-human neuronal proteins, (iii) neuronal proteins, and (iv) natively folded proteins (Fig. 1), show entire coincidence in its points of maximum, minimum and points of inflection, with the exception of interactions 8 and 9, where the natively folded proteins do not coincide with the other three groups. However, these four groups show a different profile with respect to the group of natively unfolded proteins, and amyloidogenic proteins. This is illustrated by Fig. 2 which compares the polarity profiles of natively unfolded, natively folded, and amyloidogenic proteins, showing that these profiles are dissimilar, since their points of inflection and maximum/minimum are not coincidental in any of the 16 polar interactions, for the three groups of proteins (Fig. 2).

Only one of the 15 amyloidogenic proteins is not associated with the other two groups (see Table 10, item 7).

The data mining analysis (see Section "Linear matches") did not provide any reliable pattern on the linear sequences of peptides and proteins. In all cases the patterns found in a particular group of peptides were also repeated in the other groups.

Assuming that the statistical test used and the extent of the samples are appropriate, we show that the similarity in the three groups compared in Fig. 1 correlates with the position of the occurrence of the inflection points, and that the lack of similarity between the three groups in Fig. 2 is also verified in the statistical test.

## DISCUSSION

In the past few years, advances in molecular biology, proteomics and bionformatics have combined to improve our understanding of the amyloidoses as a conformational disease. Isolation of the protein components of natural amyloids and the chemical characterization of these components are indispensable investigative tools, because modern classification of amyloidosis is based on the nature of the precursor of the protein that form the fibrillar deposits. Although these proteins are unrelated and diverse, all produce amyloid deposits with a common cross-β structure and similar fibrillar morphology. The number of recognized amyloidogenic proteins is ever expanding, and there are more than 30 amyloid proteins in the AmyPDB database (June 06, 2014) (Pawlicki *et al*., 2008). These proteins have the capacity to acquire more than one spatial conformation and have been recognized as the causative agents of various amyloid diseases, posing increasing clinical difficulties in formulating a correct diagnosis, appropriate treatment, asses prognosis and offer genetic counsel when appropriate.

In this study, polarity index method has shown to be an effective discriminant in the identification of intrinsically disordered (natively unfolded), natively folded, amyloidogenic and neuronal proteins. Therefore, we think that the method can have the following applications: (i) to automate the subsystem that extracts the "template" of the group of proteins/peptides in training, becoming a self-learning algorithm; (ii) to establish a website to enable any user to test any group of proteins and peptides in FASTA format, and (iii) to enable the method to be executed under parallel computing, to explore the total combinatorial divergence of proteins/peptides of a certain length, ($20^n$, where $n < 13$ is the maximum length of the peptide or protein), this will allow to scale this method toward understanding "shortcuts" that nature "found" in the construction of functional proteins and peptides.

An important issue is to understand the reasons behind the effectiveness of polarity, the simple physicochemical property, to differentiate proteins in different structural groups. The two graphs included in this work point out a high correlation between the polar profile of the studied groups and the localization and concavity around the inflection points. If the matrices used here were symmetric, some of these points will surely be catastrophic bifurcation points. However, the matrices are not symmetric, at least not under this four polar group classification. This is a subject this team is currently working on, apart from exploring the construction of an incidence matrix based on seven polarity groups (Koolman & Rohm, 1996).

## CONCLUSIONS

The discriminative efficiency of the polarity index method aimed at the identification of natively unfolded, natively folded, and amyloidogenic proteins, makes it a useful computational tool as a first filter in the analysis of these protein groups, effectively reducing the number of experimental tests in laboratory. The method also allows the identification of other protein groups, such as the human neuronal proteins by their polar profile, opening the possibility to differentiate human neurons by their proteins.

### Availability

The source programs are given as "supplementary material". The sets of natively **unfolded** proteins, natively **folded** proteins, and **amyloidogenic** proteins are given as Appendix section, at the end of this manuscript.

### 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, VU, JACG, JLS, and TB. Mathematical analysis: CP, and JLS. Medical analysis and discussion: CP, VU, and JACG. Documentation appendixes: CP, LT, MLS, and AMA. Data mining: CP, JAG, MAE, and AMR. Results discussion: CP, JLS, VU, JACG, and TB.

### Acknowledgements

The authors want to thank Concepción Celis Juárez for proof-reading and also acknowledge the support of Computer Science Department at Institute for Nuclear Sciences at the National Autonomous University of Mexico. We gratefully acknowledge the financial support of the Mexican-French bilateral research grant CONACYT (188689) — ANR (12-IS07-0006).

## REFERENCES

Abrahamson M, Grubb A, Olafsson I, Lundwall A (1987) Molecular cloning and sequence analysis of cDNA coding for the precursor of the human cysteine proteinase inhibitor cystatin C. *FEBS Lett* **216**: 229–233.

Ackerman SJ, Corrette SE, Rosenberg HF, Bennett JC, Mastrianni DM, Nicholson-Weller A, Weller PF, Chin DT, Tenen DG (1993) Molecular cloning and characterization of human eosinophil Charcot-Leyden crystal protein (lysophospholipase). Similarities to IgE binding proteins and the S-type animal lectin superfamily. *J Immunol* **150**: 456–468.

Aiba H, Mori K, Tanaka M, Ooi T, Roy A, Danchin A (1984) The complete nucleotide sequence of the adenylate cyclase gene of Escherichia coli. *Nucleic Acids Res.* **12**: 9427–9440.

Albrecht JC, Nicholas J, Biller D, Cameron KR, Biesinger B, Newman C, Wittmann S, Craxton MA, Coleman H, Fleckenstein B, Honess RW (1992) Primary structure of the herpesvirus saimiri genome. *J Virol* **66**: 5047–5058.

Andreola A, Bellotti V, Giorgetti S, Mangione P, Obici L, Stoppini M, Torres J, Monzani E, Merlini G, Sunde M (2006) Conformational switching and fibrillogenesis in the amyloidogenic fragment of apolipoprotein A-I. *J Biol Chem* **278**: 2444–24451.

Arya SK, Gallo RC (1986) Three novel genes of human T-lymphotropic virus type III: inmune reactivity of their products with sera from acquired immune deficiency syndrome patients. *Proc Natl Acad Sci USA* **83**: 2209–2213.

Auron PE, Webb AC, Rosenwasser LJ, Mucci SF, Rich A, Wolff SM, Dinarello CA (1984) Nucleotide sequence of human mono-

cyte interleukin 1 precursor cDNA. *Proc Natl Acad Sci USA* **81**: 7907−7911.
Biesiadka J, Sikorski MM, Bujacz G, Jaskolski M (1999) Crystallization and preliminary x-ray structure determination of Lupinus luteus PR10 protein. Acta Crystallogr. *D Biol Crystallogr* **55**: 1925−1927.
Bredt DS, Hwang PM, Glatt CE, Lowenstein C, Reed RR, Snyder SH (1991) Cloned and expressed nitric oxide synthase structurally resembles cytochrome P-450 reductase. *Nature* **351**: 714−718.
Buckanovich RJ, Posner JB, Darnell RB (1993) Nova, the paraneoplastic Ri antigen, is homologous to an RNA-binding protein, is specifically expressed in the developing motor system. *Neuron* **1**: 657−672.
Bujacz G, Jaskolski M, Alexandratos J, Wlodawer A, Merkel G, Katz RA, Skalka AM (1995) High-resolution structure of the catalytic domain of avian sarcoma virus integrase. *J Mol Biol* **253**: 333−346.
Burton ZF, Gross CA, Watanabe KK, Burgess RR (1983) The operon that encodes the sigma subunit of RNA polymerase also encodes ribosomal protein S21, DNA primase in *E. coli* K12. *Cell* **32**: 335−349.
Bussey H, Storms RK, Ahmed A, Albermann K, Allen E, Ansorge W, Araujo R, Aparicio A, Barrell BG, Badcock K, Benes V, Botstein D, Bowman S, Brueckner M, Carpenter J, Cherry JM, Chung E, Churcher CM, Coster F, Davis K, Davis RW, Dietrich FS, Delius H, DiPaolo T, Dubois E, Duesterhoeft A, Duncan M, Floeth M, Fortin N, Friesen JD, Fritz C, Goffeau A, Hall J, Hebling U, Heumann K, Hilbert H, Hillier LW, Hunicke-Smith S, Hyman RW, Johnston M, Kalman S, Kleine K, Komp C, Kurdi O, Lashkari D, Lew H, Lin A, Lin D, Louis EJ, Marathe R, Messenguy F, Mewes HW, Mirtipati S, Moestl D, Mueller-Auer S, Namath A, Nentwich U, Oefner P, Pearson D, Petel FX, Pohl TM, Purnelle B, Rajandream MA, Rechmann S, Rieger M, Riles L, Roberts D, Schaefer M, Scharfe M, Scherens B, Schramm S, Schroeder M, Sdicu AM, Tettelin H, Urrestarazu LA, Ushinsky S, Vierendeels F, Vissers S, Voss H, Walsh SV, Wambutt R, Wang Y, Wedler E, Wedler H, Winnett E, Zhong WW, Zollner A, Vo DH, Hani J (1997) The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XVI. *Nature* **387**: 103−105
Chardin P, Paris S, Antonny B, Robineau S, Beraud-Dufour S, Jackson CL, Chabre M (1996) A human exchange factor for ARF contains Sec7-, pleckstrin-homology domains. *Nature* **384**: 481−484.
Charles IG, Dougan G, Pickard D, Chatfield S, Smith M, Novotny P, Morrissey P, Fairweather NF (1989) Molecular cloning, characterization of protective outer membrane protein P.69 from Bordetella pertussis. *Proc Natl Acad Sci USA* **86**: 3554−3558.
Chen Y, Riley DJ, Chen PL, Lee WH (1997) HEC, a novel nuclear protein rich in leucine heptad repeats specifically involved in mitosis. *Mol Cell Biol* **17**: 6049−6056.
Chiti F, Webster P, Taddei N, Clark A, Stefani M, Ramponi G, Dobson CM (1999) Designing conditions for in vitro formation of amyloid protofilaments, fibrils. *Proc Natl Acad Sci USA* **90**: 3590−3594.
Chou MY, Li SC, Li YT (1996) Cloning, expression of sialidase L, aNeuAcalpha2-->3Gal-specific sialidase from the leech, Macrobdella decora. *J. Biol. Chem.* **271**: 19219−19224.
Christensen HE, Ramachandran S, Tan CT, Surana U, Dong CH, Chua NH (1996) Arabidopsis profilins are functionally similar to yeast profilins: identification of a vascular bundle-specific profilin, a pollen-specific profilin. *Plant J* **10**: 269−279.
Cole ST, Danos O (1987) Nucleotide sequence, comparative analysis of the human papillomavirus type 18 genome. Phylogeny of papillomaviruses, repeated structure of the E6, E7 gene products. *J Mol Biol* **193**: 599−608.
Colombo G, Fanti P, Yao C, Malluche HH (1993) Isolation, complete amino acid sequence of osteocalcin from canine bone. *J Bone Miner Res* **8**: 733−743.
Cook J, Holder (1994) Substructure Discovery Using Minimum Description Length, Background Knowledge. *J Artificial Intelligence Res* **1**: 231−255.
Cooke NE, Coit D, Shine J, Baxter JD, Martial JA *(1981)* Human prolactin. cDNA structural analysis, evolutionary comparisons. *J Biol Chem* **256**: 4007−4016.
Couto JR, Taylor MR, Godwin SG, Ceriani RL, Peterson JA (1996) Cloning, sequence analysis of human breast epithelial antigen BA46 reveals an RGD cell adhesion sequence presented on an epidermal growth factor-like domain. *DNA Cell Biol* **15**: 281−286.
Cram D S, Loh SM, Cheah KC, Skurray RA (1991) Sequence, conservation of genes at the distal end of transfer region on plasmids F, R6-5. *Gene* **104**: 85−90.
Davies GJ, Dauter M, Brzozowski AM, Bjørnvad ME, Andersen KV, Schülein M (1998) Structure of the *Bacillus agaradherans* family 5 endoglucanase at 1.6 A, its cellobiose complex at 2.0 A resolution. *Biochemistry* **37**: 1926−1932.
De Reuse H, Danchin A (1988) The ptsH, ptsI,, crr genes of the Escherichia coli phosphoenolpyruvate-dependent phosphotransferase system: a complex operon with several modes of transcription. *J Bacteriol* **170**: 3827−3837.
del Rio G, Castro-Obregon S, Rao R, Ellerby HM, Bredesen DE (2001) APAP, a sequence-pattern recognition approach identifies substance P as a potential apoptotic peptide. *FEBS Lett* **494**: 213−219.
Devlin, TM (1992) *The Textbook of Biochemistry* — 3rd Edition, Wiley-Liss Inc, NY, accessed May 16, 2014 http: //www.ann.com.au/MedSci/amino.htm
Domenjoud L, Fronia C, Uhde F, Engel W (1988) Sequence of human protamine 2 cDNA. *Nucleic Acids Res* **16**: 7733.
Drocourt D, Calmels T, Reynes JP, Baron M, Tiraby G (1990) Cassettes of the Streptoalloteichus hindustanus ble gene for transformation of lower, higher eukaryotes to phleomycin resistance. *Nucleic Acids Res* **18**: 4009.
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.
Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ (2000) Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* **11**: 161−171.
Eckner R, Birnstiel ML (1989) Cloning of cDNAs coding for human HMG I, HMG Y proteins: both are capable of binding to the octamer sequence motif. *Nucleic Acids Res* **17**: 5947−5959.
Eirin-Lopez JM, Fernanda RM, Gonzalez-Tizon AM, Martinez A, Sanchez L, Mendez J (2004) Molecular evolutionary characterization of the mussel Mytilus histone multigene family: first record of a tandemly repeated unit of five histone genes containing an H1 subtype with 'orphon' features. *J Mol Evol* **58**: 131−144.
Eschenfeldt WH, Berger SL (1986) The human prothymosin alpha gene is polymorphic, induced upon growth stimulation: evidence using a cloned cDNA. *Proc Natl Acad Sci USA* **83**: 9403−9407.
Eylar EH, Brostoff S, Hashim G, Caccam J, Burnett P (1971) Basic A1 protein of the myelin membrane. The complete amino acid sequence. *J Biol Chem* **246**: 5770−5784.
Fahnestock SR, Alexander P, Nagle J, Filpula D (1986) Gene for an immunoglobulin-binding protein from a group G streptococcus. *J Bacteriol* **167**: 870−880.
Fink AL (2005) Natively unfolded proteins. *Curr Opin Struct Biol* **15**: 35−41.
Fujiwara T, Ito K, Nakayashiki T, Nakamura Y (1999) Amber mutations in ribosome recycling factors of *Escherichia coli, Thermus thermophilus*: evidence for C-terminal modulator element. *FEBS Lett* **447**: 297−302.
Funayama N, Nagafuchi A, Sato N, Tsukita S, Tsukita S (1991) Radixin is a novel member of the band 4.1 family. *J. Cell Biol.* **115**: 1039−1048.
Gaillard JL, Berche P, Frehel C, Gouin E, Cossart P (1991) Entry of L. monocytogenes into cells is mediated by internalin, a repeat protein reminiscent of surface antigens from gram-positive cocci. *Cell* **65**: 1127−1141.
Galaktionov K, Beach D (1991) Specific activation of cdc25 tyrosine phosphatases by B-type cyclins: evidence for multiple roles of mitotic cyclins. *Cell* **67**: 1181−1194.
Garcia-Villegas MR, De La Vega FM, Galindo JM, Segura M, Buckingham RH, Guarneros G (1991) Peptidyl-tRNA hydrolase is involved in lambda inhibition of host protein synthesis. *EMBO J* **10**: 3549−3555.
Gautam A, Singh H, Tyagi A, Chaudhary K, Kumar R, Kapoor P, Raghava GP (2012) CPPsite: a curated database of cell penetrating peptides. *Database* (Oxford) **2012**: bas015, accessed May 20, 2013.
Gerber-Huber S, Nardelli D, Haefliger JA, Cooper DN, Givel F, Germond JE, Engel J, Green NM, Wahli W (1987) Precursor-product relationship between vitellogenin, the yolk proteins as derived from the complete sequence of a *Xenopus vitellogenin* gene. *Nucleic Acids Res* **15**: 4737−4760.
Gerhard DS,Wagner L,Feingold EA,Shenmen CM,Grouse LH,Schuler G, Klein SL, Old S, Rasooly R, Good P, Guyer M, Peck AM, Derge JG, Lipman D, Collins FS, Jang W, Sherry S, Feolo M, Misquitta L, Lee E, Rotmistrovsky K, Greenhut SF, Schaefer CF, Buetow K, Bonner TI, Haussler D, Kent J, Kiekhaus M, Furey T, Brent M, Prange C, Schreiber K, Shapiro N, Bhat NK, Hopkins RF, Hsie F, Driscoll T, Soares MB, Casavant TL, Scheetz TE, Brown-stein MJ, Usdin TB, Toshiyuki S, Carninci P, Piao Y, Dudekula DB, Ko MS, Kawakami K, Suzuki Y, Sugano S, Gruber CE, Smith MR, Simmons B, Moore T, Waterman R, Johnson SL, Ruan Y, Wei CL, Mathavan S, Gunaratne PH, Wu J, Garcia AM, Hulyk,SW, Fuh E, Yuan Y, Sneed A, Kowis C, Hodgson A, Muzny DM, McPherson J, Gibbs RA, Fahey J, Helton E, Ketteman M, Madan A, Rodrigues S, Sanchez A, Whiting M, Madari A, Young AC, Wetherby KD, Granite SJ, Kwong PN, Brinkley CP, Pearson RL, Bouffard GG, Blakesly RW, Green ED, Dickson MC, Rodriguez AC, Grimwood J, Schmutz J, Myers RM, Butterfield YS, Griffith M, Griffith OL, Krzywinski MI, Liao N, Morin R, Palmquist D, Petrescu AS, Skalska U, Smailus DE, Stott JM, Schnerch A, Schein JE, Jones SJ, Holt RA, Baross A, Marra MA, Clifton S, Makowski K.A, Bosak S, Malek J (2004) The status, quality, expansion of the NIH full-length

cDNA project: the Mammalian Gene Collection (MGC). *Genome Res* **14**: 2121–2127.
Ghosh S, Gifford AM, Riviere LR, Tempst P, Nolan GP, Baltimore D (1990) Cloning of the p50 DNA binding subunit of NF-kappa B: homology to rel, dorsal. *Cell* **62**: 1019–1029.
Gibbs S, Lohman F, Teubel W, van de Putte P, Backendorf C (1990) Characterization of the human spr2 promoter: induction after UV irradiation or TPA treatment, regulation during differentiation of cultured primary keratinocytes. *Nucleic Acids Res* **18**: 4401–4407.
Gibbs S, Lohman F, Teubel W, van de Putte P, Backendorf C (1990) Characterization of the human spr2 promoter: induction after UV irradiation or TPA treatment, regulation during differentiation of cultured primary keratinocytes. *Nucleic Acids Res* **18**: 4401–4407.
Gill DR, Hatfull GF, Salmond GP (1986) A new cell division operon in *Escherichia coli*. *Mol Gen Genet* **205**: 134–145.
Gilmour J, Liang N, Lowenstein JM (1997) Isolation, cloning, characterization of a low-molecular-mass purine nucleoside-, nucleotide-binding protein. *Biochem J* **326**: 471–477.
Goedert M, Wischik CM, Crowther RA, Walker JE, Klug A. (1988) Cloning, sequencing of the cDNA encoding a core protein of the paired helical filament of Alzheimer disease: identification as the microtubule-associated protein tau. *Proc Natl Acad Sci USA* **85***:* 4051–4055.
Goedert M, Wischik CM, Crowther RA, Walker JE, Klug A. (1988) Cloning, sequencing of the cDNA encoding a core protein of the paired helical filament of Alzheimer disease: identification as the microtubule-associated protein tau. *Proc Natl Acad Sci USA* **85***:* 4051–4055.
Güssow D, Rein R, Ginjaar I, Hochstenbach F, Seemann G, Kottman A, Ploegh HL (1987) The human beta 2-microglobulin gene. Primary structure, definition of the transcriptional unit. *J Immunol* **139**: 3132–3128.
Hall KU, Collins SP, Gamm DM, Massa E, DePaoli-Roach AA, Uhler MD (1999) Phosphorylation-dependent inhibition of protein phosphatase-1 by G-substrate. A Purkinje cell substrate of the cyclic GMP-dependent protein kinase. *J Biol Chem* **274**: 3485–3495.
Kang J, Lemaire HG, Unterbeck A, Salbaum JM, Masters CL, Grzeschik KH, Multhaup G, Beyreuther K, Müller-Hill B (1987) The precursor of Alzheimer's disease amyloid A4 protein resembles a cell-surface receptor. *Nature* **325**: 733–736.
Kawashima S, Kanehisa M (2000) AAindex: amino acid index database. *Nucleic Acids Res* **40** (Web Server issue): W199–W204.
Koolman, J., Rohm, K-H, *Colour Atlas of Biochemistry*, Thieme, Stuttgart, 1996, accessed May 16, 2014 http: //www.ann.com.au/MedSci/amino.htm
Kwiatkowski DJ, Stossel TP, Orkin SH, Mole JE, Colten HR, Yin HL (1986) Plasma, cytoplasmic gelsolins are encoded by a single gene, contain a duplicated actin-binding domain. *Nature* **323**: 455–458.
Hansen JC, Lu X, Ross ED, Woody RW (2006) Intrinsic protein disorder, amino acid composition, histone terminal domains. *J Biol Chem* **281**: 1853–1856.
Harper JW, Adami GR, Wei N, Keyomarsi K, Elledge SJ (1993) The p21 Cdk-interacting protein Cip1 is a potent inhibitor of G1 cyclin-dependent kinases. *Cell* **75**: 805–816.
Hay RE, Woods WD, Church RL, Petrash JM (1987) cDNA clones encoding bovine gamma-crystallins. *Biochem Biophys Res Commun* **146**: 332–338.
Hemmingsen SM, Woolford C, van der Vies SM, Tilly K, Dennis DT, Georgopoulos CP, Hendrix RW, Ellis RJ (1988) Homologous plant, bacterial proteins chaperone oligomeric protein assembly. *Nature* **333**: 330–334.
Hew CL, Wang NC, Joshi S, Fletcher GL, Scott GK, Hayes PH, Buettner B, Davies PL (1988) Multiple genes provide the basis for antifreeze protein diversity, dosage in the ocean pout, *Macrozoarces americanus*. *J Biol Chem* **263**: 12049–12055.
Hu RJ, Watanabe M, Bennett V (1992) Characterization of human brain cDNA encoding the general isoforma of beta-spectrin. *J Biol Chem* **267**: 18715–18722.
Iacangelo AL, Grimes M, Eiden LE (1991) The bovine chromogranin A gene: structural basis for hormone regulation, generation of biologically active peptides. *Mol Endocrinol* **5**: 1651–1660.
Inokuchi K, Mutoh N, Matsuyama S, Mizushima S (1982) Primary structure of the ompF gene that codes for a major outer membrane protein of *Escherichia coli* K-12. *Nucleic Acids Res* **10**: 6957–6968.
Inouye M, Imada M, Tsugita A (1970) The amino acid sequence of T4 phage lysozyme. IV. Dilute acid hydrolysis, the order of tryptic peptides. *J Biol Chem* **245**: 3479–3484.
Izadpanah A, Gallo RL. (2005) Antimicrobial peptides. *J Am Acad Dermatol* **52**: 381-390; quiz 391–392.
Janssen DB, Pries F, van der Ploeg J, Kazemier B, Terpstra P, Witholt B (1989) Cloning of 1,2-dichloroethane degradation genes of Xanthobacter autotrophicus GJ10, expression, sequencing of the dhlA gene. *J Bacteriol* **171**: 6791–6799.
Jauris-Heipke S, Fuchs R, Motz M, Preac-Mursic V, Schwab E, Soutschek E, Will G, Wilske B (1993) Genetic heterogenity of the genes coding for the outer surface protein C (OspC), the flagellin of *Borrelia burgdorferi*. *Med Microbiol Immunol* **182**: 37–50.
Jerse AE, Yu J, Tall BD, Kaper JB (1990) A genetic locus of enteropathogenic *Escherichia coli* necessary for the production of attaching, effacing lesions on tissue culture cells. *Proc Natl Acad Sci USA* **87**: 7839–7843.
Ji H, Liu YE, Jia T, Wang M, Liu J, Xiao G, Joseph BK, Rosen C, Shi YE (1977) Identification of a breast cancer-specific gene, BCSG1, by direct differential cDNA sequencing. *Cancer Res* **57**: 759–764.
Jung A, Sippel AE, Grez M, Schutz G (1980) Exons encode functional, structural units of chicken lysozyme. *Proc Natl Acad Sci USA* **77**: 5759–5763.
Karlsen S, Hough E, Olsen RL (1998) Structure, proposed amino-acid sequence of a pepsin from atlantic cod (Gadus morhua). *Acta Crystallogr D Biol Crystallogr* **54**: 32–46.
Kiefer MC, Bauer DM, Barr PJ (1989) The cDNA, derived amino acid sequence for human osteopontin. *Nucleic Acids Res* **17**: 3306.
Koch C, Moll T, Neuberg M, Ahorn H, Nasmyth K (1993) A role for the transcription factors Mbp1, Swi4 in progression from G1 to S phase. *Science* **261**: 1551–1557.
Koenig M, Monaco AP, Kunkel LM (1988) The complete sequence of dystrophin predicts a rod-shaped cytoskeletal protein. *Cell* **53**: 219–228.
Konecki DS, Benedum UM, Gerdes HH, Huttner WB (1987) The primary structure of human chromogranin A, pancreastatin. *J Biol Chem* **262**: 17026–17030.
Kretzschmar HA, Stowring LE, Westaway D, Stubblebine WH, Prusiner SB, Dearmond SJ (1986) Molecular cloning of a human prion protein cDNA. *DNA* **5**: 315–324.
Kukluk J, You C, Holder L, Cook D (2007) Learning node replacement graph grammars in metabolic pathways. international conference on bioinformatics. *Computational Biology (BIOCOMP-07).*
Lee JS, An G, Friesen JD, Isono K (1981) Cloning, the nucleotide sequence of the genes for *Escherichia coli* ribosomal proteins L28 (rpmB), L33 (rpmG). *Mol Gen Genet* **184**: 218–223.
Lee WH, Shew JY, Hong FD, Sery TW, Donoso LA, Young LJ, Bookstein R, Lee EY (1987) The retinoblastoma susceptibility gene encodes a nuclear phosphoprotein associated with DNA binding activity. *Nature* **329**: 642–645.
Leffers H, Nielsen MS,, ersen AH, Honore B, Madsen P, Vandekerckhove J, Celis JE (1993) Identification of two human Rho GDP dissociation inhibitor proteins whose overexpression leads to disruption of the actin cytoskeleton. *Exp Cell Res* **209**: 165–174.
Li S, Finley J, Liu ZJ, Qiu SH, Chen H, Luan CH, Carson M, Tsao J, Johnson D, Lin G, Zhao J, Thomas W, Nagy LA, Sha B, DeLucas LJ, Wang BC, Luo M (2002) Crystal structure of the cytoskeleton-associated protein glycine-rich (CAP-Gly) domain. *J Biol Chem* **277**: 48596–48601.
Loeb JD, Davis LI, Fink GR (1993) NUP2, a novel yeast nucleoporin, has functional overlap with other proteins of the nuclear pore complex. *Mol Biol Cell* **4**: 209–222.
Luiten RG, Putterman DG, Schoenmakers JG, Konings RN, Day LA (1985) Nucleotide sequence of the genome of Pf3, an IncP-1 plasmid-specific filamentous bacteriophage of *Pseudomonas aeruginosa*. *J Virol* **56**: 268–276.
Luke MM, Sutton A, Arndt KT (1991) Characterization of SIS1, a Saccharomyces cerevisiae homologue of bacterial dnaJ proteins. *J Cell Biol* **114**: 623–638.
Maeda K, Kneale GG, Tsugita A, Short NJ, Perham RN, Hill DF, Petersen GB (1982) The DNA-binding protein of Pf1 filamentous bacteriophage: amino-acid sequence, structure of the gene. *EMBO J* **1**: 255–261.
Maeder DL, Weiss RB, Dunn DM, Cherry JL, Gonzalez JM, DiRuggiero J, Robb FT (1999) Divergence of the hyperthermophilic archaea Pyrococcus furiosus, P. horikoshii inferred from complete genomic sequences. *Genetics* **152**: 1299–1305.
Magrane M., the UniProt consortium (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database bar009* accessed Oct 21, 2013.
Martin-Verstraete I, Debarbouille M, Klier A, Rapoport G (1990) Levanase operon of Bacillus subtilis includes a fructose-specific phosphotransferase system regulating the expression of the operon. *J Mol Biol* **214**: 657–671.
McClelland M, Sanderson KE, Clifton SW, Latreille P, Porwollik S, Sabo A, Meyer R, Bieri T, Ozersky P, McLellan M, Harkins CR,Wang C, Nguyen C, Berghoff A, Elliott G, Kohlberg S, Strong C, Du F, Carter J, Kremizki C, Layman D, Leonard S, Sun H, Fulton L,Nash W, Miner T, Minx P, Delehaunty K, Fronick C, Magrini V, Nhan M, Warren W, Florea L, Spieth J, Wilson RK (2004) Comparison of genome degradation in Paratyphi A, Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid. *Nat Genet* **36**: 1268–1274.
McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. *Bioinformatics* **16**: 404–405.
McHugh KM, Lessard JL (1988) The nucleotide sequence of a rat vascular smooth muscle alpha-actin cDNA. *Nucleic Acids Res* **16**: 4167.

McHugh KM, Lessard JL, (1988) The nucleotide sequence of a rat vascular smooth muscle alpha-actincDNA. *Nucleic Acids Res* **16**: 4167.
McLaren L, Boyle S, Mason JO, Bard JB (2000) Expression, genomic characterization of protein phosphatase inhibitor-1: a novel marker for mesothelium in the Mouse. *Mech Dev* **96**: 237−241.
McLaren L, Boyle S, Mason JO, Bard JB (2000) Expression, genomic characterization of protein phosphatase inhibitor-1: a novel marker for mesothelium in the Mouse. *Mech Dev* **96**: 237−241.
Mengaud J, Braun-Breton C, Cossart P (1991) Identification of phosphatidylinositol-specific phospholipase C activity in Listeria monocytogenes: a novel type of virulence factor? *Mol Microbiol* **5**: 367−372.
Mita S, Maeda S, Shimada K, Araki S (1984) Cloning, sequence analysis of cDNA for human prealbumin. *Biochem Biophys Res Commun* **124**: 558−64.
Mosqueira FG, Negron A, Ramos S, Polanco C (2012) Biased versus unbiased randomness in homo-polymers, copolymers of amino acids in the prebiotic world. *Acta Biochim Pol* **59**: 543−547.
Mosselman S, Höppener JW, Lips CJ, Jansz HS (1989) The complete islet amyloid polypeptide precursor is encoded by two exons. *FEBS Lett* **247**: 154−158.
Nagase T, Ishikawa K, Miyajima N, Tanaka A, Kotani H, Nomura N, Ohara O (1998) Prediction of the coding sequences of unidentified human genes. IX. The complete sequences of 100 new cDNA clones from brain which can code for large proteins *in vitro*. *DNA Res* **5**: 31−39.
Nakashima T, Higa H, Matsubara H, Benson AM, Yasunobu KT (1966) The amino acid sequence of bovine heart cytochrome *c*. *J Biol Chem* **241**: 1166−1177.
Nanmori T, Nagai M, Shimizu Y, Shinke R, Mikami B (1993) Cloning of the beta-amylase gene from *Bacillus cereus*, characteristics of the primary structure of the enzyme. *Appl Environ Microbiol* **59**: 623−627.
Ogata M, Sawada M, Fujino Y, Hamaoka T (1995) cDNA cloning, characterization of a novel receptor-type protein tyrosine phosphatase expressed predominantly in the brain. *J Biol Chem* **270**: 2337−2343.
Ogawa T, Ishii C, Kagawa D, Muramoto K, Kamiya H (1999) Accelerated evolution in the protein-coding region of galectin cDNAs, congerin I, congerin II, from skin mucus of conger eel (*Conger myriaster*). *Biosci Biotechnol Biochem* **63**: 1203−1208.
Oikawa S, Imai M, Ueno A, Tanaka S, Noguchi T, Nakazato H, Kangawa K, Fukuda A, Matsuo H (1984) Cloning, sequence analysis of cDNA encoding a precursor for human atrial natriuretic polypeptide. *Nature* **309**: 724−726.
Oki M, Nishimoto T (1998) A protein required for nuclear-protein import, Mog1p, directly interacts with GTP-Gsp1p, the *Saccharomyces cerevisiae* ran homologue. *Proc Natl Acad Sci USA* **95**: 15388−15393.
Okuda T, Hirai H, Valentine VA, Shurtleff SA, Kidd VJ, Lahti JM, Sherr CJ, Downing JR (1995) Molecular cloning, expression pattern,, chromosomal localization of human CDKN2D/INK4d, an inhibitor of cyclin D-dependent kinases. *Genomics* **29**: 623−630.
Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, Dunker AK (2005) Comparing, combining predictors of mostly disordered proteins. *Biochemistry* **44**: 1989−2000.
Oliva R, Dixon GH (1989) Chicken protamine genes are intronless. The complete genomic sequence, organization of the two loci. *J Biol Chem* **264**: 12472−12481.
Olsen SR, Uhler MD (1991) Inhibition of protein kinase-A by overexpression of the cloned human protein kinase inhibitor. *Mol Endocrinol* **5**: 1246−1256.
Oshima T, Aiba H, Baba T, Fujita K, Hayashi K, Honjo A, Ikemoto K, Inada T, Itoh T, Kajihara M, Kanai K, Kashimoto K, Kimura S, Kitagawa M, Makino K, Masuda S, Miki T, Mizobuchi K, Mori H, Motomura K, Nakamura Y, Nashimoto H, Nishio Y, Saito N, Sampei G, Seki Y, Tagami H, Takemoto K, Wada C, Yamamoto Y, Yano M, Horiuchi T (1996) A 718-kb DNA sequence of the *Escherichia coli* K-12 genome corresponding to the 12.7–28.0 min region on the linkage map. *DNA Res* **3**: 137−155.
Page AP, MacNiven K, Hengartner MO (1996) Cloning, biochemical characterization of the cyclophilin homologues from the free-living nematode *Caenorhabditis elegans*. *Biochem J* **317**: 179−185.
Palm W, Hilschmann N (1975) [The primary structure of a crystalline monoclonal immunoglobulin kappa-type L-chain, subgroup I (Bence-Jones protein Rei.);isolation, characterization of the tryptic peptides; the complete amino acid sequence of the protein; a contribution to the elucidation of the three-dimensional structure of antibodies, in particular their combining site (author's transl)]. *Hoppe-Seyler's Z Physiol Chem* **356**: 167−191.
Palomeque-Messia P, Englebert S, Leyh-Bouille M, Nguyen-Disteche M, Duez C, Houba S, Dideberg O, Van Beeumen J, Ghuysen JM (1991) Amino acid sequence of the penicillin-binding protein/DD-peptidase of Streptomyces K15. Predicted secondary structures of the low Mr penicillin-binding proteins of class A. *Biochem J* **279**: 223−230.
Patti JM, Jonsson H, Guss B, Switalski LM, Wiberg K, Lindberg M,, Hook M (1992) Molecular characterization, expression of a gene encoding a *Staphylococcus aureus* collagen adhesin. *J Biol Chem* **267**: 4766−4772.
Pause A, Belsham GJ, Gingras AC, Donzé O, Lin TA, Lawrence JC Jr, Sonenberg N (1994) Insulin-dependent stimulation of protein synthesis by phosphorylation of a regulator of 5'-cap function. *Nature* **371**: 762−767.
Pawlicki, Le Béchec A, Delamarche C (2008) AMYPdb: a database dedicated to amyloid precursor proteins. *BMC Bioinformatics* **10**: 9: 273, accessed Mar 10, 2014.
Pederson DM, Welsh LC, Marvin DA, Sampson M, Perham RN, Yu M, Slater MR (2001) The protein capsid of filamentous bacteriophage PH75 from *Thermus thermophilus*. *J Mol Biol* **309**: 401−421.
Pelletier I, Pfeifer O, Altenbuchner J, van Pee KH (1994) Cloning of a second non-haem bromoperoxidase gene from Streptomyces aureofaciens ATCC 10762: sequence analysis, expression in Streptomyces lividans, enzyme purification. *Microbiology (Reading, Engl)* **140**: 509−516.
Pentecost BT, Wright JM, Dixon GH (1985) Isolation, sequence of cDNA clones coding for a member of the family of high mobility group proteins (HMG-T) in trout, analysis of HMG-T-mRNA's in trout tissues. *Nucleic Acids Res* **13**: 4871−4888.
Pepys MB (2006) Amyloidosis. *Annu Rev Med* **57**: 223−241.
Perna NT, Plunkett G III, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Posfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157: H7. *Nature* **409**: 529−533.
Piatkowski D, Schneider K, Salamini F, Bartels D (1990) Characterization of five abscisic acid-responsive cdna clones isolated from the desiccation-tolerant plant craterostigma plantagineum, their relationship to other water-stress genes. *Plant Physiol* **94**: 1682−1688.
Plunkett G 3rd, Burland V, Daniels DL, Blattner FR (1993) Analysis of the *Escherichia coli* genome. III. DNA sequence of the region from 87.2 to 89.2 minutes. *Nucleic Acids Res* **21**: 3391−3398.
Polanco C, Buhse T, Samaniego JL, Castañón−González JA (2013) Detection of selective antibacterial peptides by the Polarity Profile method. *Acta Biochim Pol* **60**: 183−189.
Polanco C, Castañón−González JA, Samaniego JL [Letter to the Editor] Arabi YM, Arifi AA, Balkhy HH, Najm H, Aldawood AS, Ghabashi A, Hawa H, Alothman A, Khaldi A, Raiy B (2014) Clinical course, outcomes of critically ill patients with middle east respiratory syndrome coronavirus infection. *Ann Intern Med DOI:* 10.7326/M13−2486.
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 (2014b) Protein misfolding, congophilia, oligomerization,, defective amyloid processing in preeclampsia. *Sci Transl Med* **6***:* 245ra92. DOI: 10.1126/scitranslmed. 3008808.
Polanco C, Castañón-González JA, Uversky VM, Samaniego JL, Buhse T (Letter to the Editor) G Vogel G (2014d) Delays hinder Ebola genomics. *Science* **346**: 684−685 DOI: 10.1126/science.346.6210.684.
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, 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 DOI: 10.1155/2012/585027.
Polanco C, Samaniego JL, Castañón−González JA, Buhse T (2014) Polar profile of antiviral peptides from AVPpred Database. *Cell Biochem Biophys* **70**: 1469−1477. DOI: 10.1007/s12013−014−0084−4.
Polanco C, Samaniego-Mendoza JL, Buhse T, Castañón-González JA, Arias-Estrada M (2013a) Computational model of abiogenic amino acid polymerization to form dipeptides. *Acta Biochim Pol* **61***:* 253–258.
Polanco C, Samaniego JL, Castañón-González JA, Buhse T, Sordo ML (2013b) Characterization of a possible uptake mechanism of selective antibacterial peptides. *Acta Biochim Pol* **60**: 629−633.
Polanco C, Samaniego-Mendoza JL, Buhse T, Castañón-González JA, Leopold-Sordo M (2014a) Polar Characterization of Antifungal Peptides from APD2 Database. *Cell Biochem Biophys* **70**: 1479−1488 DOI: 10.1007/s12013−014−0085−3.
Polanco C., Samaniego-Mendoza JL, Castañón-González JA, Buhse T [Letter to the Editor] Howard SJ, Hopwood S, Davies SC (2014c) Antimicrobial Resistance: A Global Challenge. *Sci Transl Med* DOI: 10.1126/scitranslmed. 3009315.
Przylas I, Tomoo K, Terada Y, Takaha T, Fujii K, Saenger W, Strater N (2000) Crystal structure of amylomaltase from thermus aquaticus, a glycosyltransferase catalysing the production of large cyclic glucans. *J Mol Biol* **296**: 873−886.
Rawas A, Muirhead H, Williams J (1996) Structure of diferric duck ovotransferrin at 2.35 A resolution. *Acta Crystallogr D Biol Crystallogr* **52**: 631-640.

*Real E, Rain JC, Battaglia V, Jallet C, Perrin P, Tordo N, Chrisment P, D'Alayer J, Legrain P, Jacob Y. (2004) Antiviral drug discovery strategy using combinatorial libraries of structurally constrained peptides. J Virol* **78***: 7410–7417.*

Reczek D, Berryman M, Bretscher A (1997) Identification of EBP50: A PDZ-containing phosphoprotein that associates with members of the ezrin-radixin-moesin family. *J. Cell Biol.* **139**: 169–179.

Redfern OC, Harrison A, Dallman T, Pearl FM, Orengo CA (2007) CATHEDRAL: a fast, effective algorithm to predict folds, domain boundaries from multidomain protein structures. *PLoS Comput. Biol* 2007;3: e232

Reid GA (1988) Sequence polymorphisms in the yeast gene encoding aspartyl tRNA synthase. *Nucleic Acids Res.* **16**: 1212.

Remacha M, Saenz-Robles MT, Vilella MD, Ballesta JP (1988) Independent genes coding for three acidic proteins of the large ribosomal subunit from Saccharomyces cerevisiae. *J Biol Chem* **263**: 9094–9101.

Robbins PW, Trimble RB, Wirth DF, Hering C, Maley F, Maley GF, Das R, Gibson BW, Royal N, Biemann K (1984) Primary structure of the Streptomyces enzyme endo-beta-N-acetylglucosaminidase H. *J Biol Chem* **259**: 7577–7583.

Robinson MS (1989) Cloning of cDNAs encoding two related 100-kD coated vesicle proteins (alpha-adaptins). *J Cell Biol* **108**: 833–842.

Robinson MS (1990) Cloning, expression of gamma-adaptin, a component of clathrin-coated vesicles associated with the Golgi apparatus. *J Cell Biol* **111**: 2319–2326.

Rogers MJ, Ohgi T, Plumbridge J, Soll D (1988) Nucleotide sequences of the *Escherichia coli* nagE, nagB genes: the structural genes for the N-acetylglucosamine transport protein of the bacterial phosphoenolpyruvate: sugar phosphotransferase system, for glucosamine-6-phosphate deaminase. *Gene* **62**: 197–207.

Sakakibara M, Mukai T, Hori K (1985) Nucleotide sequence of a cDNA clone for human aldolase: a messenger RNA in the liver. *Biochem Biophys Res Commun* **131**: 413–420.

Saporito SM, Smith-White BJ, Cunningham RP (1988) Nucleotide sequence of the xth gene of *Escherichia coli* K-12. *J Bacteriol* **170**: 4542–4547.

Sauer RT, Krovatin W, Poteete AR, Berget PB (1982) Phage P22 tail protein: gene, amino acid sequence. *Biochemistry* **21**: 5811–5815.

Scharf M, Engels J, Tripier D (1989) Primary structures of new 'isohirudins'. *FEBS Lett* **255**: 105–110.

Schauer AT, Carver DL, Bigelow B, Baron LS, Friedman DI (1987) lambda N antitermination system: functional analysis of phage interactions with the host NusA proteína. *J Mol Biol* **194**: 679–690.

Schluter G, Celik A, Obata R, Schlicker M, Hofferbert S, Schlung A, Adham IM, Engel W (1996) Sequence analysis of the conserved protamine gene cluster shows that it contains a fourth expressed gene. *Mol Reprod Dev* **43**: 1–6.

Schnier J, Kitakawa M, Isono K (1986) The nucleotide sequence of an Escherichia coli chromosomal region containing the genes for ribosomal proteins S6, S18, L9, an open reading frame. *Mol Gen Genet* **204**: 126–32.

Schor SL, Schor AM (2001) Phenotypic, genetic alterations in mammary stroma: implications for tumour progression. *Breast Cancer Res.* **3**: 373–379.

Seedorf K, Krammer G, Durst M, Suhai S, Rowekamp WG (1985) Human papillomavirus type 16 DNA sequence. *Virology* **145**: 181–185.

Seufert W, Jentsch S (1990) Ubiquitin-conjugating enzymes UBC4, UBC5 mediate selective degradation of short-lived, abnormal proteins. *EMBO J* **9**: 543–550.

Shimizu-Matsumoto A, Itoh K, Inazawa J, Nishida K, Matsumoto Y, Kinoshita S, Matsubara K, Okubo K (1996) Isolation, chromosomal localization of the human cone cGMP phosphodiesterase gamma cDNA (PDE6H). *Genomics* **32**: 121–124.

Siegel S (1970) Estadistica no Paramétrica. Trillas pp 69–74.

Sillitoe I, Dibley M, Bray J, Addou S, Orengo C (2005) Assessing strategies for improved superfamily recognition. *Protein Sci* **14**: 1800–1810.

Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R, Yeats C, Thornton JM, Orengo CA (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res* **41**(Database issue): D490-8. doi: 10.1093/nar/gks1211.

Shortle D (1983) A genetic system for analysis of staphylococcal nuclease. *Gene* **22**: 181–189.

Shoulders CC, Kornblihtt AR, Munro BS, Baralle FE (1983) Gene structure of human apolipoprotein A1. *Nucleic Acids Res* **11**: 2827–2837.

Sievers SA, Karanicolas J, Chang HW, Zhao A, Jiang L, Zirafi O, Stevens JT, Münch J, Baker D, Eisenberg D (2011) Structure-based design of non-natural amino-acid inhibitors of amyloid fibril formation. *Nature* **475**: 96–100.

Sipe JD, Colten HR, Goldberger G, Edge MD, Tack BF, Cohen AS, Whitehead AS (1985) Human serum amyloid A (SAA): biosynthesis, postsynthetic processing of preSAA, structural variants defined by complementary DNA. *Biochemistry* **24**: 2931–2936.

Smith DB, Davern KM, Board PG, Tiu WU, Garcia EG, Mitchell GF (1989) Mr 26,000 antigen of Schistosoma japonicum recognized by resistant WEHI 129/J mice is a parasite glutathione S-transferase. *Proc Natl Acad Sci USA* **83**: 8703–8707.

Srivastava A, Lusby EW, Berns KI (1983) Nucleotide sequence, organization of the adeno-associated virus 2 genome. *J Virol* **45**: 555–564.

Stahl HD, Kemp DJ, Crewther PE, Scanlon DB, Woodrow G, Brown GV, Bianco AE, Anders RF, Coppel RL (1985) Sequence of a cDNA encoding a small polymorphic histidine-, alanine-rich protein from *Plasmodium falciparum*. *Nucleic Acids Res* **13**: 7837–7846.

Stewart AF, Willis IM, Mackinlay AG (1984) Nucleotide sequences of bovine alpha S1-, kappa-casein cDNAs. *Nucleic Acids Res* **12**: 3895–3907.

Suva LJ, Winslow GA, Wettenhall REH, Hammonds RG, Moseley JM, Diefenbach-Jagger H, Rodda CP, Kemp BE, Rodriguez H, Chen EY, Hudson PJ, Martin TJ, Wood WI (1987) A parathyroid hormone-related protein implicated in malignant hypercalcemia: cloning, expresión. *Science* **237**: 893–896.

Takami H, Nakasone K, Takaki Y, Maeno G, Sasaki R, Masui N, Fuji F, Hirama C, Nakamura Y, Ogasawara N, Kuhara S, Horikoshi K (2000) Complete genome sequence of the alkaliphilic bacterium Bacillus halodurans, genomic sequence comparison with *Bacillus subtilis*. *Nucleic Acids Res* **28**: 4317–4331.

Takamizawa A, Mori C, Fuke I, Manabe S, Murakami S, Fujita J, Onishi E, Andoh T, Yoshida I, Okayama H (1991) Structure, organization of the hepatitis C virus genome isolated from human carriers. *J Virol* **65**: 1105–1113.

Talarico TL, Ray PH, Dev IK, Merrill BM, Dallas WS (1992) Cloning, sequence analysis,, overexpression of *Escherichia coli* folK, the gene coding for 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase. *J Bacteriol* **174**: 5971–5977.

Tamiya T, Lamouroux A, Julien JF, Grima B, Mallet J, Fromageot P, Menez A (1985) Cloning, sequence analysis of the cDNA encoding a snake neurotoxin precursor. *Biochimie* **67**: 185–189.

Tanaka Y, Tsujimura A, Fujita N, Isono S, Isono K (1989) Cloning, analysis of an Escherichia coli operon containing the rpmF gene for ribosomal protein L32, the gene for a 30-kilodalton protein. *J Bacteriol* **171**: 5707–5712.

Teesalu T, Sugahara KN, Ruoslahti E (2013) Tumor-penetrating peptides. *Front Oncol* **3**: 216.

Thakur N, Qureshi A, Kumar M (2012) AVPpred: collection, prediction of highly effective antiviral peptides. *Nucleic Acids Res* W199–W204, accessed March 10, 2013.

Thompson LH, Brookman KW, Jones NJ, Allen SA, Carrano AV (1990) Molecular cloning of the human XRCC1 gene, which corrects defective DNA strand break repair, sister chromatid Exchange. *Mol Cell Biol* **10**: 6160–6171.

Timberlake KC (1992) *Chemistry.* 5th edn, Haper-Collins Publishers Inc, NY, accessed May 16, 2014 http: //www.ann.com.au/MedSci/amino.htm

Tomaschewski J, Gram H, Crabb JW, Ruger W (1985) T4-induced alpha-, beta-glucosyltransferase: cloning of the genes, a comparison of their products based on sequencing data. *Nucleic Acids Res.* **13**: 7551–7568.

Tompa P (2002) Intrinsically unstructured proteins. *Trends Biochem Sci* **27**: 527–533.

Uversky V.N. (2009) Intrinsic disorder in proteins associated with neurodegenerative diseases. *Frontiers in Bioscience* **14**: 5188–5238.

Uversky V.N. (2010a) Mysterious oligomerization of the amyloidogenic proteins. *FEBS Journal* **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, 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.

Uversky VN, Oldfield CJ, Dunker AK (2008a) Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* **37**: 215–246.

Vallins WJ, Brand NJ, Dabhade N, Butler-Browne G, Yacoub MH, Barton PJ (1990) Molecular cloning of human cardiac troponin I using polymerase chain reaction. *FEBS Lett* **270**: 57–61.

Vaughan JM, Rivier J, Spiess J, Peng C, Chang JP, Peter RE, Vale W (1992) Isolation, characterization of hypothalamic growth-hormone releasing factor from common carp, *Cyprinus carpio*. *Neuroendocrinology* **56**: 539−549.

Vidal R, Frangione B, Rostagno A, Mead S, Révész T, Plant G, Ghiso J (1999) A stop-codon mutation in the BRI gene associated with familial British dementia. *Nature* **399**: 776−781.

Wallon G, Kryger G, Lovett ST, Oshima T, Ringe D, Petsko GA (1997) Crystal structures of *Escherichia coli, Salmonella typhimurium* 3-isopropylmalate dehydrogenase, comparison with their thermophilic counterpart from *Thermus thermophilus*. *J Mol Biol* **266**: 1016−1031.

Wang G, Li X, Wang Z (2009) APD2: the updated antimicrobial peptide database, its application in peptide design. *Nucleic Acids Res* **37**: D933−D937.

Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction, functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **337**: 635−645.

Warmke JW, Ganetzky B (1994) A family of potassium channel genes related to eag in *Drosophila*, mammals. *Proc Natl Acad Sci USA* **91**: 3438−3442.

Wasenius VM, Saraste M, Salven P, Eramaa M, Holm L, Lehto VP (1989) Primary structure of the brain alpha-spectrin. *J Cell Biol* **108**: 79−93.

Watanabe K, Kitamura K, Iha H, Suzuki Y (1990) Primary structure of the oligo-1,6-glucosidase of Bacillus cereus ATCC7064 deduced from the nucleotide sequence of the cloned gene. *Eur J Biochem* **192**: 609−620.

Watson DC, Wong NC, Dixon GH (1979) The complete amino-acid sequence of a trout-testis non-histone protein, H6, localized in a subset of nucleosomes, its similarity to calf-thymus non-histone proteins HMG-14, HMG-17. *Eur J Biochem* **95**: 193−202.

Weber FE, Dyer JH, Lopez Garcia F, Werder M, Szyperski T, Wuthrich K, Hauser H (1998) In pre-sterol carrier protein 2 (SCP2) in solution the leader peptide 1-20 is flexibly disordered,, residues 21-143 adopt the same globular fold as in mature SCP2. *Cell Mol Life Sci* **54**: 751−759.

Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. *J Mol Biol* **293**: 321−331.

Xing Y, Higuchi K. (2002) Amyloid fibril proteins. *Mech Ageing Dev* **123**: 1625−1636.

Xue B, Dunker AK, Uversky VN (2012) Orderly order in protein intrinsic disorder distribution: Disorder in 3500 proteomes from viruses, the three domains of life. *J Biomolecular Structure, Dynamics* **30**: 131−142.

Xue B, Williams RW, Oldfield CJ, Dunker AK, Uversky VN (2010a) Archaic chaos: Intrinsically disordered proteins in Archaea. *BMC Systems Biology* **4** (Suppl 1) S1.

Xue B, Williams RW, Oldfield CJ, Goh GK-M, Dunker AK, Uversky VN (2010) Viral disorder or disordered viruses: Do viral proteins possess unique features? *Protein Peptide Lett* **17**: 932−951.

Yang C, Zhu Y, Magee DM, Cox RA (1996) Molecular cloning, characterization of the Coccidioides immitis complement fixation/chitinase antigen. *Infect Immun* **64**: 1992−1997.

Yang RC, MacKenzie CR, Narang SA (1988) Nucleotide sequence of a Bacillus circulans xylanase gene. *Nucleic Acids Res* **16**: 7187.

You C, Holder L, Cook D (2006) Application of Graph-based Data Mining to Metabolic Pathways. *Workshop on Data Mining in Bioinformatics, IEEE International Conference on Data Mining.*

Zakut-Houri R, Bienz-Tadmor B, Givol D, Oren M (1985) Human p53 cellular tumor antigen: cDNA sequence, expression in COS cells. *EMBO J* **4**: 1251−1255.

Zannis VI, McPherson J, Goldberger G, Karathanasis SK, Breslow JL (1984) Synthesis, intracellular processing,, signal peptide of human apolipoprotein E. *J Biol Chem* **259**: 5495−5499.

Zhang W, Brooun A, McCandless J, Banda P, Alam M (1996) Signal transduction in the archaeon Halobacterium salinarium is processed through three subfamilies of 13 soluble, membrane-bound transducer proteins. *Proc Natl Acad Sci USA* **93**: 4649−4654.

Zhu PP, Reizer J, Peterkofsky A (1994) Unique dicistronic operon (ptsI-crr) in Mycoplasma capricolum encoding enzyme I, the glucose-specific enzyme IIA of the phosphoenolpyruvate:sugar phosphotransferase system: cloning, sequencing, promoter analysis, and protein characterization. *Protein Sci* **3**: 2115–2128.

Zurawski G, Zurawski SM (1985) Structure of the *Escherichia coli* S10 ribosomal protein operon. *Nucleic Acids Res* **13**: 4521−4526.